TRANSPORTATION SCIENCE
Vol. 40, No. 4, November 2006, pp. 529–539
ISSN 0041-1655 | EISSN 1526-5447 | 06 | 4004 | 0529

DOI 10.1287/trsc.1060.0170

Optimal Slack Time for Schedule-Based Transit Operations

Jiamin Zhao
Oracle Corporation, Redwood Shores, California 94065, jiamin.zhao.zhao@oracle.com

Maged Dessouky
Daniel J. Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, California 90089-0193, maged@usc.edu

Satish Bukkapatnam
School of Industrial Engineering and Management, Oklahoma State University, Stillwater, Oklahoma 74078, bsatish@okstate.edu

To improve service reliability, many transit agencies include significant amounts of slack in the schedule. However, too much slack in the schedule reduces service frequency, given a fixed vehicle fleet size. We study the problem of determining the optimal slack that minimizes the passengers' expected waiting times under schedule-based control. By applying a $D/G/c$ queue model, we show that the system is stable if slack is added in the schedule. For a single-bus loop transit network, we derive convexity of mean and variance of bus delays and provide an exact solution if the travel time is exponentially distributed. For the case of multiple buses and other travel time distributions, we provide several approximation approaches and compare them to simulation results. The simulation results show that our approximations are good for interval of appropriate slack, which often contains the optimal value.

Key words: slack time; transit operation; $D/G/c$ queue
History: Received: February 2003; revision received: January 2006; accepted: April 2006.

Introduction

Many transit systems operate under a schedule-based control strategy. In such a transit system, a bus at a checkpoint stop can be dispatched only at the scheduled departure time, or after the embarkation process if it is behind schedule. That is, no early departures are allowed at the checkpoints. A schedule-based transit system provides a great convenience to those passengers who like to be aware of the schedule. In addition, for those passengers who arrive at stops randomly, their waiting time depends on both mean and variance of headways (Osuna and Newell 1972). An appropriate schedule can reduce the variation of headways significantly while only slightly increasing the average headway.

There has been extensive research on controlling transit vehicles traveling along a single line. In routes providing frequent service (headways of 10 minutes or less), the objective in schedule control is largely to ensure consistency in *headways* (time separation between vehicle arrivals or departures). Customers on short-headway lines typically do not consult schedules before arriving at their stops, and therefore arrival patterns are reasonably stationary relative to the schedule. Second, as demonstrated in Osuna and Newell (1972), average waiting time increases with the square of the coefficient of variation in the headway. In fact, waiting time can be worse than the Poisson case, because vehicles on frequent lines have a tendency to bunch. Headways on very frequent lines are inherently unstable: When a bus falls slightly behind schedule, it tends to pick up more passengers, causing it to slow further, until it eventually bunches with the trailing bus (Newell 1974; Barnett 1974, 1978; Turnquist 1978). This effect can be mitigated, to some degree, by slowing down a trailing bus when it is catching up with the preceding bus. However, the added delay for passengers already on the trailing bus limits the applicability of this and other control strategies, except at the very start of the lines.

The behavior of infrequent lines differs substantially from the behavior of frequent lines. Customers generally do consult schedules, making arrival patterns nonstationary. Therefore, waiting time is not defined by the headway, but instead by the random deviations in the bus arrivals at the stop, along with the customer's selected arrival time relative to the schedule. Finally, because late buses generally do

not pick up additional passengers, schedules tend to be much more stable. As demonstrated in Dessouky et al. (1999), these attributes, combined with slack time inserted in the schedule, lead to schedule stability.

One of the essential tasks for a schedule-based transit system is setting the departure schedules. Naturally, transit planners add slack times when making a schedule. The slack time is the difference between the scheduled and actual expected travel times. The amount of slack can greatly affect the service quality. If the slack time is insufficient, buses are unlikely to be able to catch the schedule when they fall behind, thus deteriorating the service reliability. Alternatively, a large slack time reduces the service frequency, which may also inconvenience passengers. Therefore, determining the optimal slack time that minimizes the passengers' expected waiting time involves a trade-off between the service reliability and the service frequency.

Several researchers have studied schedule-based transit operations. Bowman and Turnquist (1981) developed a model representing passenger arrival times at a bus stop as the outcome of a choice process, which is sensitive to both service frequency and service reliability. The model was constructed based on utility functions and was calibrated by empirical data. Carey (1994) performed a comprehensive study on a schedule-based transit system. He derived a set of integral equations to describe the arrival- and departure-time distributions and formulated the objective function as a combination of cost of travel and cost of deviation from schedule. He pointed out that the optimal schedule problem could then be solved numerically. However, his work was only partly analytical because his main concern was to demonstrate the feasibility of the modeling approach rather than to solve the problem. Dessouky et al. (1999) analyzed empirical data collected by the Los Angeles County–Metropolitan Transit Agency. They observed a slack ratio (slack time to scheduled travel time) of 0.25 on three bus routes converging in downtown Los Angeles; this work was extended by Dessouky et al. (2003) to develop real-time rules for schedule coordination at a terminal.

Although slack is added to the schedule, there is a lack of analytical models in the literature to guide transit planners in setting the appropriate slack levels. In this paper, we develop an analytical model to determine the optimal slack time that minimizes the passengers' expected waiting time. We provide an exact solution for a single-bus loop transit network with exponentially distributed travel time. For the case of multiple buses and other travel time distributions, we provide approximation approaches and compare them to simulation results.

1. Optimal Slack Time Problem for Schedule-Based Transit Systems

The transit system that we consider in this paper consists of a single loop with a single checkpoint. There are N buses repeatedly running along the loop. Schedule-based control is used at the checkpoint where buses either depart on schedule or after the schedule if they arrive late to the checkpoint. Many transit routes, in which buses run from one terminal to another and then return to the original one, can be modeled as a loop with two checkpoints. It can be further simplified to a single loop with a single checkpoint model if we merge the effect of the schedule-based control at the other terminal into the distribution of the round-trip travel time. In §5, we extend the model to multiple checkpoints.

We assume the round-trip travel times $RT_{k,i}$, where the subscripts k and i that represent the kth loop and the ith bus, respectively, are i.i.d. random variables. The dwell times that buses use to load and unload passengers at the stops are incorporated into $RT_{k,i}$. The assumption of i.i.d. for $RT_{k,i}$ is often questioned, especially for short-headway lines and lines with a tight schedule, because a late bus has to spend more time to serve more passengers at stops and hence increases its round-trip travel time. However, for schedule-based services with sufficient slack time, which are the focus of our paper, this assumption is reasonable since vehicle bunching is less likely. In the rest of this paper, we will use RT to represent the random variable associated with round-trip travel time. Let SH denote the scheduled headway (the interval between two consecutive scheduled departure times) and ST be the scheduled round-trip travel time. They have the relationship, $ST = N \cdot SH$. The slack ratio, s_r, is defined as $s_r = ST/(E\{RT\}) - 1$. We assume $s_r > 0$.

First, consider a system having only a single bus. Let D_k and d_k be the scheduled and actual departure times, respectively, at the checkpoint on the kth loop. The delays are defined as $l_k = d_k - D_k$, $k = 1, 2, \ldots$. For a schedule-based transit operation where buses are not allowed to depart early, the values of l_k are nonnegative and the following equation holds:

$$l_{k+1} = \max(l_k + RT_k - ST, 0). \tag{1}$$

Let $E\{l\}$ then be the expected delay when the process reaches a steady state (i.e., the expectation of the delay when k goes to infinity) where l is the delay random variable at stationary. We later show that the process is stable as long as $ST > E\{RT\}$ and this conclusion can be extended to multibus systems (Proposition 1). Let $F(\cdot)$ be the distribution function of l.

We next express the objective function of the passengers' expected waiting time as a function of the first two moments of $F(t)$ and s_r, then we associate our schedule-based transit system with a $D/G/c$

(constant arrival, general service time, multiservers) queue model.

1.1. Objective Function

For a schedule-based transit operation system, we can compute the passengers' expected waiting time in the following way:

Suppose that a passenger arrives at t, at some time between D_k and D_{k+1}. To simplify the problem, we assume that the delay of any bus is less than the scheduled headway; i.e., $D_k + l_k < D_{k+1}$ for all k. In most cases, the interval of appropriate s_r, which contains the optima, is large enough to guarantee that the above assumption is approximately true. Without losing generality, we set $D_k = 0$ and $D_{k+1} = SH$. If the bus to be dispatched at 0 has not arrived yet, i.e., $l_k \geq t$, the waiting time for this passenger is then $l_k - t$; otherwise, he or she has to wait until $SH + l_{k+1}$. Note that l_k and l_{k+1} have the same distribution in a stationary state, and the passenger's expected waiting time can be expressed as

$$\begin{aligned} E\{w \mid t\} &= (E\{l \mid l \geq t\} - t) \cdot P(l \geq t) \\ &\quad + (SH + E\{l\} - t) \cdot P(l < t) \\ &= SH \cdot P(l < t) + E\{l \mid l \geq t\} \\ &\quad \cdot P(l \geq t) + E\{l\} \cdot P(l < t) - t. \end{aligned} \tag{2}$$

Let $f_A(t)$ be the probability density function (p.d.f.) of the passengers' arrival time between 0 and SH, $\int_0^{SH} f_A(t)\,dt = 1$. The overall passengers' expected waiting time is then computed by

$$E\{w\} = \int_0^{SH} E\{w \mid t\} f_A(t)\,dt. \tag{3}$$

If passengers arrive at the stop randomly, which implies that $f_A(t) = SH^{-1}$, $0 \leq t < SH$, and after a few algebraic steps Equation (3) becomes

$$E\{w\} = \frac{1}{2} SH \cdot \left(1 + \frac{2\mathrm{Var}\{l\}}{SH^2}\right). \tag{4}$$

Equation (4) is similar to the formula developed by Osuna and Newell (1972) to compute the expected waiting time for passengers who arrive randomly at the stops:

$$E\{w\} = \frac{1}{2} E\{H\}\left(1 + \frac{\mathrm{Var}\{H\}}{E^2\{H\}}\right), \tag{5}$$

where $E\{H\}$ and $\mathrm{Var}\{H\}$ are the mean and variance of headways, respectively. Note that

$$E\{H\} = SH + E\{l_{k+1}\} - E\{l_k\} = SH \tag{6}$$

$$\mathrm{Var}\{H\} = 2\mathrm{Var}\{l\} - \lim_{k \to \infty} 2\mathrm{Cov}(l_k, l_{k+1}). \tag{7}$$

Table 1 Comparision of Variance of Delays and Variance of Headways

s_r	0.05	0.1	0.15	0.2	0.25
Variance of delays	11.78	3.38	1.15	0.328	0.127
Variance of headways	15.51	5.21	1.91	0.603	0.250
V_H/V_D	1.32	1.54	1.66	1.84	1.97

Equation (4) is an approximation of Equation (5) by ignoring the autocorrelation of the delay process. As we stated before, in most cases the approximation is good enough for the interval of appropriate s_r, which often contains the optima. (We will also verify this point by simulations later in this paper.) Table 1 shows the quality of this approximation ($\mathrm{Var}\{H\} \approx 2\mathrm{Var}\{l\}$) based on simulation results with six buses. (Please refer to §4 for simulation parameters.) It shows that when s_r is greater than 0.15, the assumption is approximately true.

There are other distributions of a passenger's arrival time. For example, for passengers who are aware of the schedule and time their arrivals at the stops, Bowman and Turnquist (1981) derived a p.d.f. of passengers' arrival times as follows:

$$f_A(t) = \frac{\exp(U(t))}{\int_0^{SH} \exp(U(\tau))\,d\tau}, \tag{8}$$

where $U(t) = aE[w(t)]^b$ is the utility of an arrival at time t, a and b are parameters from empirical data, and $E[w(t)]$ is the expected waiting time for an arrival at time t. In this situation, the representation of the passengers' expected waiting time is much more complicated.

In this paper, we will mainly focus on the problem with randomly arriving passengers. The optimization problem is formulated as follows:

$$\begin{aligned} \text{(OP)} \quad \min \ & \frac{1}{2} SH \cdot \left(1 + \frac{2\mathrm{Var}\{l\}}{SH^2}\right) \\ \text{s.t.} \ & SH = (1 + s_r) E\{RT\}/N \\ & l \sim \max(l + RT - (1 + s_r) E\{RT\}, 0), \end{aligned}$$

where $\sim$ means that the two random variables have the same distribution.

Note that although we select random passenger arrival process, which is typical in short-headway situations, to construct our objective function, the schedule-based control model represented by the constraints is applicable to both short- and long-headway situations. Our previous simulation study (Zhao, Bukkapatnam, and Dessouky 2003) showed that under short headway, a proper schedule-based control has similar performance as an even-headway control; to formulate an optimization problem with long headway, one can simply replace the objective function with an appropriate one, for example,

utilizing Equation (8) to construct an objective function. Our assumption of random passenger arrival process here is mainly to simplify the analysis work.

Also note that the objective function does not include in-vehicle passenger time because for a single-loop-single-checkpoint system there is no need to consider in-vehicle passenger time since slack has no impact on the round-trip travel time. However, when there are multiple checkpoints along the route, slacks added at those checkpoints do increase the in-vehicle passenger time.

We will later show that for a single-bus system the variance of delays is a convex function and decreases monotonously as s_r increases (Proposition 2). Then we can easily show that the objective function is convex by showing that the second-order derivative is nonnegative. Taking derivative of the objective function of (OP) provides a closed-form solution for s_r^* (the optimal slack ratio) in terms of $\mathrm{Var}\{l\}$:

$$\frac{1}{2}\frac{E\{RT\}}{N} + \frac{N}{(1+s_r)E\{RT\}}\frac{\partial \mathrm{Var}\{l\}}{\partial s_r} - \frac{N}{(1+s_r)^2 E\{RT\}}\mathrm{Var}\{l\} = 0. \quad (9)$$

Once we can compute $\mathrm{Var}\{l\}$, s_r^* can be obtained by solving the above equation. If $\mathrm{Var}\{l\}$ cannot be calculated analytically, we can still utilize any efficient search algorithm to find the optimal s_r.

Generally, it is difficult to obtain a closed-form solution for $\mathrm{Var}\{l\}$. For the case of $N = 1$ and exponentially distributed round-trip travel time, a closed-form equation for $E\{w\}$ can be obtained. For other cases, we present approximations for $\mathrm{Var}\{l\}$ as a means to find an approximation of the optimal s_r. Before presenting the solution techniques, we show the relationship between our transit system and a $D/G/c$ queue.

1.2. $D/G/c$ Queue Model for Schedule-Based Transit Systems

To demonstrate the connection between our schedule-based transit system and a $D/G/c$ queue model, let us first focus on a simple system having only one bus. Consider an FIFO queue with a single server. Let RT_k and ST be the service time for the kth customer and the interarrival time between two consecutive customers, respectively. The delay in the queue of the $(k+1)$th customer can be described by Equation (1) if l_k is designated to the kth customer's waiting time. To connect the two systems, we can think of the bus as a server and the scheduled departures as customers. The customers' arrival process is deterministic with a fixed interarrival time, ST, and the server spends a random time RT_k to serve the kth customer. From this point of view, our schedule-based transit system with a single bus behaves in the same way as a $D/G/1$ queue. Similarly, for a system having N buses, assuming that a bus can overtake others and that a scheduled departure can be served by any available bus, the system is then equivalent to a $D/G/c$ queue, where the number of servers, c, equals the number of buses, N.

The association of our transit system with a $D/G/c$ queue model enables us to directly utilize many useful results from queueing theory. One of the most important applications is to verify the convergence of the distribution of delays.

For a $GI/G/1$ FIFO queue, Lindley (1952) proved that the queue is stable if and only if the expected service time is less than the expected interarrival time except for the trivial case where both are equal constants. Here, stability means that the distribution of the customers' waiting times converges, i.e.,

$$\lim_{r\to\infty} F_r(t) = F(t) \quad \text{for all } t \geq 0,$$

where $F_r(t)$ is the distribution function of the rth customer's waiting time and $F(t)$ is the limiting distribution. Lindley also showed that the limiting distribution is independent of the initial conditions, i.e., it is unique if the distributions of service time and interarrival time are given. Kiefer and Wolfowitz (1955, 1956) proved that the same stability exists for a $GI/G/c$ FIFO queue when the total service rate is greater than the arrival rate. Applying their results to our transit system, we have the following proposition:

Proposition 1. *For a schedule-based transit system, if $ST > E\{RT\}$, then the system is stable, which means the delay distribution converges, and the limiting distribution is independent of the initial delay distribution, i.e., when $k \to \infty$, $l_k \to l$ for any l_1. If $ST \leq E\{RT\}$ (except the trivial case where $RT \equiv ST$), then the delay tends to infinity with a probability of one.*

For a single-bus system, we prove that the mean and the variance of delays are both convex functions of s_r. Hence, the expected waiting time will be also a convex function. This is an important result because it allows us to use efficient search techniques to find s_r^* that minimizes the expected waiting time.

Proposition 2. *For a single-bus system, the mean and variance of delays decrease monotonously as s_r increases. Besides, they are convex functions of s_r.*

Proof. By utilizing Spitzer's identity,

$$L(s) = E\{e^{-sl}\} = \exp\left\{\sum_{n=1}^{\infty}\left[E\left(e^{-sZ_n^+}\right) - 1\right]/n\right\},$$

where $L(s)$ is the generation function of delay and $Z_n^+ = \max\{0, X_1 + X_2 + \cdots + X_n\}$, $\{X_i\}$ are i.i.d. random

variables and $X_i \sim RT - ST$ for all i. Then,

$$E\{l\} = \lim_{s\to 0} -\frac{\partial L(s)}{\partial s}$$
$$= \lim_{s\to 0}\left\{\sum_{n=1}^{\infty} E\left(Z_n^+ e^{-sZ_n^+}\right)/n\right\}\cdot \exp\left\{\sum_{n=1}^{\infty}\left[E\left(e^{-sZ_n^+}\right)-1\right]/n\right\}$$
$$= \sum_{n=1}^{\infty}\frac{E(Z_n^+)}{n}$$

$$\mathrm{Var}\{l\} = E\{l^2\} - E\{l\}^2 = \lim_{s\to 0}\frac{\partial^2 L(s)}{\partial s^2} - E\{l\}^2 = \sum_{n=1}^{\infty}\frac{E\left(Z_n^{+2}\right)}{n}.$$

Let $G_n(t)$ be the distribution function of

$$\sum_{i=1}^{n} RT_i,$$

where $\{RT_i\}$ are i.i.d. and $RT_i \sim RT$. Let $F_n(t)$ be the distribution function of Z_n, and we have $F_n(t) = G_n(t + n\cdot ST)$. Then

$$\frac{\partial E(Z_n^+)}{\partial ST} = \frac{\partial}{\partial ST}\int_0^{\infty}(1 - F_n(t))\,dt$$
$$= -\int_0^{\infty}\frac{\partial}{\partial ST}G_n(t + n\cdot ST)\,dt$$
$$= -\int_0^{\infty} n\frac{\partial}{\partial t}G_n(t + n\cdot ST)\,dt$$
$$= -n(1 - G_n(n\cdot ST)) \le 0$$
$$\frac{\partial^2 E(Z_n^+)}{\partial ST^2} = n\frac{\partial G_n(n\cdot ST)}{\partial ST} \ge 0$$
$$\frac{\partial E\left(Z_n^{+2}\right)}{\partial ST} = \frac{\partial}{\partial ST}\int_0^{\infty} 2t(1 - F_n(t))\,dt$$
$$= -\int_0^{\infty} 2t\frac{\partial}{\partial ST}G_n(t + n\cdot ST)\,dt$$
$$= -\int_0^{\infty} 2nt\frac{\partial}{\partial t}G_n(t + n\cdot ST)\,dt$$
$$= -2n\cdot E(Z_n^+) \le 0$$
$$\frac{\partial^2 E\left(Z_n^{+2}\right)}{\partial ST^2} = -2n\frac{\partial E(Z_n^+)}{\partial ST} \ge 0. \quad \square$$

For multibus systems, based on intuition and simulation results, we believe that the mean and variance of delays are still convex functions of s_r (see the simulation results section for multibus systems).

For $N = 1$, once the system reaches its steady state, the distribution of delays satisfies the following Wiener-Hopf equation:

$$F(t) = \int_0^{\infty} F(\tau)g(t + ST - \tau)\,d\tau, \quad t \ge 0. \tag{10}$$

For $N > 1$, the integral equation has a more complicated form (Kiefer and Wolfowitz 1955), which is much more intractable for our study. In the next section, we introduce solution techniques for single-bus systems; in §3, we present a method to approximate a multibus system by a single-bus system.

2. Solution Techniques for the Single-Bus Case

In the objective function of (OP), we need to compute the variance of delays. However, in most cases it is difficult to obtain a closed-form solution. We first present the exact solution for the case where $g(t)$ is exponentially distributed. For other distributions of travel times, we present a numerical algorithm and develop some bounds and approximations on the first two moments of $F(t)$.

2.1. Solving the Wiener-Hopf Equation with an Exponential Kernel

An exact method to compute the moments of delays for a $GI/G/1$ queue is solving Equation (10) directly. Explicit solution techniques for a general Wiener-Hopf equation can be found in related references (Davies 1985; Hochstadt 1973; Gohberg and Feldman 1974). Lindly (1952) explained in detail how to solve such an equation for a queue with constant arrivals and a distribution of service time belonging to a particular family

$$\left\{g_n(t) = \frac{\lambda^{n+1}}{n!}t^n e^{-\lambda t}\right\}_{n=0}^{\infty}$$

(the family of nth order convolution of the exponential distribution). Applying this result, if the service time (the round-trip travel time in our system) has an exponential distribution, i.e., $g(t) = \lambda e^{-\lambda(t-t_0)}$, $t \ge t_0$, which has a mean of $1/\lambda + t_0$ and a variance of $1/\lambda^2$, and the interarrival time (the scheduled round-trip travel time, ST, in our system) is constant, then the delay distribution has the following form:

$$F(t) = 1 + C\exp(\mu t), \tag{11}$$

where μ is the nontrivial root of equation

$$\frac{\lambda}{\lambda + \mu} = \exp[\mu(t_0 - ST)], \tag{12}$$

and C is a constant satisfying

$$1 + \frac{\lambda C}{\lambda + \mu} = 0. \tag{13}$$

When $ST > 1/\lambda + t_0 = E\{RT\}$, we have $-\lambda < \mu < 0$ and $-1 < C < 0$. The mean and the variance of delays are

$$E\{l\} = \frac{C}{\mu} = -\frac{1}{\mu} - \frac{1}{\lambda} \tag{14}$$

and

$$\mathrm{Var}\{l\} = \frac{-C(2 + C)}{\mu^2} = \frac{1}{\mu^2} - \frac{1}{\lambda^2} = \frac{2}{\lambda}E\{l\} + E^2\{l\}, \tag{15}$$

respectively.

Equation (12) cannot be solved explicitly. To approximate the solution, we perform first- and second-order Taylor expansions on the right-hand side and solve the corresponding polynomial equations. This gives us

$$\mu \approx -\lambda + \frac{1}{ST - t_0} \quad \text{(first-order approximation)} \quad (16)$$

or

$$\mu \approx -\frac{\lambda}{2} + \frac{1 - \sqrt{(1 + \lambda(ST - t_0)/2)^2 - 2}}{ST - t_0}$$

$$\text{(second-order approximation).} \quad (17)$$

Given an example, where

$$\frac{1}{\lambda} + t_0 = 60 \quad \text{and} \quad \frac{1}{\lambda^2} = 6.4^2,$$

Figure 1 compares the exact solution to the approximations of μ.

For a general distribution of the round-trip travel time, it is hard to find an analytical solution. Moreover, in many situations, the distribution of round-trip travel time can only be empirically represented. Therefore, numerical algorithms or approximations are necessary for those cases.

2.2. Numerical Algorithms

One numerical technique to solve Equation (10) is through recursion. According to Proposition 1, the following recursive equation leads to the limiting distribution of delays no matter what the initial distribution, $F_0(t)$, is

$$F_{k+1}(t) = \int_0^\infty F_k(\tau) g(t + ST - \tau)\, d\tau, \quad t \geq 0, \quad (18)$$

and the convergence rate of the above recursion has a bound in 1-induced-norm of $g(\cdot)$.

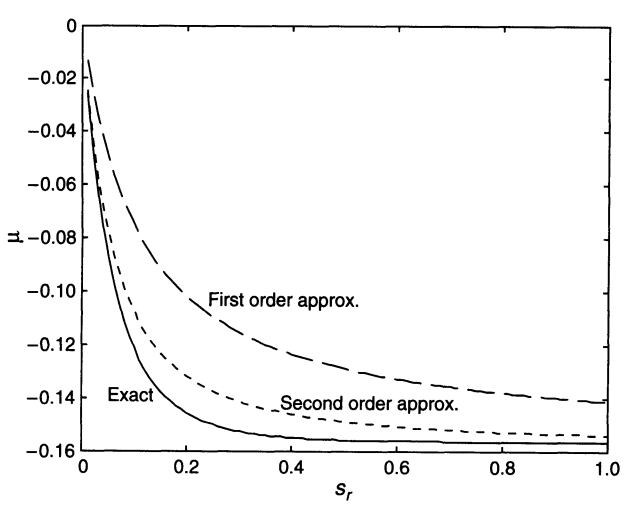

Figure 1 Comparison of Exact Solution and Approximations for Equation (12)

Another way of solving Equation (10) numerically is via solving a Toeplitz equation. Replacing the integral in Equation (10) by a numerical quadrature, we have

$$F_k = \sum_{j=0}^{\infty} F_j g_{k-j}, \quad (19)$$

where $F_i = F(i \cdot \Delta)$, $g_j = \Delta \cdot g(ST + j \cdot \Delta)$, and Δ is the discretization step. Assume $g_j \equiv 0$ for all $j < -n_1$ or $j > n_2$. Also assume that for $n > N$, $F_n \equiv 1$, Equation (10) becomes

$$\begin{bmatrix} F_0 \\ F_1 \\ \vdots \\ F_N \end{bmatrix} = \begin{bmatrix} g_0 & g_{-1} & \cdots & g_{-n_1} & & 0 \\ g_1 & g_0 & g_{-1} & & \ddots & \\ \vdots & g_1 & \ddots & \ddots & & \\ g_{n_2} & & \ddots & \ddots & & g_{-n_1} \\ & \ddots & & & \ddots & \vdots \\ 0 & & g_{n_2} & & \cdots & g_0 \end{bmatrix} \cdot \begin{bmatrix} F_0 \\ F_1 \\ \vdots \\ F_N \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ G_{-n_1} \\ \vdots \\ G_{-2} \\ G_{-1} \end{bmatrix}, \quad (20)$$

where $G_k = \sum_{j=-n_1}^{k} g_j$. Equation (20) is a Toeplitz equation, which can be solved in $O(N \log^2 N)$ time (Brent, Gustavson, and Yun 1980).

2.3. Bounds and Approximations

Many researchers have provided bounds and approximations on the moments of delays for $GI/G/1$ queues. An excellent review of these techniques can be found in Wolff's text (1988). Applying Kingman's bounds on our $D/G/1$ queue, we have

$$E\{l\} \leq \frac{\text{Var}\{RT\}}{2s_r \cdot E\{RT\}} \quad (21)$$

and

$$E\{l\} \geq \frac{E\{(RT - ST)_+^2\}}{2s_r \cdot E\{RT\}}. \quad (22)$$

The lower bound depends on the distribution of the round-trip travel time.

Obtaining bounds on variance of delays is a hard problem, especially for the upper bound. In this paper, we choose Daley, Kreinin, and Trengove's

lower bound (1992), which, for our $D/G/1$ queue, is equal to

$$\mathrm{Var}\{l\} \geq \frac{E\{(RT-ST)^3\}}{3s_r \cdot E\{RT\}} + \frac{E^2\{(RT-ST)^2\}}{4s_r^2 \cdot E^2\{RT\}} + \frac{s_r^2}{12}E^2\{RT\}. \tag{23}$$

Rao and Feldman (2001) provided an upper bound for the queue with NBUE (New Better than Used in Expectation) interarrival times. Note that a $D/G/1$ queue belongs to NBUE queues. Applying their upper bound gives

$$\begin{aligned}\mathrm{Var}\{l\} \leq{}& \frac{E\{(RT-ST)^3\}}{3s_r \cdot E\{RT\}} + \frac{E^2\{(RT-ST)^2\}}{4s_r^2 \cdot E^2\{RT\}} \\ &+ \min\left(2(1+s_r)^2E^2\{RT\}, \frac{(1+s_r)^3}{3s_r}E^2\{RT\}\right) \\ &- \frac{s_r^2}{4}E^2\{RT\}. \end{aligned} \tag{24}$$

The bounds depend on the first three moments on the distribution of the round-trip travel time. For an approximation, we can use the solution or approximations for an exponentially distributed round-trip travel time with equal mean and variance.

3. Approximate a Multibus System by a Single-Bus System

A FIFO queue with multiple servers is usually a much harder problem to solve than a queue with a single server. Therefore, it is preferred to approximate a multibus system by a single-bus system. One approximation is to ignore the case that a bus is caught up by others. Then the system is equivalent to a system having only one bus. This assumption is reasonable only when the headway is relatively large compared with the variation of travel times. When the probability that a bus overtakes others cannot be ignored, the following approximation can be used.

Consider a route having two buses, 1 and 2. Assume that Bus 1 is dispatched at the kth departure from the checkpoint, the situations for the $(k+1)$th and $(k+2)$th departures are summarized in Table 2.

The last case is almost impossible in a real situation, thus it can be ignored. To make an approximation, assume that Bus 1 and Bus 2 are dispatched exactly on time for the kth and $(k+1)$th departures, i.e., $d_k = D_k$ and $d_{k+1} = D_{k+1}$. Then

$$\begin{aligned}P(l_{k+2} \leq t) &= P(\min(D_k + RT_1, D_{k+1} + RT_2) \leq D_{k+2} + t) \\ &= P(\min(RT_1, SH + RT_2) \leq ST + t) \\ &= P(RT_1 \leq ST + t) + P(SH + RT_2 \leq ST + t) \\ &\quad - P(RT_1 \leq ST + t) \cdot P(SH + RT_2 \leq ST + t), \end{aligned} \tag{25}$$

where RT_1 and RT_2 are round-trip travel times for Bus 1 and Bus 2, respectively. Suppose $G(t) = \int_0^t g(\tau)\,d\tau$ is the distribution function of the round-trip travel time. Let

$$\begin{aligned}\hat{G}(t) &= G(t) + G(t-SH) - G(t)G(t-SH) \\ &= 1 - [1-G(t)] \cdot [1-G(t-SH)]. \end{aligned}$$

Equation (25) can be written as

$$P(l_{k+2} \leq t) = \hat{G}(ST + t). \tag{26}$$

Intuitively, we define $\hat{G}(t)$ as the distribution function of a virtual round-trip travel time and let $\hat{g}(t) = d\hat{G}(t)/dt$ be its p.d.f. Then we use a single-bus system with the modified distribution of round-trip travel time to approximate the original multibus system.

Note that the above approximation actually changes the system capacity by reducing the mean of the service time (the round-trip travel time). Besides, it assumes that the headway between two buses is exactly SH. Hence, this approximation does not work well when s_r is very small, but our simulation results show this approximation is good for a reasonably large s_r.

More generally, for a route having N buses, we can define $\hat{G}(t)$ and $\hat{g}(t)$ in a similar way. The general formula for $\hat{G}(t)$ is

$$\hat{G}(t) = 1 - \prod_{i=0}^{N-1} [1 - G(t - i \cdot SH)]. \tag{27}$$

Summarizing what we have so far, we present the following approximation algorithm to address the optimization problem (OP).

ALGORITHM.

Step 1. Approximate a multibus system to a single-bus system by Equation (27).

Step 2. Solve the Wiener-Hopf equation by the numerical algorithm described in §2.2.

Step 3. Solve (OP) by a search algorithm.

Table 2 Effects of Overtaking on Departure Orders

Overtaking between kth and $(k+2)$th departures	Dispatched bus at $(k+1)$th departure	Dispatched bus at $(k+2)$th departure
No overtaking	Bus 2	Bus 1
Bus 1 overtakes Bus 2 once	Bus 1	Bus 2
Bus 2 overtakes Bus 1 once	Bus 2	Bus 2
Bus 1 overtakes Bus 2 twice	Bus 1	Bus 1

4. Simulations

4.1. Simulations for a Single-Bus System

We first simulate the single-bus case to show the quality of the approximations for the mean and

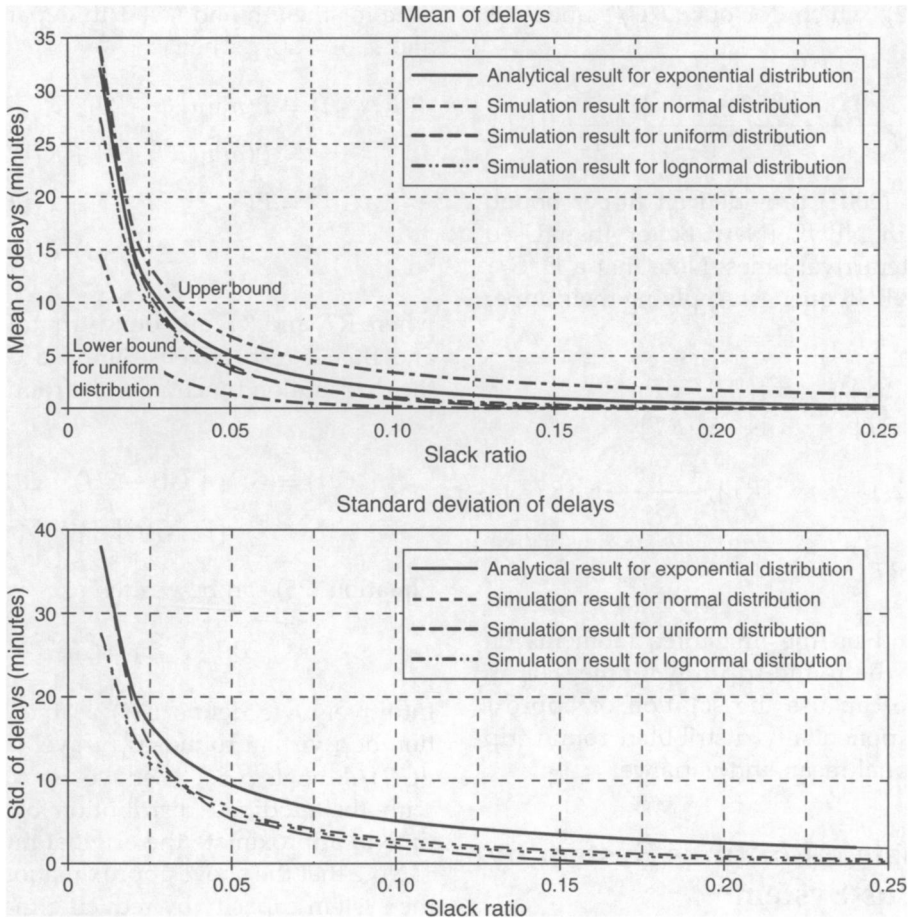

Figure 2 Comparison of Mean and Variance of Delays for a Single-Bus Case

variance of delays. We set the mean and the variance of the round-trip travel time to 60 minutes and (6.4 minutes)^2, respectively. In the simulations, different distributions of the round-trip travel time—including normal, lognormal, uniform, and exponential distributions—are tested. The results are shown in Figure 2.

The solid line represents the analytical solution for the exponential distribution. The other three lines represent simulation results for the normal, uniform, and lognormal distributions, respectively, which are very close to each other. We also plot the upper and lower bounds for the mean of delays. The lower bound, as we mentioned before, depends on the individual distribution. Here we use the lower bound for the uniform distribution to give the rough estimate of how far the bound deviates from the actual curve. We remark that the bounds for the standard deviations were not tight and did not reveal anything meaningful, so we dropped them from the graph.

Figure 2 reveals that the analytical solution for the exponential distribution can be used as a good approximation for other empirical distributions. It also shows that when s_r is less than 0.05, the mean and the variance increase rapidly as s_r decreases. When s_r is greater than 0.1, the mean and the variance have small values and neither changes much. This gives us an insight that s_r^* should be between 0.1 and 0.2. However, the actual optimal range of s_r depends on the actual parameters of a system. For example, a system having a larger variance of the round-trip travel time would require a larger s_r.

4.2. Simulations for a Multibus System

If we add more buses into the route, the delay distributions derived from the analytical model will deviate more from the simulation results, especially when the scheduled headway becomes close to the standard deviation of the round-trip travel time. However, we can use Equation (27) to modify the distribution function of the round-trip travel time to take into account the overtaking of buses. For example, if there are six buses running along the route, the distribution of the round-trip travel time is exponential with mean of 60 minutes and variance of (6.4 minutes)^2, as they were in the previous simulation. Then the mean and the variance of the modified distribution by Equation (27)

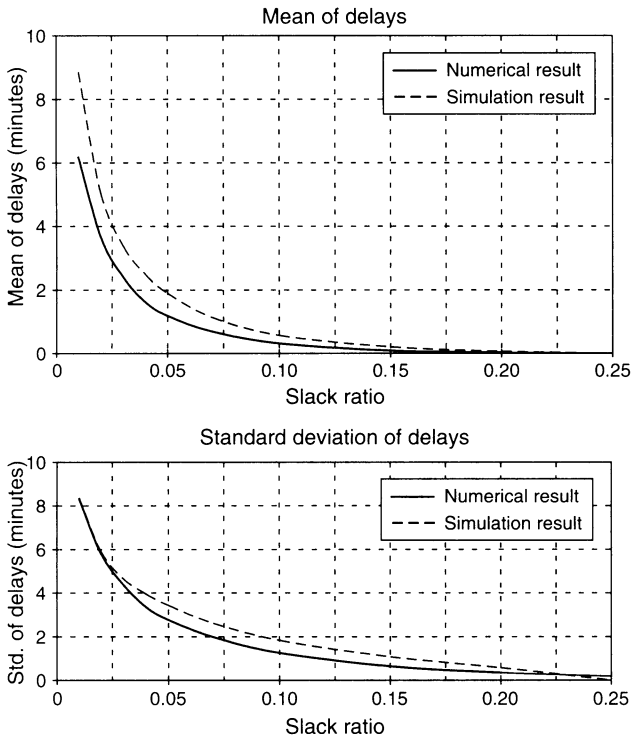

Figure 3 **Comparison of Mean and Variance of Delays for a Multibus Case**

are 59.32 minutes and (4.73 minutes)^2, respectively. Figure 3 plots the simulation result and compares it to the numerical solution using the techniques described in §2.2 with the modified distribution. There are two reasons why it is not a surprise that the numerical solution has smaller values than the simulation result. First, the modified distribution has a smaller mean, which increases the capacity of the corresponding $D/G/c$ queue model. Second, the modified distribution is based on the assumption that every dispatch is on time, which is more violated when s_r is small. Even with these assumptions, the numerical solution is still a good approximation, especially when s_r is greater than 0.1.

4.3. Comparison of Approximation Algorithm and Simulation

In this section, we will test the efficiency of the approximation algorithm presented in §3. For the same settings in the previous simulations, Figure 4 plots both the simulation and approximation results for the passengers' expected waiting time under different s_r. The comparison shows that the approximation procedure overestimates the passengers' expected waiting time when s_r is less than 0.10, but it coincides with the simulation result when s_r is greater than 0.10. For s_r^*, the simulation gives 0.10, while approximation gives 0.11.

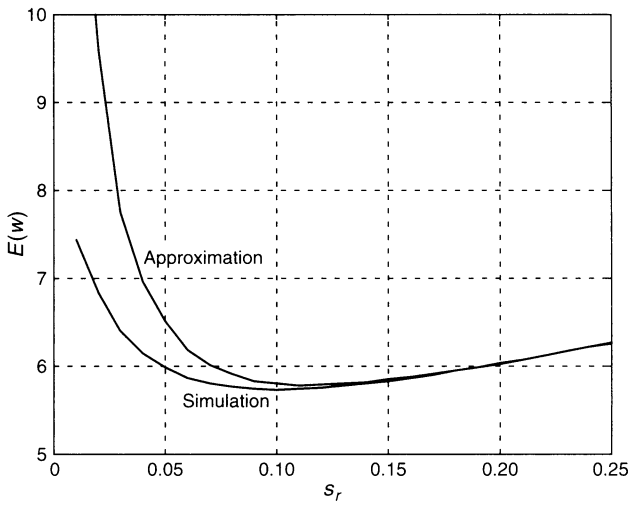

Figure 4 **Passengers' Expected Waiting Times Under Different s_r**

In Figure 5, we compare s_r^* obtained by simulation and our approximation algorithm with different standard deviations of round-trip travel time. The first graph plots the points of optimal s_r versus standard

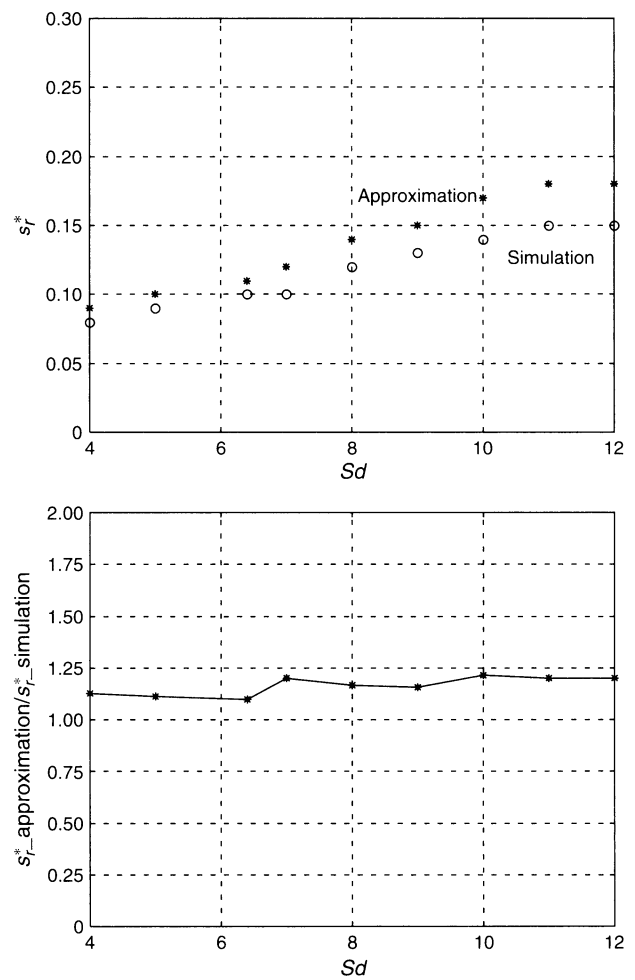

Figure 5 **s_r^* Under Different *Sd* of Round-Trip Travel Time**

deviation of round-trip travel time. The second graph shows the ratio of approximation results to simulation results. As the figure estimates, our approximation algorithm performs reasonably well in identifying the optimal s_r.

5. Extensions to Multiple Checkpoints

In many cases, a schedule-based control is implemented at multiple stops instead of at a single checkpoint. Slack time should only be added to stops where the service punctuality is most critical. This is not only because of operational simplicity, but also for the following reason, illustrated by the following proposition.

PROPOSITION 3. *Consider two control policies: One has a single checkpoint with slack time of st and another adds some checkpoints with slack times* $st_1, st_2, \ldots, st_n$, *which satisfy* $st = st_0 + st_1 + st_2 + \cdots + st_n$, *where* st_0 *is the new slack time for the original checkpoint. Assume that the travel times between any adjacent two checkpoints are independently distributed. Then the second policy has an average delay at the original checkpoint at least as large as the first policy.*

PROOF. Without losing generality, assume there are two Checkpoints 0 and 1. $g_1(t)$ and $g_0(t)$ are p.d.f.s of travel time from Checkpoint 0 to 1 and 1 to 0, respectively. ST_1 and ST_0 are the corresponding scheduled travel time. Let $g(t) = \int_0^\infty g_0(t-\tau)g_1(\tau)\,d\tau$, and $ST = ST_0 + ST_1$. Like Equation (18), the following equations hold if an on-schedule policy is applied to both checkpoints:

$$F_{k+1,0}(x) = \int_0^\infty g_0(x + ST_0 - y)F_{k,1}(y)\,dy \qquad (28)$$

$$F_{k,1}(x) = \int_0^\infty g_1(x + ST_1 - y)F_{k,0}(y)\,dy. \qquad (29)$$

A simple displacement gives

$$F_{k+1,0}(x) = \int_0^\infty g_0(x + ST_0 - y) \cdot \int_0^\infty g_1(y + ST_1 - z)F_{k,0}(z)\,dz\,dy. \qquad (30)$$

Furthermore,

$$F_{k+1,0}(x) = \int_0^\infty F_{k,0}(z) \int_0^\infty g_0(x + ST_0 - y) \cdot g_1(y + ST_1 - z)\,dy\,dz \le \int_0^\infty F_{k,0}(z)g(x + ST - z)\,dz. \quad \square$$

According to Proposition 3, it is economical and efficient to apply on-schedule control only on the most important stops along a route, such as transfer stops and stops having large passenger arrival rates. The optimal slack-time problem with multiple checkpoints can be solved through numerical algorithms similar to those discussed earlier in the paper. Note that, as mentioned earlier, in-vehicle passenger time should be taken into account when there are multiple checkpoints along the route.

6. Conclusions

For system stability, transit planners typically add some slack to the schedule. However, there exists no analytical model to guide transit planners in setting the appropriate amount of slack. In this paper, we have presented an analytical model that addresses the optimal slack time problem for a schedule-based transit operation on a single loop with a single checkpoint. The system is associated with a $D/G/c$ queue model. Results from queueing theory show that the distribution of delays converges to a limiting distribution if a positive slack time is added in the schedule. An analytical solution of distribution of delays can be obtained if there is only a single bus running along the route and the round-trip travel time is exponentially distributed. For general cases, it is difficult to obtain closed-form solutions. We provided some approximation algorithms to solve the general problem. Simulation results verified that the approximations are good.

One topic for future study is the effect of the schedule on the correlation of delays and the correlation of headways. That is, future work could account for travel time dependencies between successive buses. It is believed that a schedule can effectively reduce the correlation, especially when the slack is sufficiently large.

Acknowledgments

The authors acknowledge METRANS for its kind support of this research.

References

Barnett, A. 1974. On controlling randomness in transit operations. *Transportation Sci.* **8** 102–116.

Barnett, A. 1978. Control strategies for transport systems with nonlinear waiting costs. *Transportation Sci.* **12** 119–136.

Brent, R. P., F. G. Gustavson, D. Y. Y. Yun. 1980. Fast solution of Toeplitz systems of equations and computation of Pade approximations. *J. Algorithms* **1**(3) 259–295.

Bowman, L. A., M. A. Turnquist. 1981. Service frequency, schedule reliability and passenger wait times at transit stops. *Transportation Res. Part A* **15** 465–471.

Carey, M. 1994. Reliability of interconnected scheduled services. *Eur. J. Oper. Res.* **79** 51–72.

Daley, D. J., A. YA. Kreinin, C. D. Trengove. 1992. Inequalities concerning the waiting-time in single-server queues: A survey. *Queueing Related Models, Oxford Statist. Sci. Ser.* **9** 177–223.

Davies, B. 1985. *Integral Transforms and Their Applications*, 2nd ed. Springer-Verlag, New York.

Dessouky, M. M., R. W. Hall, A. Nowroozi, K. Mourikas. 1999. Bus dispatching at timed transfer transit stations using bus tracking technology. *Transportation Res. Part C* **7** 187–208.

Dessouky, M. M., R. W. Hall, L. Zhang, A. Singh. 2003. Real-time control of buses for schedule coordination at a terminal. *Transportation Res. Part A* **37** 145–164.

Gohberg, I. C., I. A. Feldman. 1974. *Convolution Equations and Projection Method for Their Solution. Translations of Mathematical Monographs*, Vol. 41. American Mathematical Society, Providence, RI.

Hochstadt, H. 1973. *Integral Equations*. John Wiley & Sons, New York.

Kiefer, J., J. Wolfowitz. 1955. On the theory of queues with many servers. *Trans. Amer. Math. Soc.* **78**(1) 1–18.

Kiefer, J., J. Wolfowitz. 1956. On the characteristics of the general queueing process, with applications to random walks. *Ann. Math. Statist.* **27**(1) 147–161.

Lindley, D. V. 1952. The theory of queues with a single server. *Proc. Cambridge Philosophy Soc.* **48** 277–289.

Newell, G. F. 1974. Control of pairing vehicles on a public transportation route, two vehicles, one control point. *Transportation Sci.* **9** 248–264.

Osuna, E. E., G. F. Newell. 1972. Control strategies for an idealized public transportation system. *Transportation Sci.* **6** 52–72.

Rao, B. V., Richard M. Feldman. 2001. Approximations and bounds for the variance of steady-state waiting times in a GI/G/1 queue. *Oper. Res. Lett.* **28** 51–62.

Turnquist, M. A. 1978. A model for investigating the effects of service frequency and reliability on bus passenger waiting times. *Transportation Res. Record* **663** 70–73.

Wolff, Ronald W. 1988. *Stochastic Modeling and the Theory of Queues*. Prentice Hall, NJ.

Zhao, J., S. Bukkapatnam, M. M. Dessouky. 2003. Distributed architecture for real-time coordination of bus holding in transit networks. *IEEE Trans. Intelligent Transportation Systems* **4** 43–51.