Optimality Conditions for Public Transport Schedules with Timepoint Holding

Peter G. Furth
Department of Civil & Environmental Engineering, 400 SN
Northeastern University
Boston, MA 02115
USA
e-mail: pfurth@coe.neu.edu
telephone: +1 617.373.2447; fax: +1 617.373.4419

Theo H.J. Muller
Delft University of Technology
Faculty of Civil Engineering and Geosciences
Transportation and Planning Section
Postbus 5048
2600 GA Delft
The Netherlands
e-mail: T.H.J.Muller@tudelft.nl
telephone: +31 15.278.5288; fax: +31 15.278.3179

Presented at CASPT 2006 (Leeds, UK, June 21-23, 2006)

Submitted for publication in written form on November 16, 2006; revised Aug, 2008

Published in Public Transp (2009) 1:87-102

Abstract

Holding buses to scheduled departure time at timepoints involves a tradeoff between reliability and speed, with impacts on user and operating cost. Two new measures of user cost, excess waiting time and potential travel time, are proposed. They relate to the early extreme of a bus's departure time distribution from a passenger's origin stop, and the late extreme of a bus's arrival time distribution at the destination stop. A route with long headway service is modeled assuming that segment running times are independently distributed. Operating impacts of unreliability are captured by requiring enough recovery time that delay does not systematically grow with each cycle. Based on an objective of minimizing a sum of operating cost and user costs, optimality conditions are derived for the strictness of a timepoint and for dispatching reliability at the terminal, which are related to the amount of slack within the running time schedule and within the scheduled layover. It is shown that a timepoint's optimal strictness (probability of holding) increases with the demand for boardings at the timepoint, with the effect diminishing as stops become farther from the start of the route; however, welfare benefits compared to using a uniform percentage of slack across the route may be small. It is also shown that there is no universally optimal dispatch reliability; the more slack is built into the running time schedule, the less reliable should be the dispatch from the terminal. Up to a point, as scheduled running time increases, the optimal recovery time decreases, and slack time spent holding en route substitutes one-to-one for slack time spent holding at the terminal, so that holding at timepoints does not necessarily increase operating cost.

In order to improve service reliability, bus route schedules have timepoints – intermediate points at which early buses are held until a scheduled departure time. Increasing the amount of slack time in a running time schedule will improve reliability, but hurt operating speed, affecting both riding time and operating cost. By modeling this tradeoff, one can determine optimal running time schedules (or equivalently, optimal scheduled departure times at timepoints). This work is focused on long headway routes, for which passengers time their arrivals to meet a targeted scheduled trip.

Timepoint scheduling and holding to schedule are routine practice with bus operations in many cities with most passenger train operations. However, practice in timepoint scheduling tends to be based on rules of thumb and professional judgment. Most of the literature on holding buses deals with short-headway routes, with passengers arriving independently of schedule, with an emphasis on the feedback mechanism that occurs as late buses become ever later due to the greater number of passengers waiting to board them [1, 2]. The only unreliability impact they consider is its effect on average waiting time. Timepoint scheduling for longer headway transit routes has seen relatively little attention in the literature; some recent papers are [3, 4, 5]. Our work advances the subject by using a further developed user cost function that captures the impact of unreliability on passengers, and by accounting for how unreliability propagates from one cycle to another as in [5].

User Impacts of Reliability

To model the tradeoff of reliability against speed, the first need is for a cost function that accounts for the impacts of unreliability on users. Traditional measures of reliability on long headway routes – percentage of trips that are early / on-time / late, or standard deviation of schedule deviation – are measures of operational quality, not measures of user impact. Extending our work on waiting time reliability [6], we propose two user impacts of unreliability: excess waiting time, and potential travel time (also called buffer travel time), which are related to extremes of bus arrival time and departure distributions.

User impacts related to reliability and holding are illustrated in Figure 1. The top half of the figure shows the distribution of a bus's departure time from a stop *i* and its arrival time at a stop *j* on a route without timepoint control. We treat segment running time as including dwell time at a segment's end stop, so that if there is no holding, departure time equals arrival time. The travel time of a passenger going from *i* to *j* can be divided into four components. We use the following notation:

Dep_i = departure time at stop *i*

$DepMean_i$, $Dep02_i$ = mean and 2-percentile departure time at stop *i*

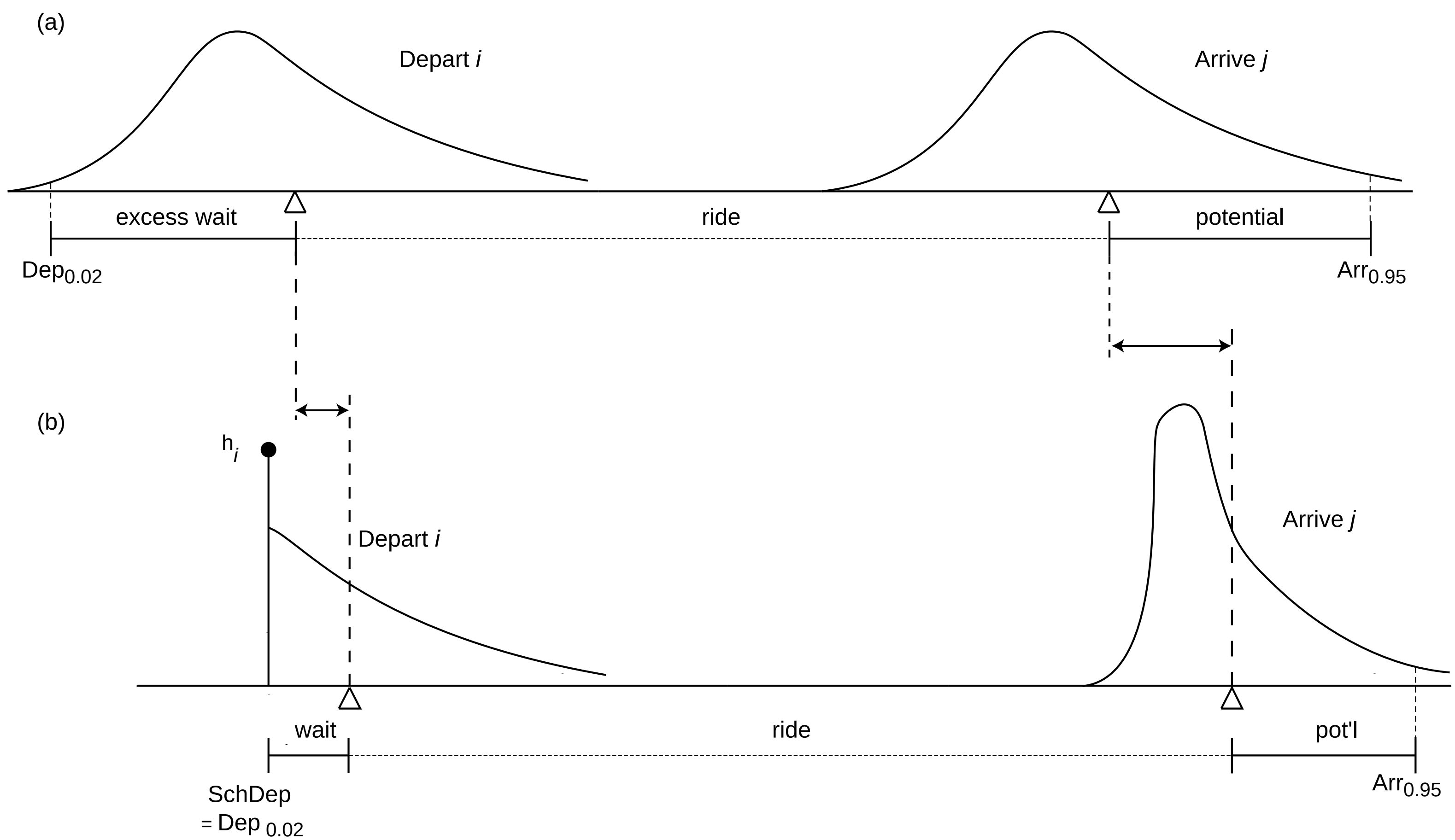

Figure 1. Impact of holding on departure and arrival time distributions and on user cost components. (a) Without timepoint control; (b) with timepoint control.

$DepSch_i$ = scheduled departure time at stop i

Arr_i = arrival time at stop i

$ArrMean_i$, $Arr95_i$ = mean and 95-percentile arrival time at stop i

Ons_i, $Offs_i$ = boardings and alightings per trip at stop i

C = scheduled cycle time

h_i = P[bus is held at stop i; also called the strictness of timepoint i

The first part of a passenger journey, not shown in the figure, includes their access time to reach the stop, plus the time spent between their arrival and $Dep02_i$, the 2-percentile departure time at i. As we argue in [6], passengers using long headway routes may be assumed to arrive before $Dep02_i$ in order to limit the chance that they miss the bus. Doing so internalizes the cost of occasional having to wait a long time for a missed bus. This part of a passenger's waiting time is omitted from the cost function because it is unavoidable, related mainly to uncertainty in access time and to the planned headway.

The second part of a passenger journey is *excess waiting time*, running from $Dep02_i$ to Dep_i, with average length $DepMean_i – Dep02_i$, which corresponds to the early tail of the departure time distribution. It is called *excess* waiting time because if service were perfectly reliable, it would be zero. The next part of a passenger journey is riding time, running from departure at i to arrival at j, with average length $ArrMean_j – DepMean_i$. The final part is related to how much time passengers have to budget for their journey. We assume that passengers, wanting to limit the probability of arriving late at their destination, budget for the 95-percentile arrival time. *Potential travel time* is the difference between their actual arrival time and their budgeted arrival time; its average length is therefore $Arr95_j – ArrMean_j$, which corresponds to the late tail of the arrival time distribution. While passengers are not actually traveling during this time, it still represents a cost to them, since time budgeted for travel, even when redeemed because the bus doesn't arrive that late, cannot be used as freely as if it were not thus encumbered.

The bottom half of Figure 1 shows how user impacts change when timepoint holding is applied at stop *i* and at intermediate stops between *i* and *j*. The departure distribution at *i* has a probability spike h_i, representing the early tail of arrival time distribution that is truncated by holding; excess waiting time is therefore reduced dramatically. Holding applied systematically along a route will tend to increase $ArrMean_j$ increases more than $DepMean_i$, resulting in an increase in average riding time. On the other hand, holding tends to increase $Arr95_j$ by a much smaller amount than it increases $ArrMean_j$, reducing potential travel time, because buses arriving very late at *j* are likely to be buses that were already late at earlier timepoints and therefore not affected by holding.

Cost Function

The objective is

$$\textit{minimize} \text{ Societal Cost} = \text{User Cost} + \text{Operating Cost} \tag{1}$$

User cost is the sum of the three components described earlier:

$$\text{Excess waiting cost} = uWait * \sum_j Ons_j * (DepMean_j - Dep02_j) \tag{2a}$$

$$\text{Riding cost} = uRide * \sum_j (Offs_j * ArrMean_j - Ons_j * DepMean_j) \tag{2b}$$

$$\text{Potential travel time cost} = uPot * \sum_j Offs_j * (Arr95_j - ArrMean_j) \tag{2c}$$

Unit costs *uWait*, *uRide*, and *uPot* should reflect traveler utility. As discussed in [6], a plausible set of relative values that is consistent with demand modeling research is 1.5 : 1.0 : 0.75, respectively.

The three user cost components shown in Eqs. 2(a)-(c) depend on four summary measures of the departure and arrival time distributions: $Dep02_j$, $DepMean_j$, $ArrMean_j$, and $Arr95_j$. The user cost function can be rearranged as a sum of components related to these four measures:

User cost =

$$- uWait * \sum_j (Ons_j * Dep02_j)$$

$$+ (uWait - uRide) * \sum_j (Ons_j * DepMean_j)$$

$$+ (uRide - uPot) * \sum_j (Offs_j * ArrMean_j)$$

$$+ uPot * \sum_j (Offs_j * Arr95_j) \quad (3)$$

A second challenge for modeling the reliability – speed tradeoff is to specify an operating cost function that accounts for the effects of reliability. Wirasinghe [3] used a penalty function for late arrivals, which only begs the question of how to weigh the penalty function. We observe that service reliability affects operating cost by determining the need for the scheduled recovery time needed to prevent lateness from growing systematically from cycle to cycle. Therefore, operating cost is assumed proportional to cycle length:

$$\text{Operating Cost} = uOp * C \quad (4)$$

where *uOp* is the cost per vehicle-hour and *C* is the sum of scheduled running time and layover or recovery time, with the constraint that *C* be long enough that mean dispatch lateness be the same at the start of a trip as at the start of the next cycle. Formally, then, we require that

$$DepMean_{n+1} = DepMean_1 + C \quad (5)$$

where $DepMean_{n+1}$ is the mean dispatch time for starting the next cycle after necessary layover and holding (if early) until the scheduled start of the next cycle.

Operations Model

To design a route running time schedule that minimizes the proposed cost function, an operations model is needed that will predict stop arrival and departure time distributions as a function of the running time schedule. The running time schedule consists of a cycle time C and a set of scheduled departure times at every timepoint. The set of stops designated as timepoints is assumed specified. By convention, scheduled departure time at stop 1 is time 0.

A route is modeled as consisting of n stops that form a cycle. Stop 1 represents departures from the start terminal, stop n represents arrivals at the end terminal (which is typically the same location as the start terminal), and stop n+1 represents departure from stop 1 on the next cycle. Segments between neighboring stops represent running time, except for the segment between stops n and $n+1$, generically called a layover segment, which represents time needed to turn the vehicle around (or deadhead to the start terminal if the start and end terminals are not at the same location), load passengers for the next trip, and give operators a minimum necessary rest. If a route has a midcyle terminal, it can be modeled as a pair of stops (one for alighting, one for boarding) separated by a layover segment.

The operations model is simplified by assuming that segment running time is independently distributed, unaffected by the lateness of a vehicle or its leader. On busy, short headway routes, this assumption is clearly invalid because of how dwell time is affected by passenger load, which depends on headway. On long headway routes, however, most passengers will use their targeted bus, and so dwell time independence is a plausible assumption. A more complex operations model may be needed if operators can speed up when running late, or if random traffic delays show significant correlation between segments. On the other hand, intentionally slowing down (“killing time”) when running early can be ignored, because it simply shifts holding a little ways upstream of the timepoint, with minimal effect on user or operating costs.

With the independence assumption, given an initial distribution of departure time at stop 1, the distributions of arrival and departure times at subsequent stops can be determined recursively. Arrival time at a stop is the sum of departure time from the previous stop and running time on the segment between them, and so the arrival time distribution is the convolution of the distributions of those two addends. The departure time distribution at a stop is the same as its arrival time distribution, except at timepoints where early arrivals are truncated to create a probability spike h_i at the scheduled departure time. Holding after the layover segment likewise applies at stop n+1, with departures held until time C. The requirement of consistent dispatch lateness (Eq. 5), along with inclusion of C in the cost function, makes the stop 1 departure time distribution endogenous.

Numerical Example

Our numerical examples use a hypothetical route with n = 6 stops linked by five running time segments and a layover segment that follows stop 6. Stops 2, …, 5 are eligible to be designated as timepoints. There are 160 ons and offs per trip distributed non-uniformly over the stops. Segment running time follows a shifted and truncated lognormal distribution with a standard deviation of 4.5 minutes on each segment. Necessary layover time has a triangular distribution of decreasing probability from a minimum of 1 minute to a maximum of 6 minutes. Aggregating over the route, the standard deviation of cycle time, without holding, is 10.1 minutes.

Arrival and departure time distributions were calculated explicitly for each stop using MatLab. Consistency between the mean stop 1 and stop n+1 departure times was obtained by running several warm-up cycles, adjusting C as needed until the means of the distributions were within 0.1 s. By searching from below for the shortest possible recovery time that avoids systematically increasing delay in the next cycle, the optimal recovery time is found for any running time schedule.

Figure 2 shows how user and operating cost varies with the scheduled running time, with running time distributed along the route in proportion to uncontrolled running time. Every eligible stop is a timepoint. Scheduled running time is expressed as a ratio to mean uncontrolled running time. When that ratio is 0.6, there is no timepoint control because buses are always running late. Cost components are shown as differences from this no-timepoint case in dollars per cycle, with unit costs of $9, $6, and $4.5 per passenger-hour for waiting, riding, and potential travel time respectively, and $80 per vehicle-hour. Cycle time, and therefore recovery time, is optimized in every case.

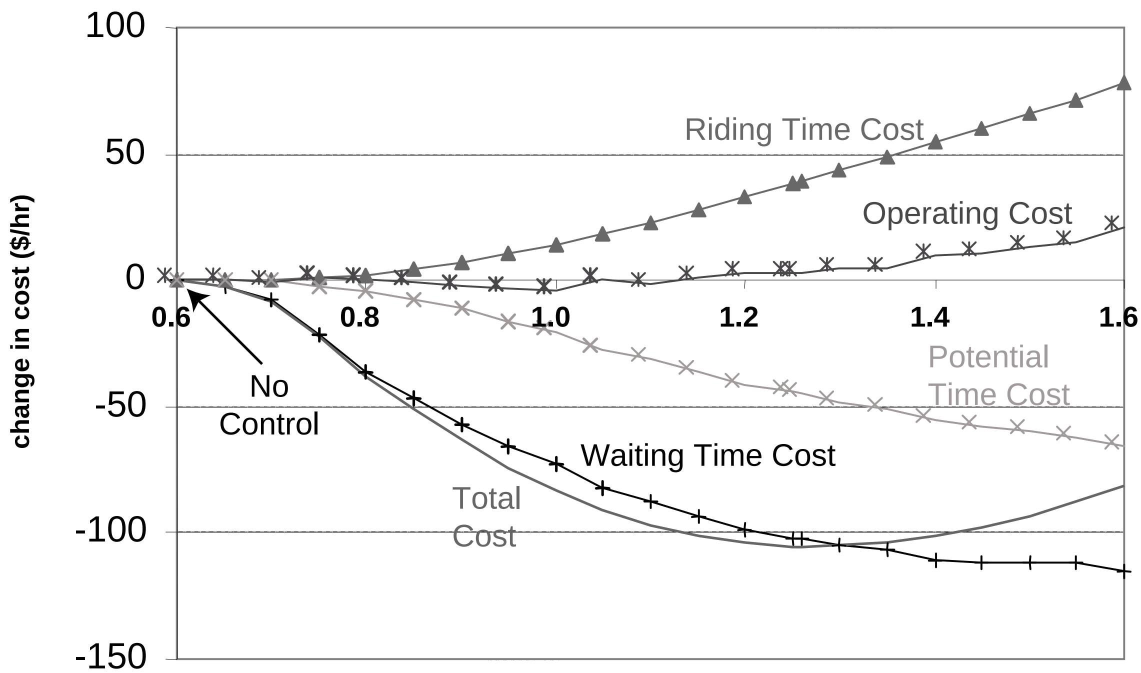

Figure 2. Cost components versus scheduled route running time. Route running time (x-axis) is given as a ratio to mean uncontrolled running time. Recovery time is optimized for each case.

Results show how the waiting cost falls dramatically as slack is added to the running time schedule. Adding slack time also reduces potential travel time, and increases riding time, as expected. Total cost has a U-shape, with an optimum when scheduled segment running time is 1.26 times the mean uncontrolled running time.

A very interesting result is that until the amount of slack in the running time schedule reaches about 35%, the optimal operating cost, which is proportional to cycle length, stays virtually unchanged. In this range, slack time added to the running time schedule is essentially subtracted, minute for minute, from scheduled recovery time. This is significant, because it suggests that timepoint holding can be introduced without necessarily increasing operating cost, contrary to conventional wisdom.

Results of more numerical experiments are given in [7]. In this paper, we focus on gaining insights into the form of an optimal running time schedule by analyzing optimality conditions.

Marginal Impacts of Schedule Variables

The control variables in the optimal running time schedule design problem are C and $SchDep_i$ for every timepoint. Thanks to the separable form of the cost function, the marginal impact of a change in any schedule variable can be formulated as the sum of its impact on operating cost and on $Dep02_j$, $DepMean_j$, $ArrMean_j$, and $Arr95_j$.

Impact of a Change in Scheduled Departure Time

Impact on 2-Percentile Departure Time

If stop i is a timepoint, it can be assumed that

$$h_i > 0.02 \tag{6}$$

and therefore that

$$Dep02_i = DepSch_i \tag{7}$$

Therefore,

$$\frac{dDep02_j}{dDepSch_i} = \begin{Bmatrix} 1 \text{ for } j = i \\ 0 \text{ otherwise} \end{Bmatrix} \tag{8}$$

Impact on Mean Departure Time

If stop i's scheduled departure time is made later by a marginal amount $dDepSch_i$, departure time at stop j will be affected only for trips that were (1) held at i, which occurs with probability h_i, and (2) arrive late at i without being held at any timepoint between i and j, by virtue of arriving late at every timepoint in between. Because of the cyclical nature of the route, the change $dDepSch_i$ applies to previous cycles as well, and so an impact can also arise from trips that were held at i in a previous cycle, arrive late at j in the current cycle, and are not held (by virtue of being late) at any timepoint or terminal after being held at i. For stops that are timepoints, let

$P[unheld_{ijk}|held_i]$ = P[given that a bus departs stop i on time in cycle m, that it arrives late at j in cycle $m + k$ without having been held in between]

with $P[unheld_{ii0}|held_i] = 1$ by definition. Summing over past cycles, let

$$P[unheld_{ij}|held_i] = \Sigma_{k \geq 0} \; P[unheld_{ijk}|held_i]$$

Because recovery time tends to limit the probability of passing through a terminal without being held, contributions to the sum from previous cycles can be expected to be small. In any event,

$$\frac{dDepMean_j}{dDepSch_i} = h_i * P[unheld_{ij} \mid held_i] \quad \text{if } j \text{ is a timepoint} \tag{9}$$

As a general trend, $P[unheld_{ij}|held_i]$ decreases as the number of intervening timepoints between i and j increases, both because each intervening point is an opportunity to "catch" an early trip, and because timepoints have schedule slack that are meant to reduce lateness. Because of the especially strong ability of terminals to absorb lateness thanks to their scheduled recovery time, one can expect $P[unheld_{ij}|held_i]$ to be small if $j < i$ unless the route has no terminal (e.g., a loop shuttle in which vehicles never empty). $P[unheld_{ij}|held_i]$ also depends on the relative strictness of timepoints i and j, increasing as timepoint i becomes stricter and as timepoint j is becomes less strict.

If j is not a timepoint, the impact to mean departure time at j will be the same as the impact to the preceding timepoint. If we define

$b(j) = j$ if j is a timepoint; otherwise, the timepoint before j

then Eq. 9 can be generalized to

$$\frac{dDepMean_j}{dDepSch_i} = h_i * P\,[unheld_{i,b(j)} \mid held_i] \tag{9a}$$

Impact on Mean Arrival Time

The marginal impact of a change in $DepSch_i$ on mean arrival time at j is the same as its impact on mean departure time at (j-1), since by assumption running time from departure at (j-1) to arrival at j is unaffected:

$$\frac{dArrMean_j}{dDepSch_i} = \frac{dDepMean_{j-1}}{dDepSch_i} \tag{10}$$

It may seem intuitive that a change in $DepSch_i$ will only affect the riding time of passengers who are traveling through i; however, this is not necessarily the case. Consider passengers traveling from j to k, with $i < j < k$. Their marginal change in riding time is

$$(dArrMean_k - dDepMean_j) = dDepSch_i * h_i * (P[unheld_{i,k-1}|held_i] - P[unheld_{ij}|held_i])$$

which will be zero only if $P[unheld_{ij}|held_i] = P[unheld_{i,k-1}|held_i]$, which in general is not the case. In fact, because $P[unheld_{i,k-1}|held_i] \geq P[unheld_{ij}|held_i]$, holding tends to decrease the riding time of passengers who board downstream of the holding point.

This apparently counterintuitive result arises because riding time is the sum of uncontrolled running time plus holding time. The former is (by assumption) unaffected by timepoint schedules; however, stricter holding at an upstream stop reduces the need for holding at later timepoints, lowering riding time for people boarding downstream.

Impact on 95-percentile Arrival Time

Let $fArr_j(t)$ and $FArr_j(t)$ be the PDF and CDF, respectively, of arrival time at j, and let t_o be the 95-percentile arrival time at stop j, that is, the time for which $FArr_j(t_o) = 0.95$. If scheduled departure time at i increases by $dDepSch_i$, trips that were held at i in either the current or a previous cycle, have not been held since, and arrive at j a moment before t_o will now arrive after t_o. The result will be to increase the fraction of arrivals after t_o, and therefore to decrease $FArr_j(t_o)$ by an amount $dFArr_j(t_o)$ given by:

$$dFArr_j(t_o) = -h_i * dDepSch_i * \Sigma_{k \geq 0} P(unheld_{ijk}) * fRun_{ijk|unheld(ijk)}(t_o + kC - DepSch_i) \tag{11}$$

where

$P(unheld_{ijk})$ = P(a trip will not be held between an on-time departure at i and its arrival, k cycles later, at j)

$fRun_{ijk|unheld(ijk)}(t)$ = probability density of running time from an on-time departure at i to arrival at j, k cycles later, given that it wasn't held at any timepoint or terminal in between.

With this decrease in $FArr_j(t_o)$, the 95-percentile arrival time at j will be changed to t_o + $dArr95_j$, where

$$dArr95_j = -dFArr_j(t_o) \,/\, fArr_j(t_o) \tag{12}$$

Combining these last two results,

$$\frac{dArr95_j}{dDepSch_i} = h_i * \frac{\sum_{k\geq 0} P(unheld_{ijk}) * fRun_{ijk|unheld(ijk)}(Arr95_j + kC - DepSch_i)}{fArr_j\left(Arr95_j\right)} \tag{13}$$

By Bayes' theorem, this marginal impact is simply

P[held at i and not held since | arrive at j "very late"]

where "very late" means in the neighborhood of $Arr95_j$. As a general trend, this marginal impact should be greatest when i is a terminal or other strict timepoint, and should fall as j becomes more distant from i, especially if a terminal separates j from i (i.e., if $j < i$), because of how intervening timepoints and especially layover points decrease the probability of arriving at j without having been held in between.

Impact of a Marginal Change in Cycle Time

A marginal increase dC in cycle time increases recovery time at the terminal, changing the stop 1 departure time by $-dC$ for the fraction $(1 - h_1)$ of trips that leave stop 1 late.

That change in departure time propagates to downstream stops until a bus is held at a timepoint. Using similar logic to before,

$$\frac{dDep02_j}{dC} = 0 \tag{14}$$

$$\frac{dDepMean_j}{dC} = -(1-h_i)*\sum_{k\geq 0}(k+1)*P[unheld_{1,b(j),k}\;|\;unheld_1] \tag{15}$$

$$\frac{dArrMean_j}{dC} = \frac{dDepMean_{j-1}}{dC} \tag{16}$$

$$\frac{dArr95_j}{dC} = -(1-h_1)*\frac{\sum_{k\geq 0}P(unheld_{1jk})*(k+1)*\,fArr_{jk|unheld(jk)}(Arr95_j + kC)}{fArr_j\left(Arr95_j\right)}$$

= P(not held at stop 1 or at any stop in between | arrive “very late” at j) (17)

where $fArr_{jk|unheld(jk}(t)$ = probability density of arriving at time t at stop j after leaving the terminal k cycles earlier without being held there or at any timepoint since. The (k+1) term appearing in Eqs. 15 and 17 reflects that fact that each time a bus cycles back to stop 1, its lateness will decrease by dC.

Of course, a marginal change in cycle time also affects operating cost:

$$\frac{d(\text{Operating Cost})}{dC} = uOp \tag{18}$$

Optimality Conditions

Because of the separable cost function, the marginal cost to users of a change in a scheduled departure time or a change in cycle time is simply a sum of the various marginal impacts shown earlier, weighted by the appropriate number of ons, offs, and unit costs, consistent with Eqs. 1, 3, and 4. For a change in scheduled departure time,

$$d(Total\ cost) / DepSch_i =$$

$$- uWait * Ons_i$$

$$+ h_i * (uWait - uRide) * \sum_j (Ons_j * P[unheld_{i,b(j)} \mid held_i])$$

$$+ h_i * (uRide - uPot) * \sum_j (Offs_j * P[unheld_{i,b(j-1)} \mid held_i])$$

$$+ h_i * uPot * \sum_j \left(Offs_j * \frac{\sum_{k \geq 0} P(unheld_{ijk}) * fRun_{ijk|unheld(ijk)}(Arr95_j + kC - DepSch_i)}{fArr_j\left(Arr95_j\right)} \right) \quad (19)$$

Setting this derivative equal to zero, one can solve for h_i, yielding an optimality condition for the strictness of a timepoint:

$$h_i = \frac{uWait * Ons_i}{\sum \left(\text{many terms involving } P(unheld)\right)} \quad (20)$$

Likewise, setting the derivative of total cost with respect to C equal to zero yields an optimality condition for dispatch reliability at stop 1, with implications for the optimal amount of recovery time:

$$h_1 = 1 - \frac{uOp}{\sum\left(\text{many terms involving } P(unheld)\right)} \tag{21}$$

We call h_1 the dispatch reliability, since it is the probability of an on-time dispatch from the terminal.

Insights into Optimal Running Time Schedules

Because the conditional probabilities related to being held at upstream stops are themselves outcomes of operating under a given timepoint schedule, marginal analysis does not afford a closed form solution. Nevertheless, having explicit formulas for marginal impacts can still be useful in developing optimizing algorithms, and for the insight they offer into optimal scheduling. We have already mentioned the result that holding tends to reduce riding time for passengers boarding downstream of a timepoint. Some further points are offered in this section.

Optimal Timepoint Strictness

From Eq. 20, one can see that a timepoint's strictness should increase with its boardings demand. This results stems from the fact that of the four marginal impacts of a change in scheduled departure time at a stop (Eqs. 8, 9, 10, and 13), the impact on waiting time is direct and concentrated on passengers boarding at that stop, while the other impacts are spread over many stops. There is no such concentrated effect from alightings demand, which therefore has far less influence on timepoint schedules.

One corollary of this result is that if there are to be a limited number of timepoints, stops with heavier boarding rates are preferred. A second is that compared to a policy of distributing schedule slack uniformly along a route, an optimal schedule will shift running time to segments ending at stops with heavy boardings. This phenomenon can be seen in Figure 3, which presents, for the example route described earlier, a comparison of an optimal schedule against a schedule in which running time slack is distributed uniformly (with an optimized percentage slack). The excess boardings index shown for each stop is proportional to the amount by which a stop's boardings exceed average boardings per stop. A stop's running time adjustment is defined as how much longer is the optimal scheduled running time for the segment ending at that stop compared to the uniform-slack running time. For example, at stop 4, a heavy boardings stop, the optimal running time for the segment ending at stop 4 is 3.25 minutes longer than what it would be if a uniform optimized percentage of slack were applied to the route, while stop 3, with less than the average boardings demand, has a negative adjustment to its segment's running time.

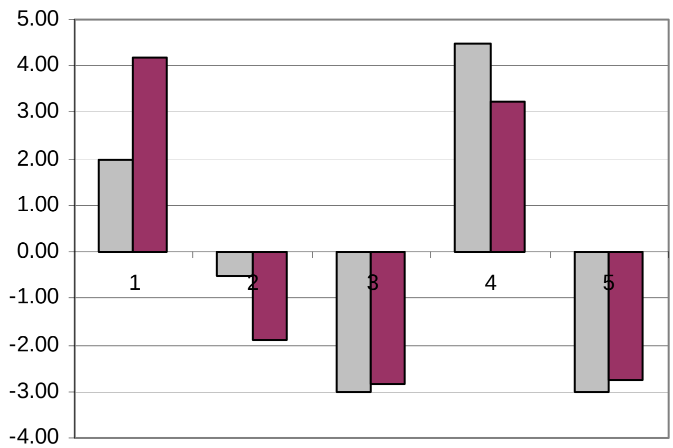

Figure 3. Excess boardings (gray bar, dimensionless) and running time adjustment (dark bar, in minutes) by stop.

The location of a stop along the route also has an effect on its optimal timepoint strictness. The denominator of Eq. 20, a sum involving probabilities that buses arrive at downstream stops without having been held since the subject stop, tends to be greatest for the early stops on a route, because later stops have fewer downstream stops they can affect without cycling through the layover point. Therefore, closeness to the start of the route tends to amplify the effect that boardings demand has on segment running time. This effect can be seen in Figure 3, where the size of the running time adjustments (whether positive or negative) relative to excess boardings is far greater at the first two stops that at the later stops.

For our example route, the welfare difference between using schedules optimized by timepoint and schedules with only an optimized percentage slack was small but not negligible. Compared to operating with no timepoints, the incremental benefit of optimizing each timepoint individually is 5.2%, or the equivalent of 0.34 minutes of riding time per passenger.

Holding at Timepoints versus at Dispatch Terminals

The optimality condition for dispatch reliability (Eq. 21) shows that there is no universally optimal dispatch reliability; instead, optimal dispatch reliability depends on how strict timepoint control is along the route. The more control is exercised at timepoints, the less need for control at the dispatch terminal, and therefore the smaller is the optimal dispatch reliability. This trend is illustrated for our example route in Figure 4. As the scheduled running time increases, optimal dispatch reliability falls from 92% when there is no timepoint control to 84% at the optimal scheduled running time (1.26 times the mean uncontrolled running time), and lower still as more slack in inserted into the schedule.

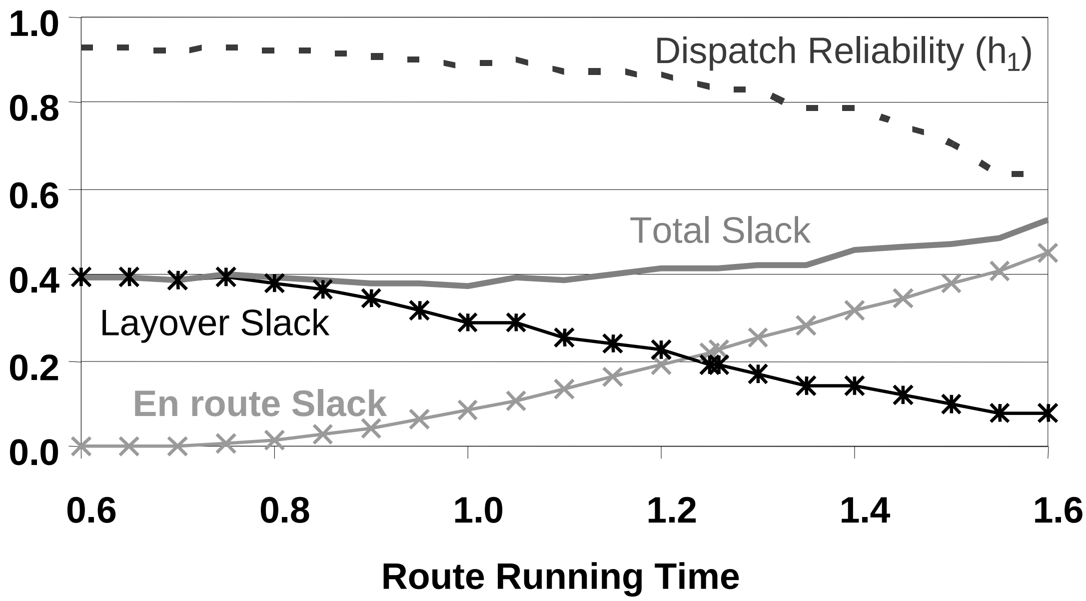

Figure 4. Optimal dispatch reliability and slack time (expected time spent holding) spent en route and during layover versus scheduled running time. Running time and slack time are expressed as a ratio to mean uncontrolled running time. Recovery time is optimized for each case.

Further insight into optimal schedules can also be seen in Figure 4 by examining how the expected amount of time spent holding ("slack") varies with scheduled running time. As scheduled running time increases, so does expected time spent holding at timepoints ("en route slack"); but at the same time, almost in perfect 1-to-1 substitution, the amount of holding occurring at the terminal ("layover slack") decreases, with the result that the total amount of slack in the schedule stays virtually unchanged, until the scheduled running time reaches a rather high level.

This is an important result that runs counter to conventional wisdom. Many people tend to separate decisions about needed recovery time from en-route slack time, and imagine therefore that imposing stricter timepoints must necessarily lead to longer cycle times and therefore greater operating cost. However, when timepoints are made stricter, two effects diminish the optimal recovery time. First, the arrival distribution at the final stop becomes tighter, making less recovery time necessary to achieve the same probability of on-time dispatch. Second, the optimal dispatch reliability is itself lower.

An important practical consideration that can limit application of this result is the need, often contractual, for operators to have sufficient rest time during layovers. Our numerical experiments suggest that on routes with high running time variability, the optimal recovery time will often be sufficient to satisfy operator rest needs as typically expressed by both labor contracts and scheduling practice. However, on routes with high levels of reliability, the optimal recovery time may be too small to meet operator rest needs. In such a case, operator layover needs will govern the scheduled recovery time, and optimal scheduled running time will be smaller, because with more slack built into the layover, less will be needed en route.

Optimal Number of Timepoints

The fraction of stops that are treated as timepoints is a point of difference between U.S. practice and practice in the Netherlands and many other European countries. In the U.S., it is typical for about 5 to 10% of the bus stops serve as timepoints, which are the only stops shown in published schedules. In the Netherlands, scheduled departure times are published for every stop, and every stop with boardings is treated as a timepoint, except where space to hold buses is lacking. The nature of the cost function is such that it can only help to increase the number of stops that are designated as timepoints, assuming scheduled departure times are optimized for each point, because the solution space permits very early scheduled departures that have virtually the same effect as no control. However, one would expect that there are diminishing benefits to increasing the number of timepoints.

For the simple route we modeled, the benefits of increasing the number of timepoints can be seen in Figure 5. In every case, the highest demand stops were selected as timepoints, and departure times at each timepoint and recovery time at the terminal are optimized.

One can see how the benefits continue to accrue with each added timepoint, although most of the benefit is realized when two of the four eligible stops are timepoints.

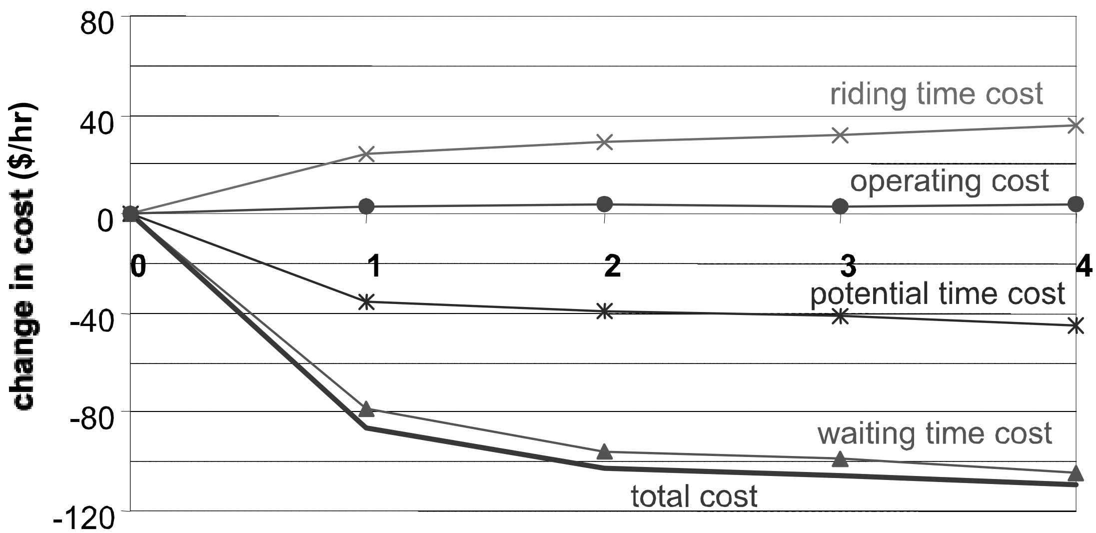

Figure 5. Cost versus number of timepoints on a 6-stop route.

Conclusion

A proposed framework for measuring user costs related to unreliability permits one to explore the tradeoff between reliability and speed inherent in the design of running time schedules with timepoints. User costs are related to mean and extreme values of the arrival and departure time distributions at each stop. Optimality conditions for timepoint strictness and dispatch reliability were derived. They show that timepoints should be stricter where boardings are greater, with this effect strongest for stops near the start of the route, and that dispatch reliability should be worse as timepoint control becomes more strict. Up to a point, slack time added to a schedule en route (i.e., at timepoints) simply substitutes for slack time at layover, meaning than timepoint holding does not necessarily increase cycle length or operating cost. Benefits increase with the number of timepoints, but with diminishing returns.

References

1. Osuna EE, Newell GF (1972) Control strategies for an idealized public transportation system. Transportation Sci 6: 52-72
2. Hill SA (2008) Holding strategies in a bus route model. arXiv:0709.0078v1 [physics.soc-ph]
3. Jhao J, Dessouky M, Bukkapatnam S (2006) Optimal slack time for schedule-based transit operations. Transportation Sci. 40:529-539
4. Wirasinghe SC (2003) Initial planning for urban transit systems. In: Lam WHK, Bell MG (eds) Advanced Modeling for Transit Operations and Service Planning. Pergamom, pp 1-30
5. Wirasinghe SC, Liu G (1995) Determination of number and locations of time, points in transit schedule design – Case of a single run. Annals of Operations Research 60 : 161-191
6. Furth PG, Muller ThHJ (2006) Service reliability and hidden waiting time: Insights from AVL data. Transportation Research Record 1955: 79-87
7. Furth PG, Muller ThHJ (2007) Service reliability and optimal running time schedules. Transportation Research Record 2034: 55-61.