Using Time-Varying Tolls to Optimize Truck Arrivals at Ports

Xiaoming Chen

Department of Civil and Environmental Engineering
University of Utah, Salt Lake City, UT 84112-0561
E-mail: xiaoming.chen@utah.edu

Xuesong Zhou

Department of Civil and Environmental Engineering
University of Utah, Salt Lake City, UT 84112-0561
Email: zhou@eng.utah.edu
(Corresponding Author)

George F. List

Department of Civil, Construction and Environmental Engineering
North Carolina State University, Raleigh, North Carolina, 27695
Email: gflist@ncsu.edu

Abstract
An analytical point-wise stationary approximation model is proposed to analyze time-dependent truck queuing processes with stochastic service time distributions at gates and yards of a port terminal. A convex nonlinear programming model is developed which minimizes the total truck turn time and discomfort due to shifted arrival times. A two-phase optimization approach is used to first compute a system-optimal truck arrival pattern, and then find a desirable pattern of time-varying tolls that leads to the optimal arrival pattern. Numerical experiments are conducted to test the computational efficiency and accuracy of the proposed optimization models.

Keywords: port management; first best toll pricing; time-dependent queuing model, stochastic service time distribution

1. Introduction

Containerization has brought vast improvements in port handling efficiency, lowering freight transportation costs and greatly boosting trade flows. This great growth in international trade has created many challenges for seaports. Many seaport facilities currently are running at or near capacity, trucks have to wait at gates or at container transfer yards for an extended period of time. These truck delays considerably downgrade the overall productivity of the freight transportation system, and the environmental impacts associated with trucks idling at seaports also become a major concern for port operators and neighboring communities. Although seaports can increase their transfer and storage capacity by adding cranes, or by increasing yard storage density and/or space, the required equipment investments are generally very large, and seaport expansion is also constrained by land availability. To fully utilize the increasingly tight capacity, seaport authorities have sought to deploy a wide range of innovative technologies and strategies, such as truck appointment systems and time-of-day congestion mitigation fees, to achieve better demand management.

An efficient truck appointment mechanism, as shown in a number of previous studies (e.g. Morais and Lord, 2006; Huynh, 2005; Huynh and Hutson, 2005; Huynh and Walton, 2008), holds the promise of reducing truck waiting time and decreasing truck-induced emissions by setting an upper bound, namely an arrival quota, for each time window. A successful truck appointment system gains truck driver support by providing clear benefits, e.g. guaranteed entry times, reduced queue lengths and shorter truck turn times. Therefore, the time-varying length of the truck queue and delays at the gates and yards are essential measures of effectiveness for truck appointment strategies; these performance metrics need to be estimated accurately. For example, underestimating the truck turn times can lead to difficulties in maintaining the desirable level-of-service for truck drivers, and further reduce the attractiveness of the appointment mechanism. On the other hand, spreading out the truck arrival pattern too much can underutilize limited equipment and labor resources and cause inconvenience for the truck drivers. To examine the relationship between truck turn times and crane availability/deployment at terminal yards, Huynh (2005) developed a simulation-based framework to model detailed movements of trucks and yard cranes. By further combining this simulation model with a mathematical programming model, Huynh and Walton (2008) evaluated the effect of truck arrival patterns on truck turn times and crane utilization rates through a heuristic search process.

Extending from the concept of road pricing (e.g., Florian and Hearn, 1995; Sullivan and El Harake, 1998; Supernak et al., 2003), the idea of congestion tolls have been used to provide an incentive for truckers to alter their arrival times, thereby improving system performance. For example, the Pier Pass Program (FHWA, 2009) implemented in the Port of Los Angeles and the Port of Long Beach aims to reduce truck traffic during peak periods on major highways around the ports. In this market incentive approach, the loaded containers by truck entering or exiting the marine terminals during Monday through Friday between 3:00 AM and 6:00 PM are charged a traffic mitigation fee with a fixed toll rate. Holguín-Veras et al. (2005 and 2006) highlighted and examined impacts of time-of-day pricing initiative on the behavior of commercial carriers, based on findings from an empirical study conducted for the Port Authority of New York and New Jersey.

Two major toll pricing strategies have been widely investigated. *First-best* toll pricing assumes that every arc in a network is tollable, while *second-best* toll pricing assumes only some of the arcs can be tolled. First-best toll solutions are often found using a marginal societal cost-based pricing strategy (e.g., Arnott and Small 1994). Another effective option is the toll set approach, developed by Hearn and Ramana (1998) for the static traffic assignment problem with fixed demands. Yildirim and Hearn (2005) extended it to consider static tolled equilibrium with variable demands. Second-best toll strategies are often found using bilevel programming models and mathematical programs with equilibrium constraints (Yang and Lam, 1996, Larsson and Patriksson, 1998 and Lawphongpanich and Hearn, 2004).

This study assumes that every arrival time slot can be tolled and the toll can be different for each time slot. Moreover, the toll set approach is used to compute time-varying tolls that optimize truck arrivals at the port. The solution procedure that finds the time-varying tolls has two stages:

1. For the time-dependent queuing system at the port, find a target truck arrival pattern d^* that minimizes total waiting time and discomfort.
2. Obtain time-varying tolls ω^* that lead to the optimized truck arrival pattern d^* while also minimizing the average toll rate.

The fact that the average toll rate is being minimized is tied to the reality that there are many time-varying toll patterns that lead to the same optimal arrival pattern; the problem is under-specified; and each time-varying toll pattern has a different average toll.

Mathematically, for target solution d^* obtained from the first phase, the feasible region for the time-varying toll patterns $W(\omega)$ is essentially a polyhedron, based on the Karush–Kuhn–Tucker (KKT) optimality conditions for tolled equilibrium (see Hearn and Ramana (1998) for a more detailed proof). To minimize the average toll paid, we construct a linear programming model to find a desirable time-varying toll pattern that leads to the derived system-optimal arrival pattern.

The remainder of this paper is organized as follows. In Section 2, we review different approaches for modeling time-dependent queuing systems. Section 3 presents a set of point-wise fluid-based approximation functions to

represent time-dependent queues with stochastic service times. Corresponding to the first stage in the proposed solution framework, Section 4 presents a nonlinear convex optimization model that derives system-optimal arrival patterns. Section 5 then presents a strategy for finding a time-varying toll pattern that leads to the system-optimal solution obtained from Section 4 while minimizing the average toll rate. Numerical experiments are conducted in Section 6 to illustrate the computational efficiency and accuracy of the proposed optimization models in evaluating the demand management strategies at seaports.

2. Review on time-dependent queuing models

The simplest way to represent time-dependent queues is to subdivide the time period of interest into shorter time intervals, and then apply steady-state queuing theory formulas within each. However, the underlying steady-state assumption does not always hold, especially when the system is over-saturated during peak periods. In addition, the above problem decomposition scheme considers each time slot separately and ignores the impact of residual queues from previous time intervals. In a sophisticated but numerically intensive solution algorithm developed by Hengsbach (1975) and presented later in the textbook by Larson and Odoni (1981), the system state at each interval is described by a probability vector that covers the states of 0 through K customers in the system (K is a sufficiently large number), and the queue evolution pattern is estimated by solving a system of first-order differential equations.

Many computer simulation-based approaches (e.g. Monte Carlo simulation methods) have been extensively used to effectively describe time-dependent queue dynamics under oversaturated situations. On the other hand, a large number of simulation runs (based on different random number seeds) are usually needed to generate average system performance within a desirable confidence interval. This limitation prevents simulation-based methods from being seamlessly incorporated into a mathematical programming framework that requires tractable analytical functions.

As shown in a comparative study by Nie and Zhang (2005), there are four types of macroscopic link flow models available to describe traffic congestion on road networks with time-dependent capacity: (i) analytical travel time functions (e.g. Friesz et al., 1993); (ii) link exit-flow models (e.g. the pioneering dynamic traffic assignment models by Merchant and Nemhauser, 1978); (iii) point-queue model (e.g. Smith, 1984); and (iv) cell transmission models (Ziliaskopoulos 2000). However, the above models do not allow for different service time distributions for the servers (i.e., headway distributions at roadway bottlenecks). Specifically, in an enhanced link exit-flow models proposed by Nie and Zhang (2005), two types of discharge rates (i.e., number of vehicles staying at the last cell and maximum service rate) are used to distinguish light traffic from oversaturated conditions. In the point-queue models and cell transmission models, only deterministic maximum discharge rates are calculated from the mean service rates and they hold as constraints on the exiting flows. Kachani and Perakis (2006) proposed a deterministic fluid dynamics model to investigate dynamic pricing and inventory control strategies in areas of logistics and distribution systems, where an analytical travel time function is used to propagate flow through links.

It is theoretically important and practically useful to explicitly take into account service time distributions in our proposed seaport truck flow optimization model. First, there are a wide range of service time distributions that characterize the underlying servers, for instance, Gamma distributions for container re-handling time by transfer cranes (Kim and Kim 2002), triangular distributions for serving trucks at yards and gates of seaports (Huynh, 2005). It should be recognized that different service time distributions with the same mean service rate can lead to different capacity utilization ratios (shown in Eq. (5) for M/M/1 systems, Eq. (6) for M/G/1 systems, and Fig. 2 in Section 3), especially when the queue system is not oversaturated. To improve system-wide efficiency, one strategy is to increase the capacity utilization rates at the bottlenecks. To do this, logistics system managers have to both reduce the mean service time duration and reduce the variability in the service times. As a result, a desirable queuing model for logistics queuing systems should be sensitive to both the mean and variability of the underlying service time distributions.

In the fields of electrical engineering and management science, a "point-wise" fluid-based approximation model has received increasing attention in the last decade in terms of modeling time-dependent queuing systems. In a pioneering study (Green and Kolesar, 1991), the analysis timeframe is divided into small time intervals and then stationary approximations are used to calculate performance measures in each one, based on the service rates and the time-dependent demand rates. The demand rates consider both the remaining queue from the last analysis interval and the arrival rate for the current interval. Their study also empirically examined the accuracy of this computationally efficient approximation in multi-server queuing systems with exponential service times and periodic Poisson arrival processes. Whitt (1991) further verified that Green and Kolesar's model (1991) is asymptotically correct as the service and arrival rates increase with fixed instantaneous traffic intensity. By combining the point-wise fluid-based approximation model with flow balance equations, Wang et al. (1996) developed a system of nonlinear differential equations to numerically represent queue evolution in telecommunication networks. Along this line, Stolletz (2008) developed a "stationary backlog-carryover approximation" approach for time-dependent queuing systems with Markovian arrivals and exponentially distributed service times. The time-dependent capacity utilization ratio is estimated based on the probability of blocking. Focusing on delay estimation in airport runway systems, Stolletz (2008) further extended the backlog-carryover approximation method to describe time-dependent queuing systems with general service time distributions.

3. Fluid-based approximation functions for modeling time-dependent queues

Since the existing fluid-based approximation studies have provided valuable insights to model non-stationary queuing systems, this section focuses on how to incorporate this method into a mathematical program. The analytical fluid-based approximation method represents a branch of point-wise fluid-based approximation for queuing systems with Poisson arrival processes.

Fig. 1 illustrates a network flow representation of the analytical fluid-based approximation scheme. The analysis time horizon is decomposed into a sequence of time intervals $t = 1, 2, \ldots, T$. Each node represents a system state at one of these intervals. Customers are modeled as being the fluid and for every node there is one (vertical) incoming arc to receive them and one (vertical) outgoing arc to discharge them. The horizontal arcs maintain the fluid balance between consecutive intervals. Given the time-dependent arrival rate in each time interval λ_t from the Poisson process and maximum service rate s_t, the fluid-based approximation algorithm captures the dynamic evolution of the queue, representing it in terms of a system state x_t, the *average* number of customers in the system at interval *t*. The actual discharge rate v_t is determined by the maximum service rate and the capacity utilization ratio ρ_t. An important assumption of the fluid-based approximation model is that the capacity utilization ratio ρ_t further depends on the prevailing system state and the pre-specified variation coefficient of the service time distribution C_s. All of the above variables and parameters are floating-point numbers, rather than integers.

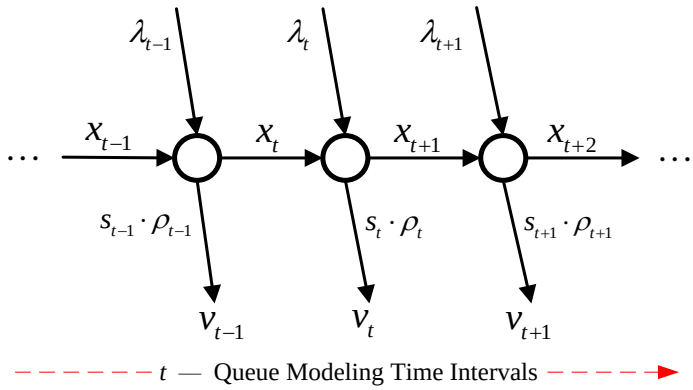


Fig. 1. Illustration of fluid-based approximation scheme

The analytical fluid-based approximation model is comprised of three major components, namely, a flow balance function, an exit flow function, and a time-dependent capacity utilization ratio function.

Flow balance function:

$$x_{t+1} = x_t + \lambda_t - v_t \qquad \forall t \tag{1}$$

Exit flow function:

$$v_t - s_t \times \rho_t \le 0 \qquad \forall t \tag{2}$$

Eq. (1) ensures that flow balance exists during the state transition from one time interval to the next, i.e. the change in the system state is equal to the average number of arrivals minus departures. The key to updating the system state x_t lies in estimating the outgoing flow, and Eq. (2) calculates the actual discharge rate based on the estimated capacity utilization ratios.

Essentially, the fluid-based approximation model utilizes the analytical results of stationary queuing models. It does this by relating the capacity utilization ratio ρ_t to the current state of the system x_t, based on a steady-state queuing relationship.

For a single server with exponentially distributed inter-arrival times and service times (an M/M/1 queue), the steady-state queuing formula is well-known as

$$x = \frac{\rho}{1-\rho} \tag{3}$$

The well-known Pollaczek-Khintchine (P-K) formula (Green and Kolesar, 1991), written as Eq. (4), can be used to estimate average number of customers in a queuing system with a general service time distribution (M/G/1 queue).

$$x = \rho + \frac{\rho^2 \cdot (1 + C_s^2)}{2 \cdot (1-\rho)} \tag{4}$$

By re-arranging Eq. (3), a capacity utilization ratio function for exponentially distributed service times (M/M/1

queue) can be formulated as

$$\rho_t = \frac{x_t}{x_t + 1} \qquad \forall t \tag{5}$$

Likewise, based on Eq. (4), the capacity utilization ratio function for a queuing system with a general service time distribution (M/G/1 queues) can be derived as:

$$\rho_t = \frac{x_t + 1 - \sqrt{(x_t)^2 + 2 \times (C_s)^2 \times x_t + 1}}{1 - (C_s)^2} \qquad \forall t \tag{6}$$

Eqs. (3) and (5) are used when $C_s = 1$; while Eqs. (4) and (6) are used when $C_s \neq 1$. Eqs. (3-6) are only applicable when $\rho_t < 1$. It should be noticed that the time-dependent utilization ratio in this proposed model, ρ_t, is calculated based on the prevailing queue length at time t, which is different from the (time-invariant) steady-state utilization ratio used in a stationary queuing model. That is, $\rho =$ arrival rate/departure rate over the whole planning horizon. As a result, the proposed model can be used to represent over-saturated conditions, under which the queue length can be extremely long, the discharge flow is near full capacity, and ρ_t is close to 1.

As illustrated in Fig. 2, the average capacity utilization ratio monotonically increases with the average number of customers in the system. Under low system load, the service capacity is not fully utilized due to voids in the customer arrival pattern. As the system load increases, the capacity utilization ratio rises gradually to 1. Moreover, as the variance in the service time distribution increases, the model yields a lower capacity utilization ratio for the same arrival rate.

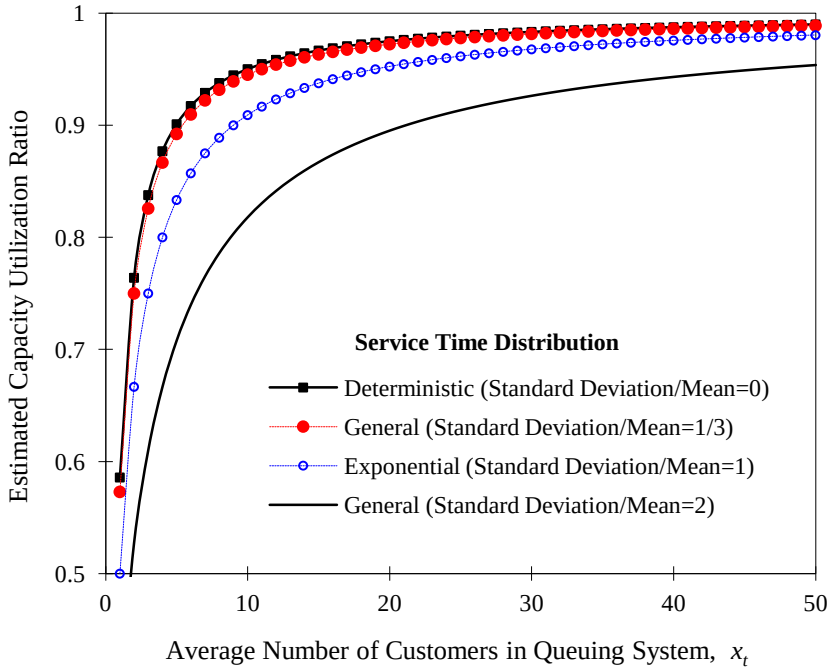


Fig. 2. Capacity utilization ratio as a function of average number of customers for different service time distributions

The capacity utilization ratio functions successfully reproduce the underlying mechanism. That is, with the same mean discharge rate, the utilization rate of the full capacity is dependent on the prevailing queue length and service time distributions. Under oversaturated conditions, in particular, the server is always busy so its capacity is fully utilized. When the server is under-saturated, capacity is not fully utilized because the server is sometimes idle when no customers are in queue due to the randomness in the arrival headways. When designing effective time-of-day pricing strategies, we need to accurately estimate the system performance for both under- and over-saturated situations, as well as the transitions between these two states. If we only use a deterministic maximum service capacity as a constraint, we can overestimate the discharge rate and underestimate the traffic congestion when the system is not fully saturated.

The capacity utilization ratio functions are concave under a wide range of input parameters, as shown in Fig. 2. Since the second-order derivative of ρ_t with respect to x_t is always negative for the interval $x_t \geq 0$, Eq. (5) for an M/M/1 queuing process can be easily proved to be concave. Lemma 1 below aims to show that, in Eq. (6) for an M/G/1 process, ρ_t is also a concave function of queue length x_t.

Lemma 1: $\rho_t = \dfrac{x_t + 1 - \sqrt{(x_t)^2 + 2 \cdot (C_s)^2 \cdot x_t + 1}}{1 - (C_s)^2}$ **is a concave function of** x_t.

Proof: For simplicity of notation, let us ignore constant denominator $1-(C_s)^2$ and focus on $\rho'_t = x_t + 1 - \sqrt{(x_t)^2 + 2\cdot(C_s)^2 \cdot x_t + 1}$, where we denote $M = (x_t)^2 + 2\cdot(C_s)^2 \cdot x_t + 1$.

$\rho'_t(x_t)$ is a continuous and differentiable function, and the first-order derivative of $\rho'_t(x_t)$ is

$$\frac{\mathrm{d}\,\rho'_t(x_t)}{\mathrm{d}\,x_t} = 1 - M^{-\frac{1}{2}} \cdot [x_t + (C_s)^2]$$

The second-order derivative of $\rho'_t(x_t)$ can be calculated as

$$\begin{aligned}
\frac{\mathrm{d}^2\rho'_t(x_t)}{\mathrm{d}(x_t)^2} &= -M^{-\frac{1}{2}} + \frac{1}{2}M^{-\frac{3}{2}} \cdot \left(2x_t + 2(C_s)^2\right)\left(x_t + (C_s)^2\right) \\
&= M^{-\frac{1}{2}}\left(M^{-1} \cdot \left(x_t + (C_s)^2\right)^2 - 1\right) \\
&= M^{-\frac{1}{2}}\left(\frac{(x_t)^2 + 2\cdot(C_s)^2 \cdot x_t + (C_s)^4}{(x_t)^2 + 2\cdot(C_s)^2 \cdot x_t + 1} - 1\right)
\end{aligned} \tag{7}$$

The number of customers in the system is positive $x_t \geq 0$, so $M > 0$ and $M^{-\frac{1}{2}} > 0$.

Since $0 \leq C_s \leq 1$, $(x_t)^2 + 2\cdot(C_s)^2 \cdot x_t + (C_s)^4 < (x_t)^2 + 2\cdot(C_s)^2 \cdot x_t + 1$; and

since $\left(\frac{(x_t)^2 + 2\cdot(C_s)^2 \cdot x_t + (C_s)^4}{(x_t)^2 + 2\cdot(C_s)^2 \cdot x_t + 1} - 1\right) < 0$, the second-order derivative is negative.

Thus ρ_t is a concave function of queue length x_t.

To account for the blocking effects of finite storage capacities, one can impose an additional storage capacity constraint, written as constraint (8).

$$x_t \leq \bar{x}_{\max} \quad \forall t \tag{8}$$

where $\bar{x}_{\max}$ denotes queue storage capacity.

To illustrate the advantages of using this fluid-based approximation to capture the queuing dynamics, consider an M/M/1 queuing system with a service rate of 30 customers per hour, and sequential hourly arrival rates of 20, 25 and 20 customers per hour over a 3-hour horizon. The length of each time interval is set to 6 minutes. In Fig.3, the traditional stationary queuing model is applied separately to each time interval; and the result is compared to the fluid-based model applied to the entire analysis horizon. The stationary model fails to capture the transitions in system state (i.e. queue length) between the different demand rates, and it overestimates the maximum queue length during the short congested period. Thus, while the stationary model is more suitable for estimating the queue length when the analysis horizon is long, the fluid-based approximation model more effectively represents the queue length during dynamic conditions.

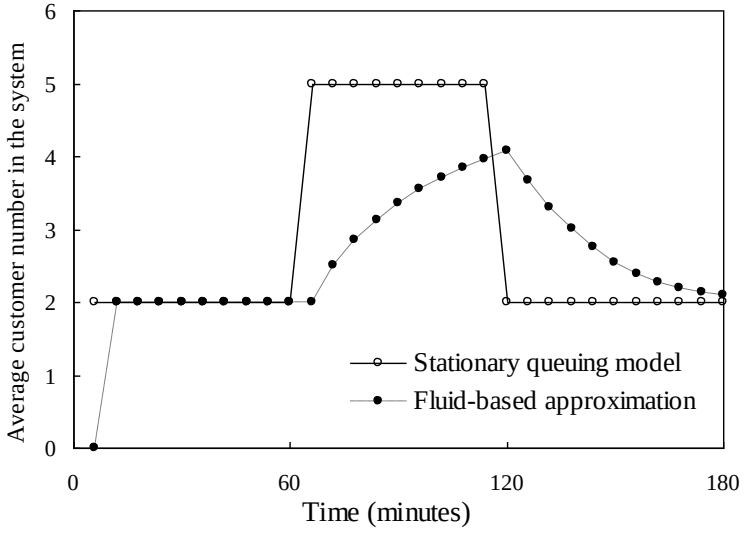


Fig. 3. Comparison between point-wise stationary approximation and stationary queuing model

4. Solution Stage I: Derive system-optimal truck arrival pattern

4.1. Problem statement

The workflow of a truck at a port can be characterized as follows. The truck arrives at a terminal gate during its preferred or assigned appointment time window, and they randomly chooses an entry gate. After finishing its entry paperwork, it proceeds to a designated yard zone to join a queue, if any, to wait for an available yard crane or straddle carrier to load/unload its container(s). Finally, it departs at an exit gate. The gates and yard zones can be modeled jointly as a two-layer queuing network, as shown in Fig. 4. In the gate layer (or yard layer), the gate lanes (or yard zones) with single waiting lines can be viewed as multiple independent M/M/1, M/G/1 or G/G/1 queues, depending on the specific distributions for the inter-arrival times and service times.

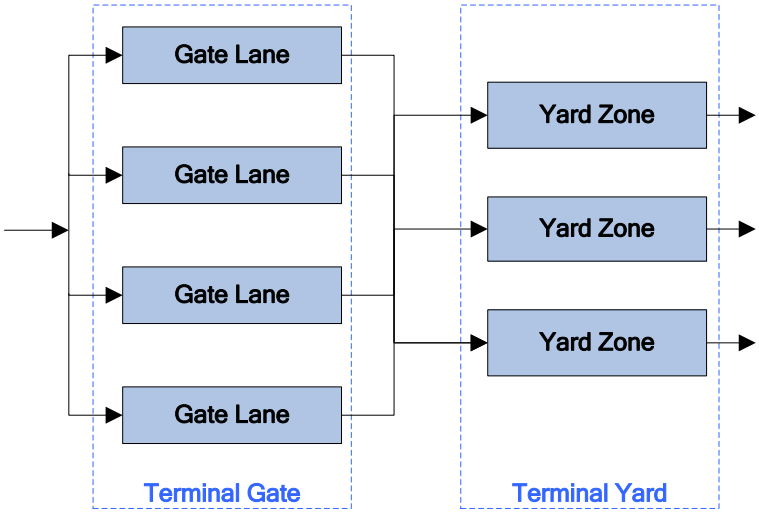


Fig. 4. Queuing network in a seaport terminal

The proposed truck arrival pattern optimization model considers two criteria that can potentially conflict: (a) minimizing the total trucks waiting at the gates and yards and (b) minimizing the difference between the truckers' assigned arrival times and their preferred arrival times. The first objective aims to improve the system-wide performance, while the second concentrates on reducing the inconvenience to the truckers due to the shifted arrival times. The major inputs for this optimization model are: 1) the preferred arrival times for the trucks, 2) the gate service plan, and 3) the yard service plan. This model further assumes that the truckers comply with their assigned arrival times and arrive at terminal gates within their designated time windows. The model also assumes that the proportions of trucks flows headed to specific yard zones are time-invariant, and that these proportions can be estimated from historical data. Finally, the model also assumes that the yard destination proportions remain constant over the entire analysis horizon. It should be remarked that, at the daily operational level, the yard proportion matrix can be time-varying, and the truck appointment system may already know the exact container location when a truck makes an appointment (say one or two days before coming to the port). However, this study focuses on a planning-level application, which considers how to determine a (stable) truck flow quota pattern across multiple days (say within one month).

Now, we define δ_j as the proportion of the truck flows assigned to zone *j* (i.e. yard destination proportion), where $\sum \delta_j = 100\%$. For simplicity, the constant percentage δ_j is used in this study, as the exact locations of containers are changed every day and typically unknown for a planning-level application.

4.2. Notations in the system-optimal inflow pattern model

The sets and subscripts, parameters and decision variables used in the system-optimal inflow pattern model are as follows:

Indices:

- p = index of truckers' preferred arrival time windows, p =1, …, *P*, where *P* is the number of arrival time windows.
- τ = index of assigned appointment time windows, τ =1, …, *P*.
- t = index of fluid-based modeling time intervals, *t*=1, …, *T*, where *T* is the number of modeling time intervals.
- i = index of terminal gate lanes, *i* = 1, …, *m*, where *m* is the number of lanes.
- j = index of terminal yard zones, *j* = 1, …, *n*, where *n* is the number of zones.

Input Parameters:

w^{a} = penalty for shifting truckers' arrival times from their preferred arrival times

α = polynomial penalty coefficient for deviation between p and τ

w^{g} = penalty coefficient for number of queuing trucks at gate lanes

w^{y} = penalty coefficient for number of queuing trucks at yard zones

$\overline{d}_{p}$ = truck demand with preferred arrival time window p

σ = number of approximation intervals in one appointment time window

$s^{g}_{i,t}$ = maximum service rate (i.e. capacity) of gate lane *i* at interval *t*

$s^{y}_{j,t}$ = maximum service rate of yard zone *j* at interval *t*

$C^{g}_{s,i}$ = coefficient of variation of service time distribution at gate lane *i*

$C^{y}_{s,j}$ = coefficient of variation of service time distribution at yard zone *j*

β = trucker maximum tolerable arrival time shifts

Decision variables:

$d_{p,\tau}$ = truck demand assigned to arrive at time τ with preferred arrival time *p*

d_{τ} = truck flow assigned in appointment time window τ

$x_{i,t}$ = average number of trucks on gate lane *i* at queuing model interval *t*

$y_{j,t}$ = average number of trucks in yard zone *j* at queuing model interval *t*

$\lambda^{g}_{i,t}$ = arrival flow rate at gate lane *i* at queuing model interval *t*

$v^{g}_{i,t}$ = actual discharge rate of gate lane *i* at queuing model interval *t*

$\rho^{g}_{i,t}$ = capacity utilization ratio of gate lane *i* at queuing model interval *t*

$\lambda^{y}_{j,t}$ = arrival flow rate at yard zone *j* at queuing model interval *t*

$v^{y}_{j,t}$ = actual discharge rate of yard zone *j* at queuing model interval *t*

$\rho^{y}_{j,t}$ = capacity utilization ratio of yard zone *j* at queuing model interval *t*

Fig. 5 illustrates the system-optimal inflow pattern model. The arcs are associated with the model variables, and the nodes represent processes related to those variables. It is important to note that, p, τ and *t* are different, but inter-related and synchronized indices of time. Both *p* and τ are the indices for the appointment time windows and *t* is the index for the fluid-based queue modeling time intervals. The duration of each τ is the same as that of each *p*, which is typically set to 15 minutes or one hour, as it is convenient for truckers to make appointments. To achieve satisfactory accuracy in the queue evolution approximation, a fine-grained time resolution, such as 1 minute, is required for *t*, and multiple queue modeling time intervals can fit within each assigned appointment time window. As an example, for an over-congested time window *p*, some of the corresponding preferred arrival demand d_{p} is shifted from time $p = \tau$ to neighboring time windows $\tau - 1$, $\tau + 1$. Subsequently, the adjusted demand flow from different preferred arrival times is aggregated to d_{τ} at each appointment time interval τ. d_{τ} is further distributed to the corresponding queue modeling intervals as input λ_{t} (within the same appointment time interval τ) to fluid-based approximation.

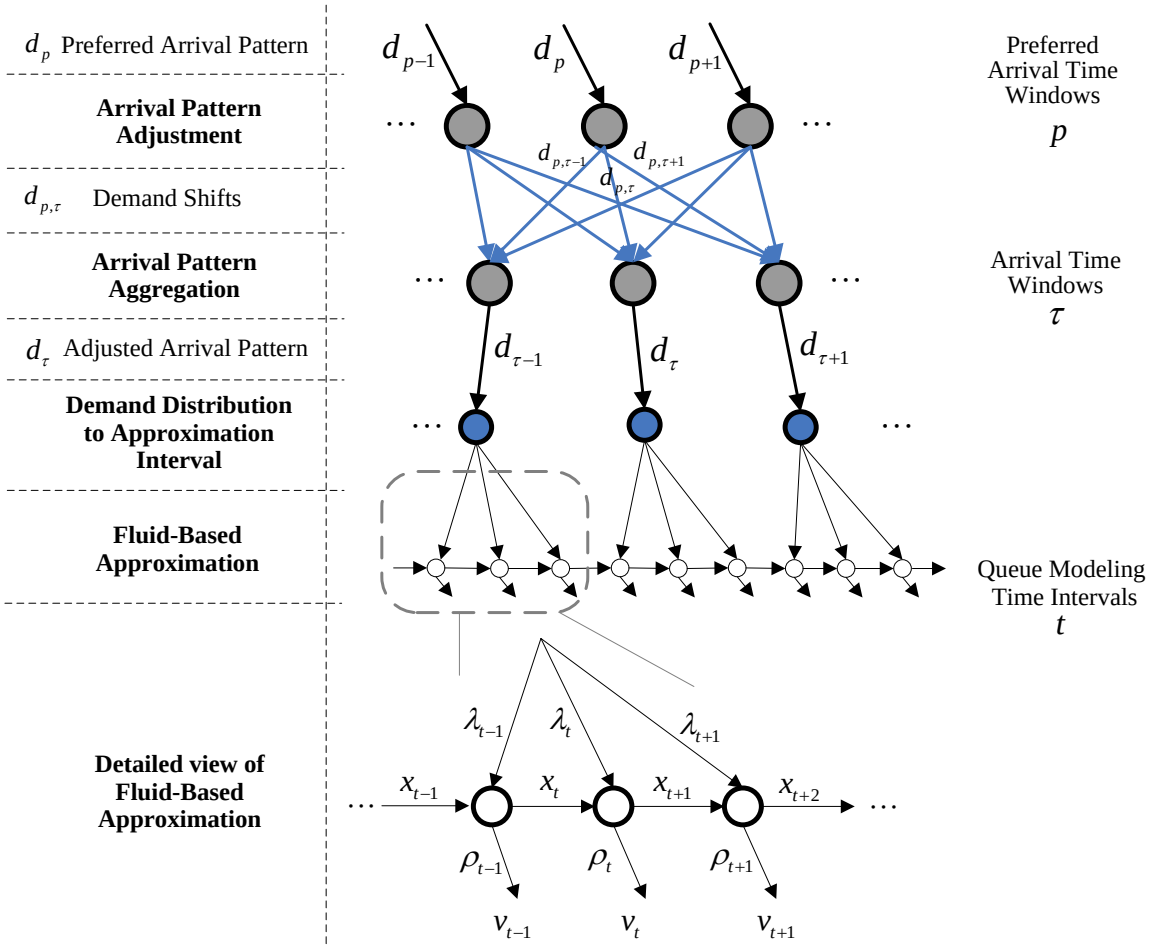


Fig. 5. Network representation of system-optimal inflow pattern determination model

4.3. Truck arrival pattern optimization model

The problem can be formulated and solved as a dynamic system optimum network flow assignment problem with side constraints.

Objective function:

$$\min z = w^{a} \times \sum_{p,\tau} \left[(p-\tau)^2 \times d_{p,\tau} \right] + w^{g} \times \sum_{t} \sum_{i} x_{i,t} + w^{y} \times \sum_{t} \sum_{j} y_{j,t} \tag{9}$$

The first summation in the objective function corresponds to the disutility of the arrival time adjustments made by the truckers. The deviation between p (preferred arrival time window) and τ (assigned arrival time window) represents a measure of trucker disutility. A quadratic function, which is smooth and is differentiable, is used to characterize how the disutility increases nonlinearly when the schedule deviation becomes large. The second and third summations describe the average number of trucks at the gate lanes and yard zones over the analysis horizon, respectively. From a multi-objective optimization perspective, port operators can use the weights w^{a}, w^{g} and w^{y} to balance the trucker disutility against system-wide performance. The calibration results on similar objective functions can be found in the studies by Small (1982), and Noland et al. (1998).

Preferred arrival pattern constraint:

$$\sum_{\tau} d_{p,\tau} = \bar{d}_p \qquad \forall p \tag{10}$$

Assigned arrival pattern constraint:

$$d_\tau = \sum_{p} d_{p,\tau} \qquad \forall \tau \tag{11}$$

Constraints (10) and (11) imply that both the assigned arrival pattern d_τ and the trucker preferred arrival pattern $\bar{d}_p$ can be expressed as linear combinations of decision variables $d_{p,\tau}$, as shown in Fig. 5. As a practical requirement, constraint (12) does not allow time shifts to exceed the truckers' maximum tolerable range β (e.g. 2 hours used in the case study below). That is, the assigned time window τ has to be within a certain range of the preferred arrival time p.

Maximum tolerable time shift constraint:

$$d_{p,\tau} = 0 \qquad \forall p, \quad |p - \tau| > \beta \tag{12}$$

Truck arrival flow distribution constraint:

$$\lambda^{\mathrm{g}}_{i,t} = \frac{d_{\tau}}{\sigma \cdot m} \qquad \forall i, \tau, t = \tau \cdot \sigma, \tau \cdot \sigma + 1, \cdots, (\tau + 1) \cdot \sigma - 1 \tag{13}$$

Constraint (13) converts the aggregated assigned arrival pattern into a sequence of arrival rates for each queuing model time interval (within an appointment window), and evenly distributes the demand among gate lanes.

Constraints for queuing process at gate lanes (Fluid-based approximation):

$$x_{i,t+1} = x_{i,t} + \lambda^{\mathrm{g}}_{i,t} - v^{\mathrm{g}}_{i,t} \qquad \forall i, t \tag{14}$$

$$v^{\mathrm{g}}_{i,t} \le s^{\mathrm{g}}_{i,t} \times \frac{x_{i,t}}{x_{i,t} + 1} \qquad \forall i, t \tag{15}$$

To capture the time-varying queue dynamics at the seaport terminal gate lanes, the analytical fluid-based approximation model is embedded as constraints (14) and (15) in the optimization framework.

Truck arrival flow constraint at yard zones:

$$\lambda^{\mathrm{y}}_{j,t} = \sum_{i} \left(v^{\mathrm{g}}_{i,t} \times \delta_{j} \right) \qquad \forall j, t \tag{16}$$

The yard arrival rates are determined by gate discharge rates and a predefined (time-invariant) yard destination proportion matrix.

Constraints for queuing process at yard zones (fluid-based approximation):

$$y_{j,t+1} = y_{j,t} + \lambda^{\mathrm{y}}_{j,t} - v^{\mathrm{y}}_{j,t} \qquad \forall j, t \tag{17}$$

$$v^{\mathrm{y}}_{j,t} \le s^{\mathrm{y}}_{j,t} \times \frac{y_{j,t} + 1 - \sqrt{(y_{j,t})^2 + 2 \cdot (C^{\mathrm{y}}_{s,j})^2 \cdot y_{j,t} + 1}}{1 - (C^{\mathrm{y}}_{s,j})^2} \qquad \forall j, t \tag{18}$$

The flow from the gate system (comprising *n* M/M/1 queues) to the yard zones is viewed as a Poisson arrival process to take advantage of the "equivalence property" of M/M/1 systems: an M/M/1 queuing system with infinite queuing capacity will have a departure process that involves an exponential distribution identical to the arrival process (Larson and Odoni, 1981). Some existing empirical results reveal that the yard services are more appropriate to be modeled as non-Poisson processes. For example, Huynh (2005) used log-normal distributions to approximate the container loading time, based on data collected at the Port of Houston's Barbours Cut Terminal. Similarly, an empirical result presented by Kim and Kim (2002) also suggests that the yard transfer operation is a non-Poisson process, and that it should be represented as an M/G/1 queue. Following these results, this study treats the yard zone queues as M/G/1, and uses the corresponding capacity utilization ratio function, shown as Eq. (18).

Notably, the proposed model has a problem structure similar to existing system-optimal dynamic traffic assignment (SO-DTA) models, especially those models addressing departure time choice or demand spreading strategies (e.g. Chow, 2009). These SO-DTA models, sometimes referred to as "normative" dynamic traffic assignment models, seek to optimize some system-wide performance metric such as total network travel time.

4.4. Further discussion about model convexity

Lemma 2: The proposed truck arrival flow optimization model has a unique optimal solution.

Proof: The standard form of a convex optimization problem can be usually written as $\min \left\{ f(x) \middle| h_i(x) = 0; g_j(x) \le 0 \right\}$, where $f(x)$ is a convex function to be minimized over variable x. The $h_i(x) = 0 \ \forall \ i$ are equality constraints for linear functions $h_i(x)$; while the $g_j(x) \le 0$ are convex inequality constraints. With the following three steps, we show that the optimization model described in Section 4.3 is a convex program and thus has a unique solution. First, the optimization model has a convex objective function (9). Second, its constraints (10-18), except constraints (15, 18), are equality constraints. Third, Eq. (15) can be rewritten as $v^{\mathrm{g}}_{i,t} - s^{\mathrm{g}}_{i,t} \times \frac{x_{i,t}}{x_{i,t} + 1} \le 0$. Since $\frac{x_{i,t}}{x_{i,t} + 1}$ is concave and the negative of a concave function is a convex function, Eq. (15) is a convex inequality. Similarly, according to Lemma 1, we can also show Eq. (18) is a convex inequality. Therefore, the above proposed model is a convex mathematical program that has a unique optimal solution. This convexity ensures a number of favorable properties pertain, such as the uniqueness of the optimal solution.

A potential complexity, however, lies in the FIFO (first-in, first-out) property typically assumed for traffic flow networks. FIFO ensures that, on average, traffic entering a network link in any period will exit that link before traffic that arrives in later time periods. The problem is that when dynamic traffic assignment models are extended to multiple-destination networks, the FIFO constraints lead to non-convexity for links that are shared by different paths (Carey,

1992). Furthermore, FIFO constraints dramatically increase the complexity of the mathematical programming model (Lasdon and Luo, 1994).

In our study, FIFO is an issue for the gate-to-yard flows. If the yard destination proportions are known for the gate lanes, the variables $v^{g}_{i,t}$ and $x_{i,t}$ can be extended to $v^{g}_{i,j,t}$ and $x_{i,j,t}$, so the FIFO constraints applied at gate lanes can be formulated as

$$\frac{v^{g}_{i,j,t}}{v^{g}_{i,j',t}} = \frac{x_{i,j,t}}{x_{i,j',t}} \qquad \forall j,t \tag{19}$$

Similarly, the FIFO constraints applied at yard zones can be formulated as

$$\frac{v^{y}_{i,j,t}}{v^{y}_{i,j',t}} = \frac{y_{i,j,t}}{y_{i,j',t}} \qquad \forall j,t \tag{20}$$

If non-identical proportions of multiple yard zone destinations pertain, it is necessary to explicitly incorporate the FIFO constraints. However, if the yard destination proportions are identical over time, as this study has assumed, then the proposed model does not need to consider the FIFO requirement explicitly, as the yard destination ratios on the right hand side of Eqs. (19) and (20) are constants, rather than being dependent on other variables.

5. Solution Stage II: Determine the best time-varying toll pattern

Given the system-optimal flow assignment model, there are two possible ways to implement the demand spreading strategy. In the first approach, the port has truckers obtain arrival time reservations on a quota-limited first-come, first-served basis. If no quota is available for a preferred/requested time window, the trucker is mandatorily assigned a new time window adjacent to the preferred time window. The second approach is to use a tolling mechanism to associate each oversaturated time window with a toll value.

Focusing on the second approach, this section proposes a toll-setting model that calculates the time-varying toll pattern, which adjusts the truckers' non-cooperative arrival time choices so they are consistent with the desired system optimal solution.

5.1. Notations in toll pricing model

The parameters and decision variables used in the toll-setting model are first defined as follows.

Input parameters based on the outputs of the optimization model for truck arrival patterns are:

$d^{*}_{p,\tau}$ = system-optimal truck demand assigned to arrive at time τ with preferred arrival time *p*, which is an output of the truck arrival pattern optimization model

μ_{τ} = average turn times of the trucks arriving during time window τ under system-optimal conditions, which can be estimated using the method to be introduced in Section 5.4.

Other input parameters are:

θ = penalty coefficient associated with arrival time shift

$\eta^{*}_{p,\tau}$ = inconvenience measure of shifting arrival time from *p* to *τ*, which can be calculated by $\theta \times (p-\tau)^2$

Decision variables are:

ω_{τ} = toll price at arrival time window τ

π_{p} = minimum possible total disutility for truckers with preferred arrival time *p*

$c_{p,\tau}$ = total disutility for the truckers who have a preferred arrival time *p*, but actually use time window τ

5.2. Quantifying the toll-based user equilibrium conditions

In this application, truckers will make arrival time decisions based on three factors: (1) the estimated time-dependent turn times, (2) the disutility of shifting their arrival time and (3) the toll they have to pay in the selected time window. By integrating the truck appointment system with congestion-based tolling, the port motives the truckers to alter their arrival time choices and improve the system-wide performance.

Each truck with a preferred arrival time *p* is assumed to have feasible arrival time alternatives near *p* (from time window $p-\beta$ to $p+\beta$). These account for the maximum tolerable arrival time shifts β. For the truckers with a preferred arrival time *p*, the disutility function for option τ can be mathematically written as

$$c_{p,\tau} = \mu_{\tau} + \eta^{*}_{p,\tau} + \omega_{\tau} = \mu_{\tau} + \theta \times (p-\tau)^2 + \omega_{\tau} \tag{21}$$

The user equilibrium principle is extensively accepted as users' behavioral representation of route choice in road networks (Florian and Hearn, 1995). By analogy, we model the truck arrival time choice as a "tolled" user equilibrium problem, in which the disutility of any unused arrival time window is equal to or greater than the disutility of any used

time window, $c^*_{p,\tau}$. It is also assumed that each trucker non-cooperatively seeks to minimize his or her disutility, and has perfect knowledge of the terminal conditions. Extending from the mathematical expressions of gap functions (Smith, 1993; Lu et al., 2009) for dynamic user equilibrium conditions, the above condition can be written for any τ, p:

$$\begin{cases} \text{If } c^*_{p,\tau} > \pi_p \text{ , then } d^*_{p,\tau} = 0 \\ \text{If } c^*_{p,\tau} = \pi_p \text{ , then } d^*_{p,\tau} > 0 \end{cases} \tag{22}$$

Based on the congestion toll set method developed by Hearn and Ramana (1998), the tolled user equilibrium conditions in Eq. (22) can be reformulated as the following linear equations,

$$d^*_{p,\tau} \times \left(\mu_\tau + \theta \times (p-\tau)^2 + \omega_\tau - \pi_p \right) = 0 \qquad \forall p, \tau, \qquad |p-\tau| \le \beta \tag{23}$$

$$\mu_\tau + \theta \times (p-\tau)^2 + \omega_\tau \ge \pi_p \qquad \forall p, \tau, \qquad |p-\tau| \le \beta \tag{24}$$

Further, Eq. (23) reduces to $\mu_\tau + \theta \times (p-\tau)^2 + \omega_\tau - \pi_p = 0$ when the assigned arrival flow is strictly positive, $d^*_{p,\tau} > 0$.

Lemma 3: Eqs. (23, 24) are linear constraints.

Proof: Parameters such as the target arrival pattern $d^*_{p,\tau}$, the average turn times μ_τ, and the arrival time shift penalty $\theta \times (p-\tau)^2$ have been given before setting up Eqs. (23, 24). Only ω_τ and π_p are to be optimized, so Eq. (23) is a linear equation involving ω_τ and π_p Eq. (24) is a linear inequality involving ω_τ and π_p.

5.3. Selecting additional objective function over toll set

Typically, more than one time-varying toll pattern leads to the same set of inflow conditions; and each has a different average toll rate. To derive a desirable time-varying toll pattern, the toll pricing problem can be formulated as a mathematical programming model subject to the tolled user equilibrium conditions expressed as Eqs. (23) and (24). The objective function can be user defined; Hearn and Ramana (1998) have compared several. For example, one objective minimizes the maximum toll at different time windows, expressed as

$$\min\left(\max(\omega_\tau)\right) \tag{25}$$

One can also minimize the number of tolled time windows with positive tolls, written as

$$\min \sum_\tau \chi_\tau \tag{26}$$

where χ_τ is a Boolean indicator for whether a toll is imposed during τ, $\chi_\tau \in \{0,1\}$, and χ_τ =1 if $\omega_\tau > 0$; χ_τ =0 if $\omega_\tau = 0$.

The objective function used in this study minimizes the sum of the toll values $\omega_\tau \ge 0$. This is the same as minimizing the un-weighted-by-volume average toll. The objective is written as:

$$\min \sum_\tau \omega_\tau \tag{27}$$

It is important to note that, since the ω_τ and π_p are the only decision variables to be optimized, the optimization model is a linear program.

Fig. 6 illustrates the proposed toll setting model. The truckers with p = 2 have three possible arrival windows. Suppose the system-optimal solution assigns 5 trucks with preferred arrival time p=2 to τ =1, and leaves the other 15 with τ =2. No trucks are assigned from p=2 to τ = 3. Since the system-optimal solution moves some of the demand to time window τ =1, the total disutility for this time window, π_2, is, $15 + \omega_1 + 10 - \pi_2 = 0$. As there are no trucks assigned to τ = 3, the total disutility for this time window is greater than or equal to π_2.

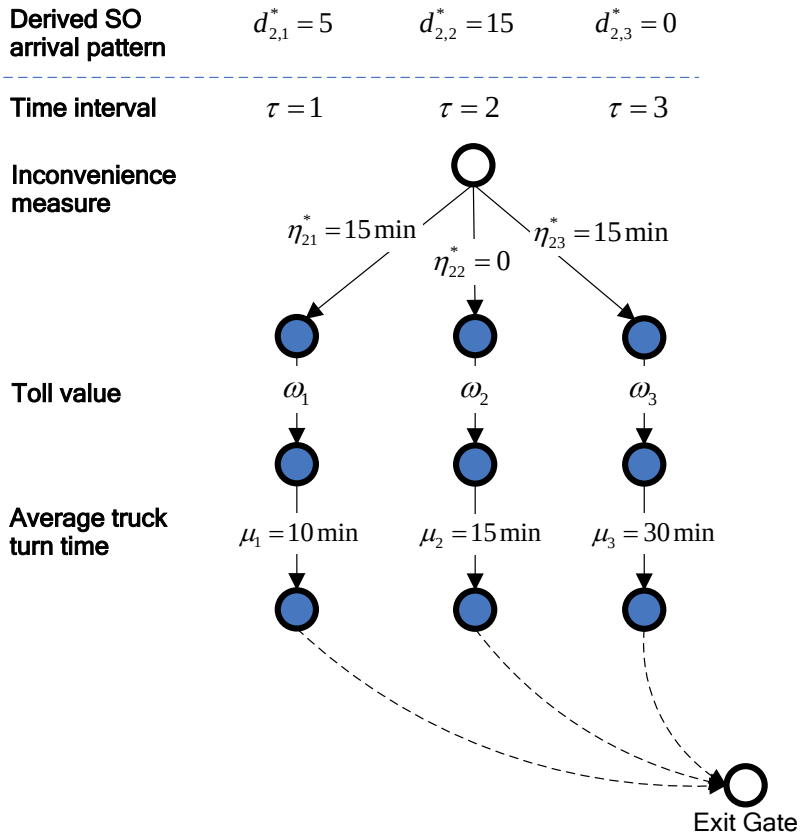


Fig. 6. Network representation of toll pricing model, where results are detailed in Table 1.

Given the input values shown in Fig. 6, where θ is set to 15, we can formulate the toll setting model as follows

$$\min(\omega_1+\omega_2+\omega_3)$$

$$\begin{cases} 15+\omega_1+10-\pi_2=0 & (because \quad d^*_{2,1}>0) \\ 0+\omega_2+15-\pi_2=0 & (because \quad d^*_{2,2}>0) \\ 15+\omega_3+30-\pi_2\geq 0 & (because \quad d^*_{2,3}=0) \end{cases}$$

This linear program can be easily solved, leading to the optimal solution as $\omega_1=0;\quad \omega_2=10;\quad \omega_3=0;\quad \pi_2=25$ with a total system charge of $\omega_1+\omega_2+\omega_3$ =10.

For the above example, we can further use Table 1 to show the interrelated arrival flow pattern and time-varying toll pattern under optimization conditions for preferred arrival time *p*=2.

Table. 1. Arrival flow and time-varying toll pattern under optimal conditions

		Arrival time interval		
		τ=1	τ=2	τ=3
User disutility	Average truck turn time, μ_τ	10	15	30
	Inconvenience due to shifted arrival time, $\theta\times(p-\tau)^2$	15	0	15
	Time-of-day toll, ω_τ	0	10	0
	Total generalized disutility, $c_{p,\tau}=\mu_\tau+\theta\times(p-\tau)^2+\omega_\tau$	10+15+0 =25	15+0+10 =25	30+15+0 =34
	Minimal generalize disutility, $\pi_{p=2}=\min(c_{p,\tau=1},c_{p,\tau=2},c_{p,\tau=3})$	25	25	25
Given input	Derived SO inflow pattern, $d^*_{p,\tau}$	5 (>0)	15 (>0)	0
Optimality condition checking		$c^*_{p,\tau}=\pi_p$	$c^*_{p,\tau}=\pi_p$	$c^*_{p,\tau}>\pi_p$

A feasible but non-optimal time-varying toll pattern is $\omega_1=5;\quad \omega_2=15;\quad \omega_3=0;\quad \pi_2=30$, which results in the same system-optimal solution for the arrival pattern but with a much higher average toll rate: $(\omega_1+\omega_2+\omega_3)/3=20/3$.

Another more complex objective function minimizes the total cost of the tolls paid by the truckers:

$$\min\sum_\tau\left[\left(\sum_p d^*_{p,\tau}\right)\times\omega_\tau\right].$$

5.4. Truck turn time estimation

To construct the tolled user equilibrium model described above, we need to numerically compute the average truck turn times (i.e. the total time of a truck spent in the port) from the system-optimal flow assignment results, because the proposed system-optimal assignment model does not explicitly incorporate truck turn times as variables. A cumulative flow counts-based method can be used, and similar methods can be found in the book chapter by Cassidy (1999) for estimating travel times in queuing systems.

First, based on the computed time-dependent arrival and departure rates by the fluid-based queuing model, the cumulative arrival counts entering the terminal system can be calculated as

$$A(t)=\sum_{t=1}^{t}\sum_i\lambda^{\mathrm{g}}_{i,t}\,, \tag{28}$$

Similarly, the departure counts exiting the terminal system can be formulated as

$$D(t)=\sum_{t=1}^{t}\sum_j v^{\mathrm{y}}_{j,t}\,. \tag{29}$$

Fig. 7 plots these cumulative count curves. Under the FIFO assumption, the horizontal distance between the corresponding arrival departure curves represents the turn time spent by a specific truck, from terminal entry gate (at time t) to exit gate. Mathematically, the turn time of this truck can be expressed as

$$TT(t)=D^{-1}\big(A(t)\big)-t\,, \tag{30}$$

where $D^{-1}(\cdot)$ is the inverse function of $D(t)$.

Under system-optimal conditions, the average turn time μ_τ during arrival time window τ can be computed as the average for multiple arriving trucks:

$$\mu_\tau=\frac{\sum_{t=\tau\cdot\sigma}^{(\tau+1)\cdot\sigma-1}TT(t)}{d_\tau} \tag{31}$$

In the cumulative count graph, the vertical distance between the arrival and departure curves at time t represents the number of trucks in the terminal. One can see in Fig. 7 that the benefit of a truck appointment system is to smooth out the arrival time pattern. Clearly, both turn time and queue length are significantly reduced by doing this.

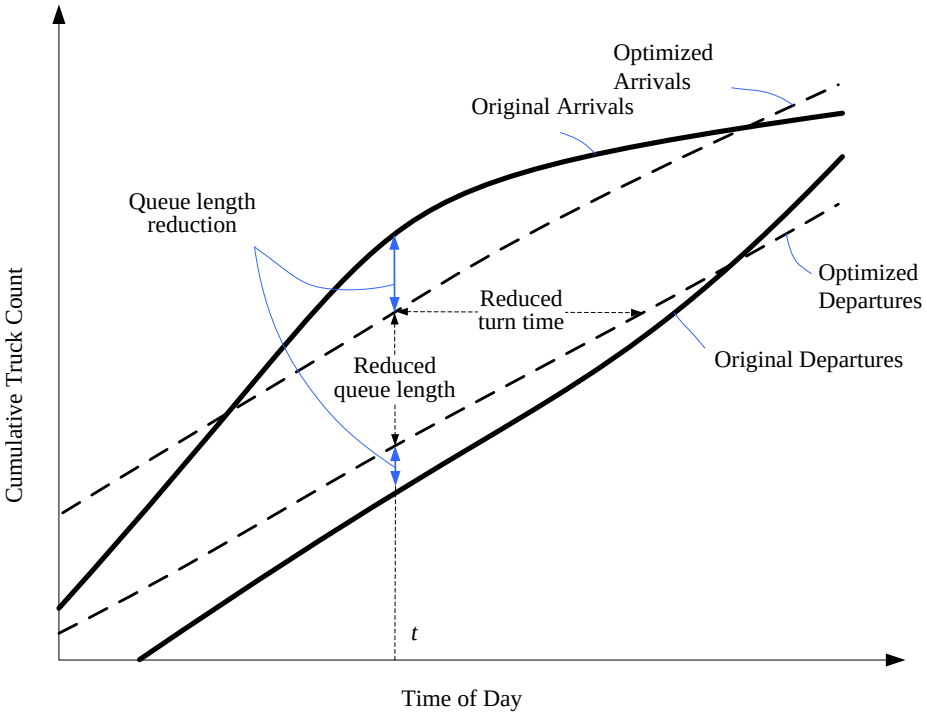


Fig. 7. Illustration of truck turn time analysis approach

6. Numerical examples

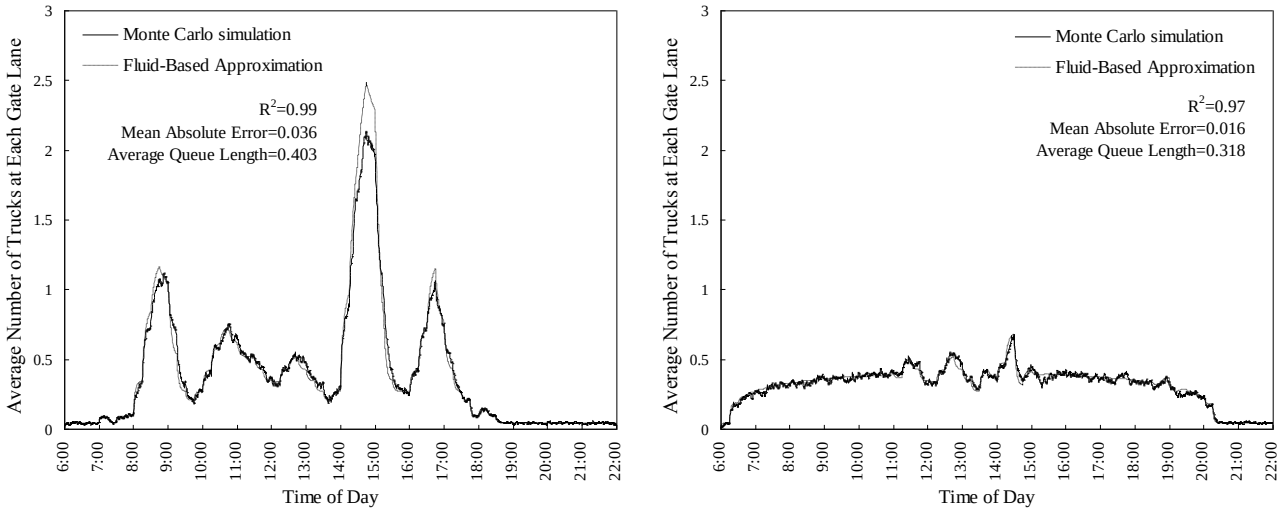
Several numerical experiments have been conducted based on a hypothetical port with 4 gate lanes and 3 yard zones, as shown in Fig. 4. The analysis horizon covers 16 hours (6:00 AM through 22:00 PM). As an input, a hypothetical preferred arrival pattern is adapted from the published time-of-day distribution of truck arrivals at the Port of Los Angeles (FHWA, 2004). A time resolution of 15 minutes is used for the appointment time windows, while 1 minute intervals are used for the fluid-based queuing model. For simplicity, Poisson arrival processes with time-dependent arrival rates are assumed at the terminal gates, and the service duration at each gate lane is assumed to follow an exponential distribution with a mean of 2 minutes. The proportion of trucks at each yard is assumed to be the same. The average service time in each yard is assumed to be 4.5 minutes, and this study tests exponential and normal distributions for yard service times. The maximum tolerable shift in arrival time is set to be ± 2 hours

The truck arrival pattern optimization model is coded and solved using the nonlinear solver, MINOS, in GAMS (Rosenthal, 2008), a high-level modeling system for mathematical programming and optimization. Second-by-second Monte Carlo simulation experiments were used to benchmark the analytical fluid-based model. The simulation code was implemented in Matlab. To reduce the sampling variance, 1000 simulation runs were performed for each scenario.

6.1. Goodness of fit and computational efficiency of fluid-based approximation

Figs. 8 and 9 compare the estimated average numbers of trucks at gates and yards between the fluid-based approximation and the Monte Carlo simulation. The squared Pearson correlation coefficients and Mean Absolute Error values are also presented in these figures. The experimental results indicate that the analytical fluid-based model produces time-dependent queue length estimates consistent with the Monte Carlo simulation model results.

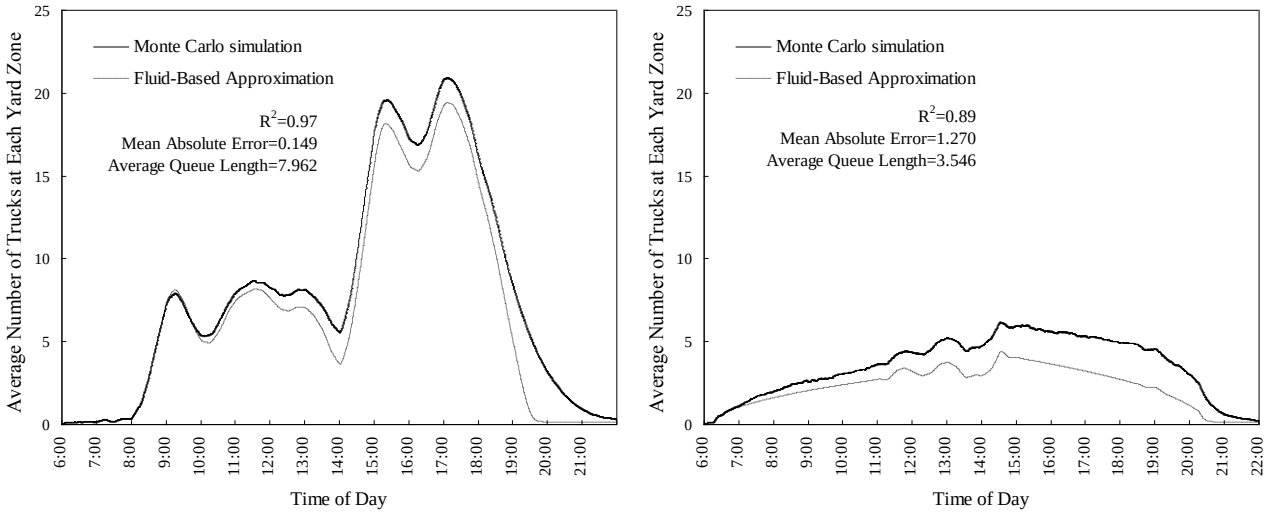
The experiments also numerically test the applicability of the analytical fluid-based model for more general service time distributions. The scenario shown in Fig. 9 (a) and (b) has Poisson arrival process and normally distributed yard service times, by which M/G/1 systems are characterized. The fluid-based model yields reasonable predictions, with mean absolute errors of 0.149 and 1.270 with respect to average queue lengths of 7.962 and 3.546. Although we do not have a mathematical proof of the error bounds, the experimental results indicate that the analytical fluid-based approximation model has a reasonable ability to characterize M/G/1 queuing systems.



(a) Arrival at preferred time windows (b) Arrival at designated time windows

(n **M/M/1** queues at gates with **Normally** distributed yard service times,)

Fig. 8. Comparison of average number of trucks at gate lanes using fluid-based model and Monte Carlo simulation



(a) Arrival at preferred time windows (b) Arrival at designated time windows

(n **M/G/1** queues at yard zones with **Normally** Distributed yard service times)

Fig. 9. Comparison of average number of trucks at yards using fluid-based model and Monte Carlo simulation

Computational efficiency is an important advantage of the analytical fluid-based approximation model. Using Matlab on a machine with an INTEL Core 2 1.83 GHz CPU and 1 GB of memory, the Monte Carlo simulations take 2.8642 seconds for a single simulation run; which means 286 seconds for 100 runs to achieve a desirable confidence interval. In contrast, the fluid-based model only uses about 0.02345 seconds to yield about an equivalent result, dramatically reducing the total computational effort by about 12200:1. This computationally-efficient feature is extremely useful in the iterative optimization algorithm that requires repeated queue estimation. The experiments conducted so far demonstrate that the CPU times needed by the MINOS solver ranged from 11.250 to 29.047 seconds to solve the proposed optimization model in various scenarios (with different average yard utilization rates, or with different arrival and service time distributions).

Another important parameter that affects the size of the optimization problem is the length of the fluid-based queuing model time intervals. Fig. 10 shows the impact of varying the interval length. Clearly, the accuracy of the model with a 1-minute interval is nearly the same as with a 1-second interval. The fluid-based approximation model with 2-minute intervals also yields reasonable results. However, with a 5-minute interval, the model fails to represent the rapid state changes effectively. Thus, a 1-minute interval achieves a sound balance between modeling accuracy and computational efficiency.

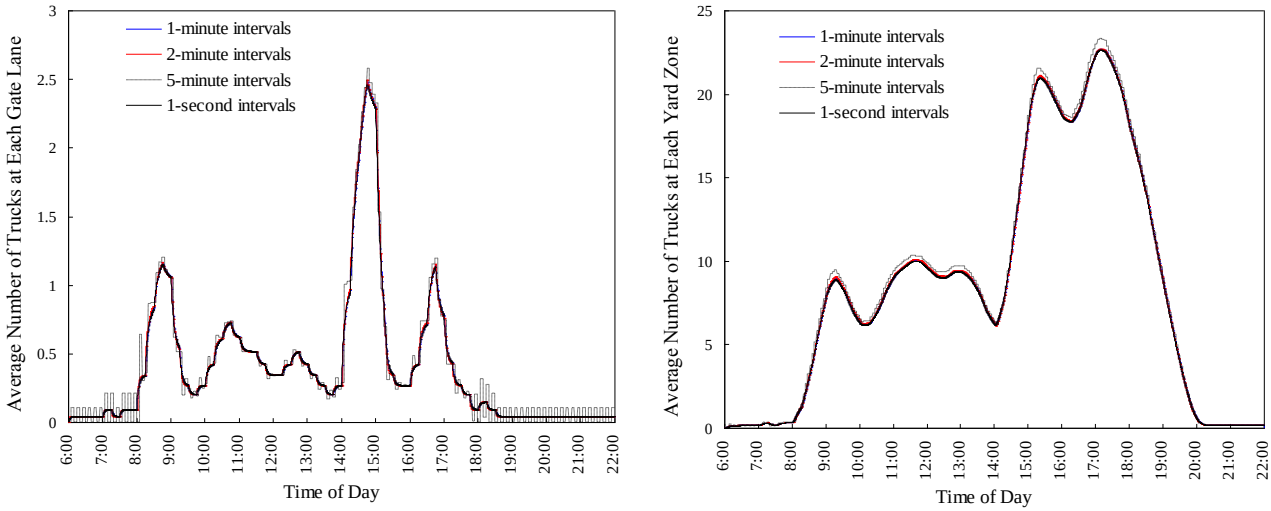
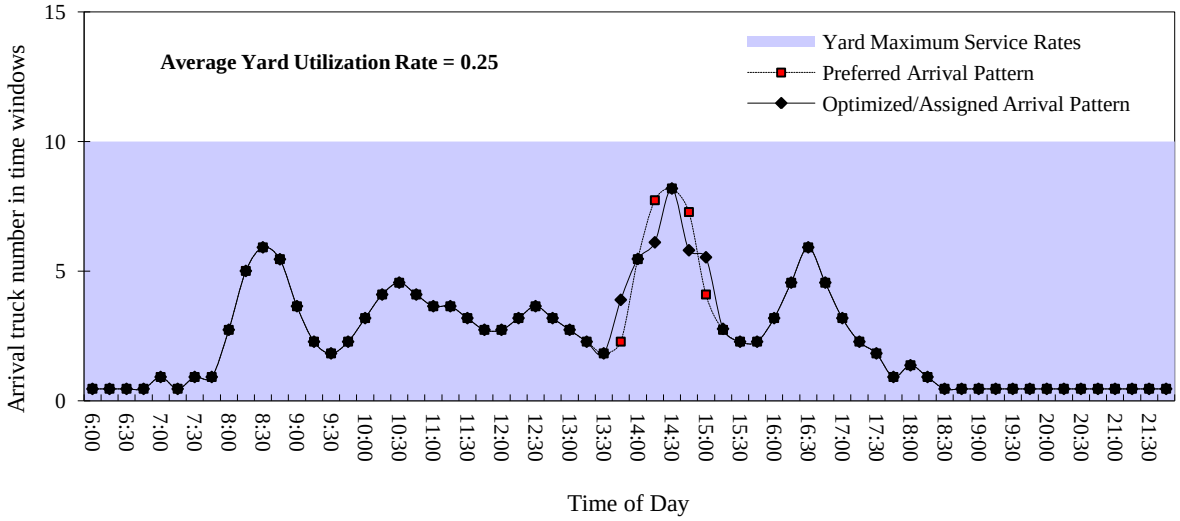


Fig. 10. Impact of fluid-based queuing model interval length on model performance

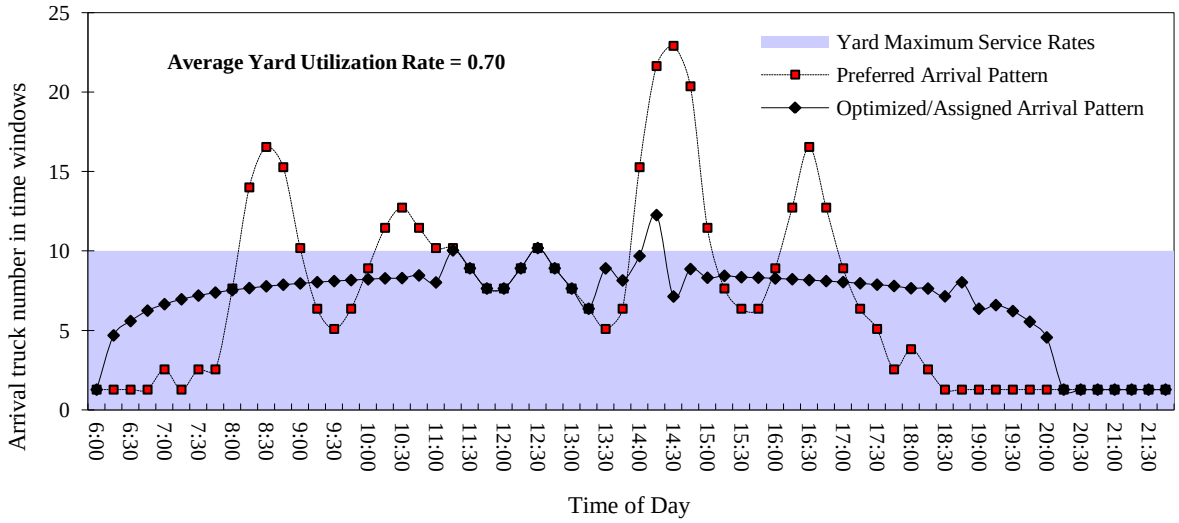
6.2. Stage I solution results: optimized truck arrival pattern

Figs. 8 and 9 also compare the terminal system performance before and after the truck arrival pattern optimization.

As the yard subsystem is typically a bottleneck, the average yard utilization rate over the analysis horizon is included in the sensitivity analysis presented in Figs. 11 and 12. Referring to Fig. 11, if the yard utilization rate is 0.25, which is a low value, the truck arrival pattern optimization model generates relatively small adjustments to the arrival pattern, so most truckers can succeed in making appointments within their preferred arrival time windows. If the yard utilization rate is 0.70, a much higher value, the preferred arrival pattern is further flattened, and more truckers at peak hours have to accept "externally assigned" arrival time windows to reduce the over-saturation and the resulting congestion.



(a) Demand adjustment under a relatively low system load



(b) Demand adjustment under a relatively high system load

Fig. 11. Impact of system load on assigned demand pattern

Fig. 12 demonstrates the system-wide effectiveness of the appointment quota optimization model. Without the appointment system, truckers arrive at their preferred arrival time windows and the average truck turn times increase sharply as the average yard utilization rate increases. In contrast, with the optimized arrival patterns, the average truck turn time increases much more slowly with the increasing system load. In particular, when the average system utilization rate approaches 1, the optimization model is able to reduce the truck turn time dramatically by more than 50% compared to the cases without the appointment system. This optimization strategy will enable the seaport operators to fully utilize the terminal capacity, however, without significant loss in level-of-service.

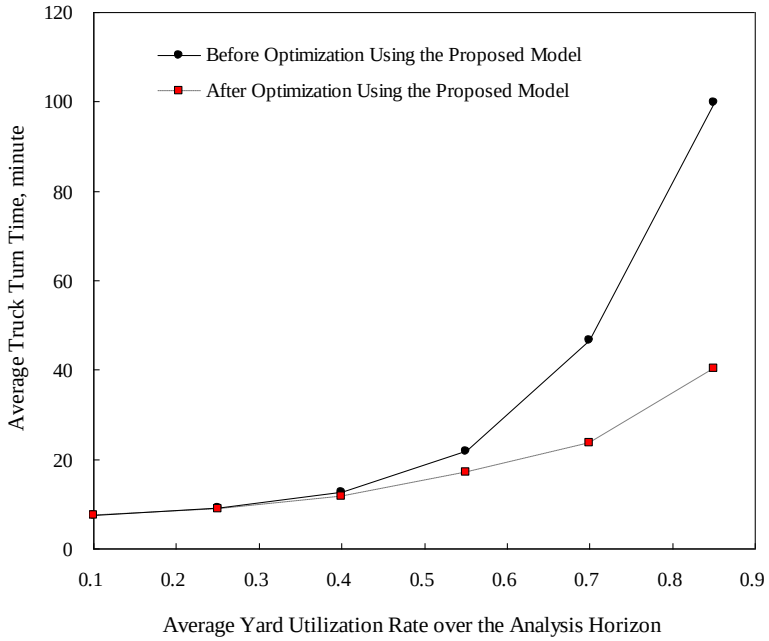


Fig. 12. Effectiveness of truck arrival pattern optimization model

6.3. Stage II solution results: optimized time-varying toll pattern

With the system-optimal demand pattern known, Fig. 13 presents the calculated average turn times given the preferred arrival times, the average turn times under the system-optimal conditions, and the time-varying toll pattern that yields those results. In the given experimental setting, there are two major peak periods: 8:00-9:00 and 14:00-17:00. The optimal time-varying toll pattern pushes the peak-period demands to off-peak periods, including early-morning, noon, and late-night periods. By implementing the time-varying toll pattern, truckers are provided with incentives to adjust their arrival time choices toward the system-optimum conditions.

There is an interesting observation from the experimental results, however. To fully achieve the system optimal condition, the authority might need to set up an extremely high toll in an over-congested time period, and a non-zero

toll even in an uncongested time slot. For example, in order to reserve sufficient capacity for the demands shifting from the highly congested period (e.g. 14:30-14:45), the stair-shaped time-varying toll pattern has a positive toll value at 17:45-18:00 so that it can shift part of demands from 17:45-18:00 to 18:00-18:15. Such a phenomena is more evident under high system loads and tight capacity constraints, and the proposed SO tolling approach in this paper is similar to the first-best toll pricing framework used in road networks, which assume every road or link in a network can be tolled. A more realistic tolling method is to adopt the second-best toll pricing strategy (e.g. Yang and Bell, 1997), which assumes some of roads (i.e. time slots in our model) are not tollable.

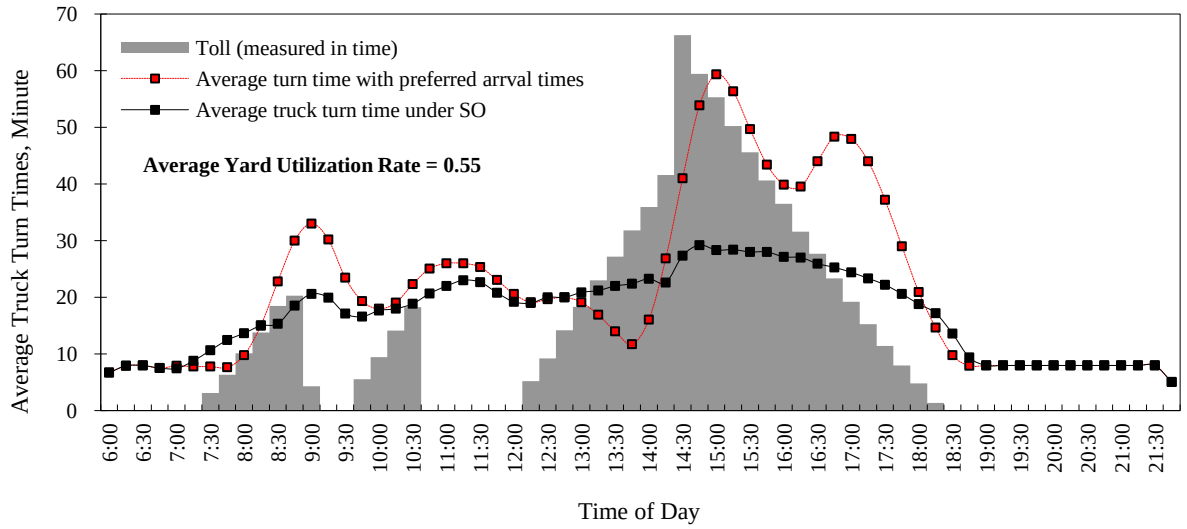


Fig. 13. Calculated time-varying toll pattern and reduced truck turn times

7. Conclusions

The paper has examined the following important theoretical modeling issues: (a) how to characterize time-dependent queuing systems with different service time distributions in a general network flow optimization framework, (b) model convexity and solution uniqueness associated with the point-wise stationary approximation method, and (c) how to derive first-best time-of-day tolling policies using a simple but effective linear programming model that represent truck behavioral responses.

Specifically, this paper makes the following contributions to the literature.

(1) It introduces fluid-based approximation functions to model stochastic time-dependent transportation queuing systems. Compared to the conventional Monte Carlo simulation method, the fluid-based approximation method dramatically improves the computational efficiency while maintaining reasonable approximation accuracy.

(2) The proposed tractable analytical fluid-based approximation functions are seamlessly incorporated into a mathematical programming model to optimize truck arrival flow pattern, and the optimization framework can be extended to improve service plans in a wide range of time-dependent transportation queuing systems, such as toll booths and security checkpoints.

(3) A toll set approach is adapted in this study to select a desirable time-varying toll pattern that leads to the system-optimal truck arrival patterns while minimizing the average toll. The presented toll pricing model considers trucker responses to the terminal conditions and the time-varying toll pattern, and enables the use of time-dependent congestion tolls as an auxiliary tool to improve the effectiveness of port demand management.

To balance trade-offs between computational complexity and model realism, a future study will consider the need to further refine the time-dependent yard zone destination proportions, which will require explicit FIFO constraints. Due to the absence of an analytically tractable relationship between the capacity utilization ratio and the queue lengths for G/G/1 queues, the fluid-based approximation model may have a limited ability to consider more general queuing systems with non-Poisson arrival rates. Future research will also focus on (a) adapting useful analytical results on queuing systems with priorities to address the impact of walk-in trucks, (b) considering time-dependent destination yard zone information and distinguishing pick-up and drop-off trucks, and (c) using the second-best tolling method to establish a more realistic and implementable pricing strategy.

Acknowledgements

This research has benefited from the discussion with Dr. Nathan Huynh at the University of South Carolina. The authors would also like to thank two anonymous referees for their constructive suggestions. The authors are responsible for all the results and opinions expressed in this paper.

References

Arnott, R., Small, K., 1994. The economics of traffic congestion. *American Scientist* 20 (2), pp. 123–127.

Carey, M., 1987. Optimal time-varying flows on congested networks. *Operations Research* 35(1), pp. 58–69.

Carey M., 1992. Nonconvexity of the dynamic assignment problem. *Transportation Research Part B* 26(2), pp. 127-133.

Cassidy, M.J., 1999. Traffic Flow and Capacity, Chapter in Transportation Engineering Handbook. Kluwer Academic Press.

Chow, A. H. F., 2009. Properties of system optimal traffic assignment with departure time choice and its solution method. *Transportation Research Part B* 43(3), pp. 325–344.

FHWA, 2004. Port of Los Angeles Baseline Transportation Study, Cambridge Systematics, Inc. Available online: <http://www.portoflosangeles.org/DOC/REPORT_Draft_Traffic_Baseline.pdf> (accessed 28.09.09.)

FHWA, 2009. FHWA Operations Support - Port Peak Pricing Program Evaluation, Report number: FHWA-HOP-09-014. Cambridge Systematics, Inc. Available online:
< http://ops.fhwa.dot.gov/publications/fhwahop09014/fhwahop09014.pdf > (accessed 28.09.09.)

Florian, M., Hearn, D., 1995. Network equilibrium models and algorithms. In: Ball MO, Magnanti TL, Monma C, Nemhauser GL (eds) Network routing: Handbooks in operations research and management science 8. Elsevier Science, Amsterdam.

Friesz, T.L., Bernstein, D., Smith, T.E., Tobin, R.L., Wie, B.W., 1993. A variational inequality formulation of the dynamic network equilibrium problem. *Operations Research* 41(1), pp. 179–191

Giuliano, G., 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 12(7), pp. 460–473.

Green, L., Kolesar, P., 1991. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science* 37(1), pp. 84–97.

Hearn, D.W., Ramana, M.V. 1998. Solving congestion toll pricing models. In: Equilibrium and Advanced Transportation Modeling, P. Marcotte, S. Nguyen (eds.), Kluwer Academic Publishers, NewYork, pp. 109–124.

Hengsbach, G., Odoni, A. R., 1975. Time-dependent estimates of delays and delay costs at major airports, Report R75-4, MIT Flight Transportation Laboratory, Cambridge, Mass.

Holguín-Veras, J., Ozbay, K., Cerreño,A., 2005. Evaluation Study of Port Authority of New York and New Jersey's Time of Day Pricing Initiative, FHWA/NJ-2005-005. Available online:
< www.rpi.edu/~holguj2/PA/Executive%20Summary.pdf > (accessed 28.09.09.)

Holguín-Veras, J., Wang, Q., Xu, N., Ozbay, K., Cetin, M., Polimeni, J., 2006. The impacts of time of day pricing on the behavior of freight carriers in a congested urban area: Implications to road pricing. *Transportation Research Part A* 40(9), pp. 744–766.

Huynh, N., Hutson, N., 2005. Mining the sources of delay for dray trucks at container terminals. *Transportation Research Record* No. 2066, pp. 41–49.

Huynh, N., 2005. Methodologies for reducing truck turn time at marine container terminals. Ph.D Dissertation, the University of Texas, Austin.

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

Kachani, S., Perakis, G., 2006. Fluid dynamics models and their applications in transportation and pricing. *European Journal of Operational Research* 170(2), pp. 496 – 517

Kim, K. H., Kim, H. B., 2002. The optimal sizing of the storage space and handling facilities for import containers. *Transportation Research Part B* 36(9), pp. 821–835.

Larson, R. C., Odoni, A. R., 1981. Urban Operations Research. Chapter 4 Englewood Cliffs, N.J.: Prentice-Hall. pp.249-251.

Lasdon, L. S., Luo, S., 1994. Computational experiments with a system optimal dynamic traffic assignment model. *Transportation Research Part C* 2(2), pp. 109–127.

Larsson, T., Patriksson, M., 1998. Side constrained traffic equilibrium models–traffic management through link tolls. In: Equilibrium and Advanced Transportation Modelling, P. Marcotte, S. Nguyen (eds.), Kluwer Academic Publishers, New York, pp. 125–151

Lawphongpanich, S., Hearn, D., 2004. An MPEC approach to second-best toll pricing. Mathematical programming, 101, pp. 33-55.

Lu, C-C., Mahmassani, H. S., Zhou, X., 2009. Equivalent gap function-based reformulation and solution algorithm for the dynamic user equilibrium problem. *Transportation Research Part B* 43, pp. 345–364.

Merchant, D.K., Nemhauser. G.L., 1978. A model and an algorithm for the dynamic traffic assignment problem. *Transportation Science* 12, pp 183–199.

Morais, P., Lord, E., 2006. Terminal appointment system study. Transport Canada, Available online: <http://www.tc.gc.ca/tdc/publication/pdf/14500/14570e.pdf> (accessed 28.08.09.)

Nie X J, Zhang H.M., 2005. A comparative study of some macroscopic link models. *Networks and Spatial Economics* 5(1), pp. 89-115.

Noland, R. B., Small, K. A., Koskenoja, P. M., Chu, X., 1998. Simulating travel reliability. *Regional Science and Urban Economics* 28 (5), 535–564.

Rosenthal, R. E., 2008. GAMS User's Guide, GAMS Development Corporation, Washington, DC, USA.
Small K. A., 1982. The scheduling of consumer activities: work trips. *American Economic Review* 72(3), pp. 467-479.
Smith, M.J. 1984. The existence of an equilibrium distribution of arrivals at a single bottleneck. *Transportation Science* 18(4), 385–394.
Smith, M. J., 1993. A new dynamic traffic model and the existence and calculation of dynamic user equilibrium on congested capacity-constrained road networks. Transportation Research Part B 27(1), pp. 49-63.
Stolletz, R., 2008. Approximation of the non-stationary M(t)/M(t)/c(t) queue using stationary queueing models: The stationary backlog-carryover approach. *European Journal of Operational Research* 190(2), pp. 478–493.
Stolletz, R., 2008. Non-stationary delay analysis of runway systems. *OR Spectrum* 30, pp. 191–213.
Sullivan, E.C., El Harake, J., 1998. California route 91 toll lanes: impacts and other observations. *Transportation Research Record* No. 1649, pp. 55–62.
Supernak, J., Kaschade, C., Steffey, D., 2003. Dynamic value pricing on I-15 in San Diego: impact on travel time and its reliability. *Transportation Research Record* 1839, pp. 45–54.
Wang, W., Tipper, D., Banerjee, S, 1996. A simple approximation for modeling non-stationary queues. Proceedings of the IEEE, Fifteenth Annual Joint Conference of the IEEE Computer Societies, Networking the Next Generation: INFOCOM 96 (1), pp. 255–262.
Whitt, W., 1991. The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Science* 37(3), pp. 307–314.
Yang, H., Bell, M.G.H., 1997. Traffic restraint, road pricing and network equilibrium. *Transportation Research Part B* 33 (4), pp.303–314.
Yang, H., Lam,W.H.K., 1996. Optimal road tolls under conditions of queueing and congestion. *Transportation Research Part A* 30(5), pp. 319–332.
Yildirim, M. B., Hearn, D. W., 2005. A first best toll pricing framework for variable demand traffic assignment problems, *Transportation Research Part B* 39(8), pp. 659-678.
Ziliaskopoulos, A. K., 2000. A linear programming model for the single destination system optimum dynamic traffic assignment problem. *Transportation Science* 34(1), pp. 37-49.