OPTIMAL HOLDING TIMES

AT TRANSFER STATIONS

March 19, 2001

Randolph Hall
Department of Industrial and Systems Engineering
University of Southern California
Los Angeles, CA 90089-0193
email: rhall@mizar.usc.edu
phone: 213-740-4894 fax: 213-740-1120

Maged Dessouky
Department of Industrial and Systems Engineering
University of Southern California
Los Angeles, CA 90089-0193
email: maged@rcf.usc.edu
phone: 213-740-4891 fax: 213-740-1120

Quan Lu
Department of Industrial and Systems Engineering
University of Southern California
Los Angeles, CA 90089-0193

OPTIMAL HOLDING TIMES AT TRANSFER STATIONS

Abstract

Most bus transit systems only offer direct service between a small fraction of the origin/destination pairs that they serve. As a consequence, many travelers must transfer between bus lines to complete their journey. In this paper schedule control policies are created to minimize transfer time under stochastic conditions. We determine how long a bus should be held at a transfer stop in anticipation of the arrival of passengers from connecting bus lines. In general, the total expected waiting time function may have multiple local minima. However, we show there exists at most one non-boundary local minimum point for the special case where arrival times of connecting buses are identically and normally distributed. An application is provided based on bus lines in Los Angeles County's Metropolitan Transportation Agency to evaluate holding time policies.

Keywords: Transportation, transit, dispatching, holding times

1. Introduction

One of the challenges in operating and designing transit systems is providing connectivity to customers with diverse travel patterns. Most bus transit systems can only offer direct service between a small fraction of the origin/destination pairs that they serve. As a consequence, many travelers must transfer between bus lines to complete their journey. These transfers lead to route circuity as well as create transfer delays. In the absence of synchronized schedules, average transfer delays can be quite large. Accounting for randomness in schedules, an average waiting time of 20 minutes or more would not be unusual in systems where schedules are unsynchronized and headways are long (30 minutes or more).

The objective of our paper is to develop methods for optimizing and controlling schedules to minimize these transfer delays. Schedule control will be reflected in the "holding time", which represents releasing a bus later than its scheduled departure time in anticipation of late arrivals of the connecting buses. The amount of time to hold a bus depends on numerous factors, including the number of passengers expected to transfer, number of passengers on board the holding bus, and the lateness of the buses. The holding times will be optimized based on the probability distribution for schedule "lateness", accounting for the likelihood of making a transfer connection and the expected waiting time for the transfer connection.

There has been limited work in developing analytical models that determine the optimal holding times at transfer stations. Hall (1985) examines transfers to and from a rail line, and develops formulas for optimal "safety margins" (i.e., the expected time between arrival of an inbound bus and an outbound train). Abkowitz et al (1987) simulate a variety of dispatching strategies at a timed transfer hub. Their simulation results on two bus lines show that a no holding strategy is best when the bus lines have unequal headways and a double holding strategy

is best when the bus lines have equal headways. Our work differs from Abkowitz et al. (1987) by developing an analytical model instead of a simulation model. Lee and Schonfeld (1991, 1992) simultaneously optimize headways and safety margins at a timed transfer terminal. Our work differs from Lee and Schonfeld (1991, 1992) by optimizing holding times instead of headway. Bookbinder and Desilets (1992) develop network models for schedule construction that account for random transfer delays, but do not focus on real-time schedule control. Knoppers and Muller (1995) find that it is beneficial to coordinate transfers when the variability of the arrival times of connecting buses is low.

A related area of research has been headway control on high-frequency lines. As shown by Osuna and Newell (1971), average waiting time is an increasing function of the headway coefficient-of-variation, for passengers that arrive at stops independently of bus arrivals. Through headway control, variability is reduced, as is average waiting time. This body of work focuses on individual routes, not taking into account transfers and transit centers. For example, Lin et al. (1995) analyze holding and stop skipping strategies under headway based and scheduled based controls. The study reveals that tight stop skipping control significantly increases the average wait time, while the most critical decision variable is the holding time. Other research in the holding problem where transfers are not considered includes the work by Newell (1974), Barnett (1974, 1978), Koffman (1978), and Abkowitz (1986).

The focus of our paper is on bus systems with long headways, for which it is desirable to synchronize arrivals and departures of buses at transfer points. These differ from short headway lines in several important respects: (1) the consequences of missing a connection are greater due to longer waits (hence, synchronization is more important), (2) there is significantly reduced dependency between arrival times of successive buses on a line, and (3) passengers are more

likely to consult a schedule prior to their arrival at stops, reducing dependency between passenger loads and bus lateness.

The remainder of this paper is divided into four sections. In Section 2, we develop analytical models that determine the optimal holding times at transfer stations with general bus arrival time distributions. Section 3 furthers the analysis for special cases based on normal distributions for bus arrival times, which are motivated by data collected for transit operations in Los Angeles County (Dessouky et al., 1999). Section 4 experimentally analyzes the developed holding time models using the probability distribution models developed in Section 3.

2. Holding time models

We consider dispatching strategies associated with a "bank" of buses, scheduled to arrive and depart at a terminal within a narrow time window. Each bank is defined by a set of bus lines, along with their scheduled arrival and departure times. We assume a one-to-one correspondence between arriving and departing buses.

Ideally, the terminal would operate in a manner that minimizes passenger and bus waiting time, and minimizes the number of passengers who miss connections (i.e., buses that arrive later than connecting buses depart). These objectives naturally conflict, in that holding a bus for a late connection may cause more delay for the passengers already on the bus than the time saved for the connecting passengers. To simplify the analysis, the models that follow have a single attribute objective function, which counts all passenger waiting time (either for successful or missed connections) identically.

A bus is considered a*vailable* when all passengers that are present at the terminal have boarded the bus. Suppose that at time 0 a bus is available for dispatch but that another set of

buses (numbered $i = 1,2,...,N$ in order of arrival) are late and have not yet arrived. B represents the number of passengers currently on board, and $M_1, M_2,..., M_N$ represents the number of transferring passengers that will arrive from the late buses, with arrival times $t_1, t_2, ..., t_N$. The problem is then to determine the dispatching time, t_d, for the holding bus that minimizes the total waiting time among passengers. Additionally, let τ be the departure time of the next bus that has the same route of the holding bus. Clearly, the optimal dispatching time for the holding bus will be some time between the current time and the departure time of the next bus on the line, τ. We assume $\tau \geq t_i$ for all i. The problem is then to determine the dispatching time, t_d, for the holding bus that minimizes the total waiting time among passengers, $W(t_d)$.

Comment:

Comment: The formula

Comment:

Comment: The formula

$$\min_{t_d} \; W(t_d) = \min_{t_d} \left(t_d B + \sum_{t_i \leq t_d} (t_d - t_i) M_i + \sum_{t_i > t_d} (\tau - t_i) M_i \right) \qquad (1)$$

In order to keep the model tractable while accounting for delay to the downstream passengers, the parameter B can be multiplied by a factor to account for passengers originating at the stop along with passengers that board at later stops. Future research can focus on developing system-wide models that explicitly account for delay to the downstream passengers.

If the parameters of Eq. (1) were known in advance, the minimum of $W(t_d)$ would occur when t_d equals some t_i ,$i=1,2,...,N$,or t_d equals 0. The derivative of $W(t_d)$ with respect to t_d equals:

$$W'(t_d) = B + \sum_{t_i \leq t_d} M_i > 0, \qquad t_d \neq t_i$$

$W'(t_d)$ is strictly positive for all values of t_d other than t_i, $i=1,2,...,N$. $W(t_d)$ exhibits a discontinuity at t_i, when the function declines by the value $(\tau - t_i)M_i$. The function as a whole

exhibits a saw tooth shape and the optimal solution cannot occur at any time other than at some t_i or 0. Hence, an optimal solution could be found through enumeration of all t_i, $i=1,2,...,N$. It should be noted that the derivative $W'(t_d)$ increases over time, as it equals the number of passengers currently on-board the bus, which naturally accumulates as connecting buses arrive at the stop. Furthermore, the magnitude of the discontinuities tend to decline over time, as the remaining time until the next bus $(\tau - t_i)$ declines. As a consequence, it is eventually preferable to dispatch the bus, as the holding delay imposed on the passengers who have already arrived exceeds the benefit of waiting for the connecting passengers who will be arriving late.

In reality, most of the variables in Eq. (1) are not known with certainty when it is decided whether to hold or dispatch a bus, so we wish to allow a general probability distribution for arrival times, which can be different for each connecting line. The probability distributions for these variables depend on the information available and the quality of travel time forecasts. It may be that these values vary significantly from run to run on a bus line, but they can nevertheless be forecasted accurately due to availability of tracking and communication devices. Using these distributions of arrival times, this section considers a class of strategies in which buses can be held until a pre-determined dispatch time. Such a strategy is likely to approximate an optimal holding strategy. In the first model, buses must be held until this pre-determined time whether or not connecting buses have arrived. Effectively, this amounts to inserting a slack time in the schedule. This assumption is later relaxed.

The holding time is determined from an optimal dispatching time. If the bus arrives prior to the optimal dispatch time, then it is held until this dispatch time. If it arrives after the optimal dispatch time, then it is dispatched immediately. We assume that there is a negligible likelihood that a connecting bus will arrive after time τ, where τ represents the expected time of the next

departure on the controlled bus line (i.e., the line for which the dispatch time is being optimized), and that these times are independent. From our experience, arrival times exceeding τ are indeed rare or non-existent. Let $f_i(t)$ be the probability density function for the arrival time of bus i, which we assume to be continuously differentiable. Hence, by assumption, $f_i(t)=0$, $t \geq \tau$. We also assume that passenger loads (M_i) and bus arrival times (t_i) are mutually independent, as might be expected for lines with long headways.

The objective function is expressed as follows.

$$\min_{t_d} \; W(t_d) = \min_{t_d} \left(t_d E(B) + \sum_{i=1}^{N} E(M_i) \int_{m_i}^{t_d} (t_d - t) f_i(t) dt + \sum_{i=1}^{N} E(M_i) \int_{t_d}^{\infty} (\tau - t) f_i(t) dt \right) \quad (2)$$

where $E(B)$ represents the expected number of continuing passengers on the bus line, and m_i is the minimum possible arrival time for line i. Using ∞ as the boundary for the integral in Eq. 2 simplifies subsequent derivatives. Because the density function equals zero whenever $t \geq \tau$, using ∞ as the boundary does not alter the result.

Comment: I delete the "on line j"

We assume that the waiting time for passengers who originate at the terminal is independent of t_d, as they can adjust their arrival time at the stop to the scheduled departure time. Taking the derivative of Eq. (2) with respect to t_d yields:

$$dW(t_d)/dt_d = E(B) + \sum_{i=1}^{N} E(M_i) F_i(t_d) - \sum_{i=1}^{N} E(M_i)(\tau - t_d) f_i(t_d) \quad (3)$$

The optimal value of t_d either occurs at the boundary ($t_d = 0$), or at one of the points for which Eq. (3) equals zero. In the special case where bus arrival times are identically distributed, the roots for Eq. (3) can be found by solving:

$$(\tau - t_d) f(t_d) = F(t_d) + E(B)/[E(M) * N]$$

where *E(M)* is the expected number of connecting passengers per bus.

Real-time Dispatching: Early Dispatch Allowed

The previous model assumes a bus cannot be dispatched prior to t_d in the event that all connecting buses have already arrived. We now relax this assumption, by dispatching a bus at the minimum of: (1) t_d and (2) the arrival time of the last connecting bus. Eq. (2) is modified to the following:

$$W(t_d) = [t_d E(B) + \sum_{i=1}^{N} E(M_i) \int_{m_i}^{t_d} (t_d - t) f_i(t) dt + \sum_{i=1}^{N} E(M_i) \int_{t_d}^{\infty} (\tau - t) f_i(t) dt] -$$
$$[(t_d - g) - \int_{g}^{t_d} (1 - \prod_{i=1}^{N} F_i(t)) dt][\sum_{i=1}^{N} E(M_i) + E(B)] \quad (4)$$

where the parameter *g* is the minimum *possible* arrival time of any approaching bus: $g = min(m_1, m_2, ..., m_N)$ (i.e., there is 0 probability that a connecting bus will arrive prior to time g). The added bracketed term gives t_d – E{min{t_d, arrival of last bus}}, multiplied by the expected number of passengers on board the bus (i.e., the product is the expected reduction in delay). It represents the expected waiting time savings from dispatching a bus as soon as all connecting buses have arrived.

Taking the derivative of Eq. (4) with respect to t_d yields:

$$dW(t_d)/dt_d = E(B) + \sum_{i=1}^{N} E(M_i) F_i(t_d) - \sum_{i=1}^{N} E(M_i)(\tau - t_d) f_i(t_d) - [\prod_{i=1}^{N} F_i(t_d)][\sum_{i=1}^{N} E(M_i) + E(B)] \quad (5)$$

We next prove that the optimal value of t_d for this model is contained in the interval [t^*, τ) where t^* is the optimal dispatching time for the "do not depart early" case. In cases where $t^* > 0$, the optimal dispatching for this model will be strictly greater than t^*. Note that the optimal dispatching time is greater for this model than the earlier model because the bus can be released earlier if all the buses arrive before the derived time. We first define several variables. Let $W_1(t_d)$ be the waiting time function given in Eq. (2) and $W_2(t_d)$ be the waiting time function given in Eq. (4). Define $\Delta(t_d) = W_1(t_d) - W_2(t_d)$. $\Delta(t_d)$ is a non-decreasing function since the derivative of $\Delta(t_d)$ equals

$$dW_1(t_d)/dt_d - dW_2(t_d)/dt_d = [\prod_{i=1}^{N} F_i(t_d)][\sum_{i=1}^{N} E(M_i) + E(B)]$$

This quantity is non-negative for all values of t_d, as it represents the product of two non-negative quantities.

Proposition 1

The optimal value of t_d in Eq. (4) is greater than or equal to t^*, where t^* is the optimal dispatching time for the "do not depart early case". Furthermore, if $t^* > 0$, then the optimal value of t_d in Eq. (5) will be strictly greater than t^*.

Proof:

Because t^* is optimal for $W_1(t)$, it must be true that $W_1(t) \geq W_1(t^*)$ for any feasible value of t.

Let $\varepsilon = W_1(t') - W_1(t^*) \geq 0$, and suppose that $t' < t^*$. Then the following must be true:

$$W_2(t^*) = W_1(t^*) - \Delta(t^*) < W_1(t^*) + \varepsilon - \Delta(t') = W_2(t')$$

The last statement is true because $\Delta(t)$ is a non-decreasing function (i.e., $\Delta(t') < \Delta(t^*)$) and because of the optimality of t^* for $W_1(t)$, ε must be positive. Therefore, the optimal solution for Eq. (4) must be greater than or equal to t^*.

In cases where $t^* > 0$, it is trivial to show that the optimal value of t_d in Eq. (4) is strictly greater than t^*. This is because the derivative of $W_2(t)$ evaluated at t^* is strictly less than the derivative of $W_1(t)$ and the derivative of $W_1(t)$ evaluated at any $t^* > 0$ must equal zero. This means that $W_2(t)$ continues to decline as t exceeds t^*. $\therefore$

3. Behavior of the waiting time function

Eq. (2) for $W(t_d)$ can, in general, contain multiple local minima. As mentioned earlier, this is clearly true for deterministic arrival times, for which the function exhibits a sawtooth shape and discontinuities. For continuous arrival time distributions, $W(t_d)$ has a smoother shape, yet nevertheless can increase and decrease in association with the arrival distributions for connecting buses. In this section, we examine a special case in which there exists no more than two points for which the derivative of $W(t_d)$ equals zero. Though this special case is not intended to be completely realistic, studying its properties provides insights into the behavior of $W(t_d)$ in the vicinity of forecasted arrival times for connecting busses in actual timed transfer terminals.

The section is divided into two parts. First, we examine probability distributions that have been used for modeling bus lateness. We develop a normal distribution for bus lateness that is derived from earlier work of Dessouky et al. (1999). This model converts a conditional model for lateness on a bus segment into an arrival time distribution at a transfer stop. Second, we examine the properties of $W(t_d)$ for the normal distribution.

Bus Lateness Distributions

Previous research into bus scheduling systems involved analysis of probability distributions of delay, lateness, arrival times or travel times, and covered a wide array of topics. Some papers focused on the shape of the probability distribution (e.g., Talley and Becker, 1987; Guenthner and Hamat, 1988; Strathman and Hopper, 1993). Others developed simulation models to analyze topics such as on-time performance (e.g., Abkowitz et al., 1987; Seneviratne, 1990). While the focus of these papers varies, all include analysis of probability distributions of arrival data. Most of the research has focused on developing distributions for delay or lateness. That is, the earlier studies classified the data into either a delay or a late category and analyzed

only one of these groups. For example, based on a empirical study of bus lines in Paris, Taylor (1982) shows that delay (travel) time distributions are normally distributed. Dessouky et al. (1999) in their study of arrival data from bus service in Los Angeles County model these two random variables, delay and lateness, as dependent variables. That is, the delay distribution on a segment depends on the arrival lateness to the previous stop. For example, they found that for buses with long headways the operator will drive faster than scheduled if behind schedule and vice versa. Unlike short headway lines, passengers who board long headway lines tend to consult schedules prior to arriving at stops (Okrent, 1974; Jolliffe and Hutchinson, 1975; Marguier and Ceder, 1984). Hence, late buses do not board significantly more passengers than on-time buses. Further, since higher headway buses often have slack built into their schedule, there is opportunity to make up for some of the lost time. Dessouky et al. (1999) also noted a non-negligible chance of early arrival, which is contrary to the policy of the transit agency, but nevertheless occurs.

We define lateness and delay by the following relationship:

$$L_k = A_k - S_k$$
$$D_k = L_k - L_{k-1}$$

where:

A_k = the actual arrival time of a bus at stop k

S_k = the scheduled arrival time of a bus at stop k

L_k = the lateness of a bus at stop k

D_k = the delay on the bus segment preceding stop k

Delay represents the deviation from scheduled travel time for an individual line segment whereas lateness represents the cumulative delay among all preceding segments. Alternatively, lateness can be viewed as the deviation from scheduled arrival time at a stop. Delay and lateness can be either negative or positive, though the tendency is toward being positive as bus drivers are penalized for arriving early.

Our interest is in long headway bus lines since these are the most important to synchronize at transfer stops (benefits can be negligible for frequent lines). In a study of LA County bus lines, Dessouky et al. (1999) found that the conditional distribution for delay (D_k) given lateness (L_{k-1}) is well fit by the normal distribution with variance, $VAR(D_k|L_{k-1})$, and expectation:

$$E(D_k \mid L_{k-1}) = a_k + b_k L_{k-1} \qquad (6)$$

The parameter a_k can be interpreted as the expected delay on segment k when the bus is on-time at the previous stop and the slope, b_k, is the correction factor when the bus is not on-time. Based on analysis of arrival data of large headway bus routes in Los Angeles County, Dessouky et al. (1999) found that a_k and b_k were in the order of 0.25 mins and -0.30, respectively. The variance of the conditional distribution was found to be independent of the current lateness and was on the order of 1.5 min^2. For the test data, spacing between schedule checkpoints was on the order of 5 minutes.

A negative value for b_k would indicate that a line is in control, in that a bus that falls behind schedule would run-ahead of schedule on later segments. This would occur on lines with sufficient slack built into the schedule. A positive value of b_k would indicate that a line is out of

control, as might occur on lines with frequent headway, for which passenger boarding depend on the elapsed time since the previous bus.

Whereas Dessouky et al. (1999) developed a *conditional* lateness distribution, the holding time models in Section 2 require *marginal* bus arrival/lateness time distributions. In the appendix, we derive the marginal distributions based on the conditional model through its recursive application over the bus line segments leading up to the transfer stop. We show that if the conditional lateness distribution is normal then the marginal lateness distribution will also be normally distributed.

Behavior of $W(t_d)$ for Normal Distribution

We assume here that the feasible region for the dispatch time is $[0,\tau)$, meaning a dispatch can be no sooner than the present time, and must precede the next departure time. We note that for the normal distribution that there is a finite probability that the arrival time may be outside the feasible region. Hence, similar to before, we assume that $\int_{-\infty}^{0} f_i(t)dt \approx 0$ and $\int_{\tau}^{\infty} f_i(t)dt \approx 0$. We later state conditions on the expected arrival time, variance of arrival time, and τ where the normal distribution may give arrival time probabilities that violate these assumptions.

For identically distributed normal arrival times, we show that under certain conditions there exists at most one non-boundary/local minimum, and that the optimal solution either falls at this local minimum or at the boundary point $t_d = 0$. More importantly, we show how $W(t_d)$ behaves in the vicinity of the forecasted arrival times. We now seek to prove the following proposition:

Proposition 2

If the arrival time probability distributions are identically normally distributed, then $W(t_d)$ will have at most two non-boundary/critical points. If there are two non-boundary critical points, the first will be a local maximum, and the second will be a local minimum. If there is only one critical point, it will be a minimum.

Proof:

Without loss of generality, let *S* represent the expected number of passengers on the bus if all connecting passengers are able to board the bus, and *P* be *E(B)* as a proportion of *S*. Then:

Comment:

$$E(B) = SP \qquad and \qquad \sum_{i=1}^{N} E(M_i) = S(1-P)$$

$$f_i(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad for \quad all \quad i$$

where μ is the mean of the distribution and σ is the standard deviation. Then, we get

$$dW(t)/dt = S(1-P)[P/(1-P) + F(t) - (\tau - t)f(t)]$$

$$= S(1-P)[P/(1-P) + \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx - (\tau - t)\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}]$$

The second derivative of *W(t)* is:

$$dW^2(t)/dt^2 = S(1-P)[2f(t) - (\tau - t)df(t)/dt]$$

$$= S(1-P)\frac{1}{\sqrt{2\pi}\sigma^3} e^{-\frac{(t-\mu)^2}{2\sigma^2}} [2\sigma^2 - (\tau - t)(\mu - t)]$$

The points where the second derivative equal zero (i.e., the inflection points) are defined by:

$$t_1 = \frac{\mu + \tau - \sqrt{(\mu - \tau)^2 + 8\sigma^2}}{2}$$
$$t_2 = \frac{\mu + \tau + \sqrt{(\mu - \tau)^2 + 8\sigma^2}}{2}$$

t_1 is less than or equal to t_2. Furthermore, it can be shown that $t_2 > \tau$, thus falling outside of the feasible region. Ordinarily t_1 falls inside the feasible region, though it can be negative, falling outside the feasible region. For this to occur, $\mu\tau$ must be greater than $2\sigma^2$, which is equivalent to $(\tau/\sigma) < 2(\sigma/\mu)$. For this condition to be satisfied, bus arrival time must be highly variable and/or headway must be very short. We consider these two situations below

Situation 1: First consider the situation when $\mu\tau > 2\sigma^2$, meaning t_1 is feasible. It is easy to verify that

$dW^2(t)/dt^2 < 0$ when $t \in (0, t_1)$
$dW^2(t)/dt^2 > 0$ when $t \in (t_1, \tau)$

Therefore:

I) There is at most one $t \in (t_1, \tau)$ satisfying $dW(t)/dt = 0$. Because $dW^2(t)/dt^2 > 0$ when $t \in (t_1, \tau)$, $dW(t)/dt$ is strictly increasing. Hence, there will be at most one zero point for $dW(t)/dt$, when $t \in (t_1, \tau)$ and, if it exists, it must be a local minimum of $W(t)$.

II) There is at most one $t \in (0, t_1)$ satisfying $dW(t)/dt = 0$. If it exists, it must be a local maximal point of $W(t)$.

From the conditions I and II, we may conclude that there are at most two non-boundary critical points of *W(t)*. If both exist, the first must be a local maximum and the second must be a local minimum.

Situation 2: Consider the situation when $\mu\tau \leq 2\sigma^2$, meaning both inflection points are outside the feasible region. Then $dW^2(t)/dt^2 > 0$ when $t \in [0, \tau)$. With a similar analysis as in the first case, we may find that the function of *W(t)* has at most one non-boundary critical point and, if it exists, it is a minimum. ∴

Figure 1 illustrates the behavior of $W(t_d)$ in four cases representing $\mu\tau > 2\sigma^2$ (situation 1). In Cases 2 and 3 expected waiting time initially increases, but declines as t_d approaches the forecasted arrival time of connecting buses. In Case 2, it is preferable to wait for the connecting buses, rather than dispatch them immediately; in Case 3 it is preferable to do the opposite, with $t_d^* = 0$. Case 1 is the only case where $t_d = 0$ is not a local minimum; rather, expected waiting time declines until reaching the single critical point, after which it increases. In Case 4, $t_d = 0$ is the only local minimum (and also the optimum). The pair of critical points from Case 3 have coalesced into a single inflection point because the benefit of waiting for the forecasted buses is small relative to the added waiting time for customers already on board.

We have not provided figures for situation 2. When $\mu\tau \leq 2\sigma^2$, the assumptions of the model given by Eq. (2) would be violated since the lateness distribution would either provide a non-negligible arrival time after τ, or a non-negligible arrival time prior to 0 (the present time).

In all four cases of situation 1, the optimal policy is to do one of two things: (1) leave immediately, or (2) wait until some time after the forecasted arrival of connecting buses, thus

allowing for the possibility they might be late. However, with additional flexibility in the dispatching policy (such as dispatching as soon as all connecting buses have arrived), then it could be optimal to wait somewhat longer and, in some situations, to hold a bus when it might otherwise be dispatched immediately.

In more general cases, for which multiple lines have differing lateness distributions, each additional bus line, with a unique lateness distribution, can potentially introduce additional critical points, possibly creating an up and down pattern in the objective function in the vicinity of forecasted arrival times. The decision to hold, or not to hold, still depends on assessing whether one of the local minima produces an objective function value that is smaller than $W(0)$. Furthermore, as in the simple cases of Figure 1, the objective function will behave in one of four ways: (1) its derivative will be negative at $t=0$, making it desirable to hold the bus, (2) its derivative will be positive at $t=0$, but it will later decline to a local minimum for which $W(t_d) < W(0)$, (3) its derivative will be positive at $t=0$, and all later minima will be inferior to $W(0)$, or (4) its derivative will be strictly positive for $t \geq 0$.

4. Numerical analysis of examples

We now study the behavior of the waiting time function through a numerical example. We begin with a model based on data collected from Los Angeles County. That is, we adopt the conditional normal model, and use the parameters a =0.25, b = -0.30, and $\delta^2 = 1.5$. We later test the sensitivity of the results with changes in these parameters. For the purpose of illustration, we assume identical distributions on connecting bus lines. Though not entirely realistic, this situation is most useful in illustrating the tradeoffs between holding a bus and dispatching it immediately. We further assume equal spacing between the stops of 2.5 minutes. For example,

if a bus is 5 stops away from the transfer station, its scheduled travel time to the transfer station is 12.5 minutes. The conditional lateness model from the appendix is used to convert these data into an arrival time distribution at the transfer stop. Solving the lateness equations with $k=5$, the expectation and variance of actual arrival time will be 13.2 and 2.85, respectively. As is typical in transit operations, the expected arrival time is later than the scheduled time, because of the strong disincentive against running early.

The other parameters of the holding time model are set as follows. The next departure time of the holding bus, τ, is set to 30 minutes, and

Comment:

$$E(B) = SP = 12.5 \qquad \textit{and} \qquad \sum_{i=1}^{N} E(M_i) = S(1-P) = 12.5$$

Figure 2 plots the total waiting time and the derivative of the waiting time functions based on the distances between incoming connecting buses and the transit station. It illustrates how the shape of $W(t_d)$ evolves through the four cases of Figure 1. If the connecting buses are one stop away (i.e., $k=1$ and scheduled to arrive 2.5 minutes in the future), the waiting time function behaves as in Case 1 with a single local minimum. If the connecting buses are 2 and 4 stops away, the waiting time function behaves as in Case 2 with two non-boundary critical points (maximum and minimum), and local optimal both at $t_d=0$ and a value of t_d greater than 0. It is still preferable to hold the bus, as the second local optimum yields a lower objective value. If buses are 5 to 7 stops away, the function behaves as in Case 3. It still has two non-boundary critical points (maximum and minimum), but dispatching immediately yields the lowest objective. Lastly, for $k = 8$, the function behaves as Case 4; it has no non-boundary critical point, and dispatching immediately is optimal.

Figures 3 to 7 show the sensitivity of the results to changes in different model parameters (a, δ^2, τ, b, P). The plots show the optimal dispatching time as a function of the location of the incoming connecting buses (*k*). Multiple curves are shown to represent different parameter values. For example, Figure 3 shows, for *a=1*, that a bus should be held if incoming buses are five or fewer stops away, and should not be held if incoming buses are further away. The plots also show that the optimal dispatching time increases as the parameter *a* increases, reflecting the fact that the buses will arrive later at the station. However, the point in the route in which it is beneficial to wait at the transfer stop is reached later with larger values of parameter, *a*. Summarizing the results as a function of the other parameters:

- When the connecting buses are fairly close to the transfer stop, the optimal holding time increases as the variance (δ^2) of the conditional lateness distribution increases since it is more likely that the connecting buses will arrive later. When they are further away from the transfer stop, the impact of the higher δ^2 on the optimal holding time is not as much since the coefficient of variability of the arrival time distribution decreases when $b < 0$. Similar to the parameter *a*, the point in the route in which it is beneficial to wait at the transfer stop is reached later with larger values of parameter δ^2.
- As to be expected, the optimal holding time increases as the time until the next departure increases. In this case, the point in which it is optimal to hold is reached sooner since the penalty of missing the connection of higher headway buses is greater.
- When the connecting buses are one stop away (*k=1*), the optimal holding time does not depend on the sloped component of the conditional lateness function (*b*) since the marginal lateness distribution will be independent of this parameter. For larger values of *k*, the

optimal holding time increases with smaller values of the parameter b. The point in the route in which it is beneficial to wait at the transfer stop is reached later with larger values of the parameter b since negative values of this parameter provides the bus with the capability to catch up when it is late.

- Finally, the optimal holding time increases as the fraction of continuing passengers decreases. Furthermore, the point in the route in which it is optimal to hold is reached sooner with more transferring passengers (smaller values of P) since more passengers will miss the connection.

Figure 8 plots the total waiting time and the derivative of the total waiting time functions for the case of four connecting buses where buses can depart early if all the buses have arrived (i.e., Eqs. (4) and (5)). For small values of t_d, the objective function is quite similar to Figure 2, for which buses must wait until a fixed time whether or not all connecting buses have arrived. This is to be expected, because there is a very small chance that all buses would have arrived. On the other hand, the functions are quite different for large values of t_d. Instead of being an increasing function, $W(t_d)$ decreases toward a limiting value. This too is to be expected, because if t_d is sufficiently large, the policy becomes "depart when all connecting buses have arrived". It would rarely be the case that the bus being held would be dispatched any earlier.

Comparing the two holding models, these features result in drastically different values of t_d^*. More importantly, however, the two policies result in little difference in $W(t_d^*)$: the minima of the objectives are nearly the same in both figures. Moreover, in each of the examples shown in the figure, both policies are identical in deciding whether to dispatch the bus immediately, or wait until a later time (hold for $k=1$ *to* 4; dispatch for $k = 5$ or greater). In practice, we expect

the control models would be re-evaluated at frequent intervals on the order of one minute. In this respect, the only decision that really matters is whether to hold or not to hold at each evaluation. This fundamental decision matters more than the holding time: once the bus is dispatched, there is no opportunity for reversal, but a bus that is being held can be released as soon as it is advantageous to do so.

For more general lateness distributions, the objective functions can certainly be different for both policies developed in this paper. For instance, with the second policy (allowing the holding bus to be dispatched when all incoming buses have arrived), the optimal value of t_d does not necessarily fall in one of two critical points. Nevertheless, we expect that model parameters will have similar effects on the optimal values of t_d^* and $W(t_d)$ as discussed in this section.

5. Conclusions

We have examined a class of dispatch policies in which a bus is either dispatched immediately or held until a predetermined time in anticipation of connecting buses. The objective function, representing expected waiting time among all bus riders, can have multiple non-boundary local optima, which occur in the vicinity of forecasted arrival times. Depending on several factors -- number of passengers already on the bus, expected passengers on connecting buses, the time until the next bus departure and the lateness distributions – it can either be optimal to dispatch immediately or hold until one of these local optima. For identically and normally distributed lateness, the objective function can exhibit four different behaviors, as shown in Figure 1. In some situations a local maximum precedes the forecasted arrival time, whereas a local minimum follows the forecasted arrival time.

Based on data collected from transit operations in Los Angeles County, we developed probability distributions for arrival time and travel time that explicitly account for their dependency. As the expected time until bus arrival increases, the expected passenger waiting time function evolves through the four cases of Figure 1. Initially, under Cases 1 and 2, holding the bus is preferred, but when the forecasted lateness is sufficiently large, $W(t_d)$ behaves as in Cases 3 and 4, for which the bus should be dispatched immediately.

More flexible policies offer some potential for improvement. For instance, by allowing buses to depart as soon as all connecting buses have arrived, rather than wait until a pre-determined dispatch time, expected waiting time can be reduced. A consequence of the policy is that the optimal value of t_d increases, though buses will depart prior to t_d if all connecting buses have arrived. Nevertheless, the practical difference between the policy is small. The decision of whether to hold a bus or dispatch it immediately is likely to be the same in most situations, which is the fundamental choice when models are re-evaluated after short intervals. In deciding whether to hold or dispatch a bus, it is certainly possible to create more complex models that explicitly account for more future events. For instance, the policy could be to wait until some subset of connecting buses has arrived. Whether such policies produce substantial reductions in waiting time is a subject for future research.

Acknowledgements

The research reported in this paper was partially supported by Partners for Advanced Transit Highways (PATH), and the California Department of Transportation (CALTRANS). Our appreciation goes to the Los Angeles County Metropolitan Transportation Authority for their cooperation in providing data.

Appendix

A Normal Model for Bus Lateness

Whereas Dessouky et al. (1999) developed a *conditional* lateness distribution, the holding time models in Section 2 require *marginal* bus arrival/lateness time distributions. This section derives the marginal distribution based on the conditional model through its recursive application over the bus line segments leading up to the transfer stop.

Given that $f(D_k \mid L_{k-1})$ is normally distributed with expectation given by Eq. (6) and variance given by $VAR(D_k \mid L_{k-1})$, we now solve for the marginal densities, $f(D_k)$ and $f(L_k)$, the covariance, $COV(D_k, L_k)$, and the correlation, $\rho(D_k, L_k)$. Assume without loss of generality that $L_0 = 0$. Then, $L_1 = D_1$ which by definition is a normal random variable with $E(D_1)=a_1$ and $VAR(L_1) = VAR(D_1)$.

Before we derive $f(D_2)$ and $f(L_2)$, we state some general properties of bivariate normal random variables. If (X,Y) is a bivariate normal random vector, then the random variable $X|Y$ is also a normal random variable where

$$E(X \mid Y) = E(X) + \rho(X,Y)(STD(X)/STD(Y))(Y - E(Y)) \quad (7)$$

$$VAR(X \mid Y) = VAR(X)(1 - \rho(X,Y)^2) \quad (8)$$

Since L_1 and $D_2|L_1$ are normal random variables, $f(D_2, L_1)$ is a bivariate normal probability density based on the above definition. Hence, D_2 and L_2 are also both normal random variables since $L_2=L_1+D_2$. Furthermore, from Eqs. (7) and (8), we get:

$$E(D_2 \mid L_1) = E(D_2) + \rho(D_2, L_1)STD(D_2)/STD(L_1)(L_1 - E(L_1)) \quad (9)$$

$$VAR(D_2 \mid L_1) = VAR(D_2)(1 - \rho(D_2, L_1)^2) \quad (10)$$

where,

$$E(D_2) = E(E(D_2 \mid L_1)) = a_2 + b_2 E(L_1) = a_2 + b_2 a_1$$

We next solve for the variances of the normal probability densities $f(D_2)$ and $f(L_2)$. From Eq. (6), we also have

$$E(D_2 \mid L_1) = a_2 + b_2 L_1 \tag{11}$$

Rearranging Eq. (10), we got

$$VAR(D_2) = VAR(D_2 \mid L_1)/(1 - \rho(D_2, L_1)^2) \tag{12}$$

Grouping the terms in front of the L_1 expression in Eqs. (9) and (11) and substituting into Eq. (12) results in the following quadratic expression.

$$b_2 = \rho(D_2, L_1) STD(D_2)/STD(L_1) = \frac{\rho(D_2, L_1) STD(D_2 \mid L_1)}{\sqrt{1 - \rho(D_2, L_1)^2}\, STD(L_1)} \quad \Rightarrow$$

$$(1 + b_2^2 \frac{VAR(L_1)}{VAR(D_2 \mid L_1)}) \rho(D_2, L_1)^2 = b_2^2 \frac{VAR(L_1)}{VAR(D_2 \mid L_1)}$$

Hence,

$$\rho(D_2, L_1) = \frac{b_2 STD(L_1)}{\sqrt{VAR(D_2 \mid L_1) + b_2^2 VAR(L_1)}} \tag{13}$$

Substituting Eq. (13) into Eq. (12), we get:

$$VAR(D_2) = VAR(D_2 \mid L_1) + b_2^2 VAR(L_1) \tag{14}$$

For the parameters of the density $f(L_2)$, we get

$$E(L_2) = E(D_2) + E(L_1) = a_2 + b_2 a_1 + a_1 \tag{15}$$

$$VAR(L_2) = VAR(D_2) + VAR(L_1) + 2\rho(D_2, L_1)STD(D_2)STD(L_1) \quad (16)$$
$$= VAR(D_2 \mid L_1) + VAR(L_1)(1 + b_2)^2$$

The same iterative argument can be made to show that $f(D_k)$ and $f(L_k)$ are normal probability density functions for any k with the following recursive parameters. We note that these findings are consistent with the empirical study of Taylor (1982) who showed that delay data collected from bus lines in Paris followed a normal distribution.

$$E(D_k) = a_k + b_k E(L_{k-1})$$

$$VAR(D_k) = VAR(D_k \mid L_{k-1}) + b_k^2 VAR(L_{k-1})$$

$$\rho(D_k, L_{k-1}) = \frac{b_k STD(L_{k-1})}{\sqrt{VAR(D_k \mid L_{k-1}) + b_k^2 VAR(L_{k-1})}}$$

Under the assumption that the conditional delay distribution is identical for all line segments (as we showed, this assumption is not required for the random variables D_k and L_k to be normally distributed), the above recursive equations can be expanded as follows. That is, $a_k = a$, $b_k = b$ and $VAR(D_k \mid L_{k-1}) = \delta^2$ for all k.

$$E(D_k) = a + b(a\sum_{j=0}^{k-2}(1+b)^j) \qquad for \quad k \geq 2$$

$$VAR(D_k) = \delta^2(1+b^2\sum_{j=0}^{k-2}(1+b)^{2j}) \qquad for \quad k \geq 2$$

$$COV(D_k, L_{k-1}) = \delta^2 b\sum_{j=0}^{k-2}(1+b)^{2j} \qquad for \quad k \geq 2$$

$$E(L_k) = a\sum_{j=0}^{k-1}(1+b)^j \qquad for \quad k \geq 1$$

$$VAR(L_k) = \delta^2\sum_{j=0}^{k-1}(1+b)^{2j} \qquad for \quad k \geq 1$$

References

Abkowitz, M., Eiger, A., & Engelstein, I. (1986). Optimal control of headway variation on transit routes. *Journal of Advanced Transportation,* 20 (1), 73-88.

Abkowitz, M., Josef, R., Tozzi, J., & Driscoll, M.K. (1987). Operational feasibility of timed transfer in transit systems. *Journal of Transportation Engineering,* 113 (2), 168-177.

Barnett, A. (1974). On controlling randomness in transit operations. *Transportation Science,* 8 (2), 102-116.

Barnett, A. (1978). Control strategies for transport systems with nonlinear waiting costs. *Transportation Science,* 12 (2), 119-136.

Bookbinder, J.H., & Desilets, A. (1992). Transfer optimization in a transit network.. *Transportation Science,* 26 (2), 106-118.

Dessouky, M.M., Hall, R., Nowroozi, A., & Mourikas, K. (1999). Bus dispatching at timed transfer transit stations using bus tracking technology. *Transportation Research,* 7C (4), 187-209.

Guenthner, R.P., & Hamat, K. (1985). Distribution of bus transit on-time performance. *Transportation Research Record,* 1202, 1-8.

Hall, R.W. (1985). Vehicle scheduling at a transportation terminal with random delay en route. *Transportation Science,* 19 (3), 308-320.

Jolliffe, J.K., & Hutchinson, T. P. (1975). A behavioral explanation of the association between bus and passenger arrivals at a bus stop. *Transportation Science,* 9 (3), 248-292.

Knoppers, P., & Muller, T. (1995). Optimized transfer opportunities in public transport. *Transportation Science,* 29 (1), 101-105.

Koffman, D. (1978). A simulation study of alternative real-time bus headway control strategies. *Transportation Research Record,* 663, 41-46.

Lin, G.S., Liang, P., Schonfeld, P., & Larson, R. (1995). Adaptive control of transit operations, U.S. Department of Transportation, Report No. MD-26-7002.

Lee, K.K.T., & Schonfeld, P. (1991). Optimal slack times for timed transfers at a transit terminal. *Journal of Advanced Transportation,* 25 (3), 281-308.

Lee, K.K.T., & Schonfeld, P. (1992). Optimal headway and slack times at multiple route timed-transfer terminals. Transportation Studies Center Working Paper 92-22, University of Maryland, College Park.

Marguier, P.H.J., & Ceder, A. (1984). Passenger waiting strategies for overlapping bus routes. *Transportation Science,* 18 (3), 207-230.

Newell, G.F. (1974). Control of pairing vehicles on a public transportation route, two vehicles, one control point. *Transportation Science,* 8 (3), 248-264.

Okrent, M.M. (1974). Effects of transit service characteristics on passenger waiting time. M. S. Thesis, Northwestern University, Department of Civil Engineering, Evanston, Illinois.

Osuna, E.E., & Newell, G.G. (1972). Control strategies for an idealized public transportation system. *Transportation Science,* 6 (1), 52-72.

Seneviratne, P.N. (1990). Analysis of on-time performance of bus service using simulation. *Journal of Transportation Engineering,* 116 (4), 517-531.

Strathman, J.G., & Hopper, J.R. (1993). Empirical analysis of bus transit on-time performance. *Transportation Research,* 27A (2), 93-100.

Talley, W.K., & Becker, A.J. (1987). On-time performance and the exponential probability distribution. *Transportation Research Record,* 1198, 22-26.

Taylor, M.A.P. (1982). Travel time variability – the case of two public modes. *Transportation Science,* 16 (4), 507-521.