Title
How to Improve Bus Service

Author
Daganzo, Carlos F

Publication Date
2008-08-12

How to Improve Bus Service

Carlos F. Daganzo
WORKING PAPER
UCB-ITS-VWP-2008-6

UC Berkeley Center for Future Urban Transport
A **VOLVO** Center of Excellence

August 2008

HOW TO IMPROVE BUS SERVICE

Carlos F. Daganzo
Institute of Transportation Studies
University of California, Berkeley, CA 94720

(August 12, 2008)

ABSTRACT

Bus schedules cannot be easily maintained on busy lines with short headways: Experience shows that buses offering this type of service usually arrive irregularly at their stops, often in bunches. Although transit agencies build slack into their schedules to alleviate this problem, their attempts often fail because practical amounts of slack cannot prevent large localized disruptions from spreading system-wide. This paper describes a more resilient control scheme that overcomes this problem. The method also produces even headways with less slack than the conventional approach. Thus, buses can run faster and be more productive.

1. INTRODUCTION

This paper examines a new way of delivering reliable transit service. The focus is on high frequency transit lines where out-of-vehicle delay is given by the headways irrespective of the schedule. It is well known from experience and theory (e.g., Newell, 1977) that collective bus motion for these types of systems is unstable; i.e., that even if one starts with perfectly even headways, they invariably become irregular, and if enough time passes buses bunch up. The reason for this instability is that if a disruption causes a bus to slow up relatively to the bus it follows, the bus encounters more passengers along the way, and these extra passengers delay it further. Conversely, the next bus has a tendency to catch up.

To fight this problem, transit agencies insert slack into their schedules, and require buses (or transit vehicles) to depart on time at predefined control points along the route. The slack is calculated so buses can make up time lost due to random recurrent travel disruptions between control points, but slack reduces the commercial speed of the buses, so it cannot be too big. To limit the total amount of bus delay, control points are spaced widely so that typical routes include only a few.

Unfortunately, even considerable amounts of slack cannot guarantee on time performance in the real world where large disruptions often occur. For example, if a single transit vehicle suffers an uncommonly large delay (e.g., a transit station power failure, a brief mechanical malfunction or some other incident) so that its headway grows beyond a critical value, then enough extra passengers could arrive along its route to force it to inexorably fall further and further behind schedule. The unfortunate bus would be eventually caught by subsequent buses, and if they traveled in a loop all would eventually cluster. Generalized disruptions, such as heavy traffic or a snow storm, are even more problematic because they can push large groups of vehicles behind schedule. If this happens the delayed vehicles would try to go as fast as they can to catch up with the schedule and in doing so would tend to bunch again, disrupting themselves and the rest of the system. In this case the schedule becomes useless; counterproductive in fact.

This paper shows that if buses are controlled in an adaptive way, based on information from the buses they follow rather than a fixed schedule, they can produce even headways and fast service, without the resiliency problems of schedule-based control. Section 2 below defines the terms and presents the basic strategy, Sec. 3 examines its performance, and Sec. 4 discusses the results and presents an example.

2. THE ANALYSIS FRAMEWORK

2.1 Definitions and Background

The object of our analysis is for now a single bus (or transit) line operated with a regular schedule under time-independent conditions. This schedule is defined by the times $t_{n,s}$ at which each bus n is expected to arrive at a series of control points s lying along the line. The schedules are of the form:

$$t_{n,s+1} = t_{0,0} + nH + \sum_0^s p_i \qquad \text{for} \quad n, s = 0, 1, 2\ldots \tag{1}$$

where H is the service headway and p_i is the target travel time for the segment between points i and i+1, which is common to all buses. Note that p_i is indexed by the point at the beginning of its segment and that (1) allows for spatially inhomogeneous routes. Note too that the transit agency can reuse the buses once they reach the end of the line, so the index n does not really refer to individual buses but to bus *runs*. A brief review of known, useful facts now follows.

2.1.1 Uncontrolled motion: In a deterministic and stationary world with no random variations, p_s could be set equal to c_s, the average travel time from s to s+1 including the delay due to stops when the headway is H. Thus, the bus motion would simply be:

$$t_{n,s+1} = t_{n,s} + c_s\,. \tag{2}$$

In reality, however, random disturbances due to traffic, passenger needs and the vagaries of bus drivers give rise to errors $\varepsilon_{n,s} = (a_{n,s} - t_{n,s})$ between the actual bus arrival times $a_{n,s}$ and those scheduled. As a result, actual headways, $h_{n,s} = (a_{n,s} - a_{n\text{-}1,s})$, also differ from the target, and this affects the travel times since longer headways imply more passengers to be served. To model these effects in a simple way we shall assume that the average uncontrolled time from s to s+1, which we denote $u_{n,s}$, is linear in $h_{n,s}$, i.e. that:

$$u_{n,s} = c_s + \beta_s(h_{n,s} - H) \quad \text{for some} \quad \beta_s \geq 0. \tag{3}$$

The constant β_s is a dimensionless constant expressing the marginal increase in expected bus delay arising from a unit increase in headway—since a longer headway results in additional passenger moves, which delay the bus. For most bus lines, alighting moves are quick so that the increased bus delay depends mostly on boarding moves. The constant β_s can then be interpreted intuitively. Note that the expected bus delay from s to s+1due to these boarding moves is the product of three factors: the passenger arrival rate between our two control points, the headway $h_{n,s}$, and the average marginal delay per boarding move. Since this product is $\beta_s h_{n,s}$ by definition, we see that β_s is the expected number of passenger arrivals in segment (s, s+1) during the average marginal delay induced by one boarding move. Values of β_s can range from 10^{-2} to 10^0 depending on demand levels and the length of the segment. For example, line 44 of SF-Muni which is familiar to the author exhibits $\beta_s \approx 10^{-1}$ during the rush hour and $\beta_s \approx 10^{-2}$ on weekends for segments spanning one stop.

We also assume that the actual travel time for segment (s, s+1), $U_{n,s}$, includes a random noise term caused by the aforementioned random disturbances. Because this random noise term only becomes known upon the bus' arrival at s+1 we label it, $\nu_{n,s+1}$. Thus, $U_{n,s} = u_{n,s} + \nu_{n,s+1}$ so that:

$$U_{n,s} = c_s + \beta_s(h_{n\,s} - H) + \nu_{n,s+1} \quad \text{for some } \beta_s \geq 0. \tag{4}$$

This noise term is assumed to have zero mean, variance $\sigma_{s+1}{}^2$, and to be independent of h_{ns}.

All the assumptions we have made in connection with (3) and (4) are reasonable as long as the headways are not allowed to deviate much from H, and should be especially accurate if the control points are closely spaced and buses do not skip stops. Thus, the stochastic law of motion for an uncontrolled bus running close to schedule is:

$$a_{n,s+1} = a_{n,s} + U_{n,s} = a_{n,s} + c_s + \beta_s(a_{n,s} - a_{n\text{-}1,s} - H) + \nu_{n,s+1}. \tag{5}$$

We have already mentioned that this type of motion is unstable; i.e., that headways increasingly deviate from the target as time passes, until buses bunch up. The reason is that the terms including β_s act like forces that attract the pair of buses on opposite sides of a headway shorter than H and repel them when the headway is longer. To illustrate this effect Table I shows the ratio of the RMSE in the deviations from the schedule ($a_{n,s} - t_{n,s}$) observed at the end of segment s of a homogeneous route, and the RMSE that would have been observed at the same location if β was 0. The table assumes that the noise terms are uncorrelated with identical variance (u.i.v.) and that the buses are initially dispatched without error. Thus, as shown by the first row, the ratio in question must be 1 independent of β for s =1. Note the strong effect of β on this amplification factor.

Table I: Amplification of the RMSE in deviations from the schedule due to the attraction parameter β. Noise is u.i.v.

	β = .01	β = .03	β = .1	β = .3	β = 1	β = 3
s-1= 0	1	1	1	1	1	1
s-1 = 1	1	1	1.1	1.2	1.8	3.6
s-1 = 2	1	1	1.1	1.4	3.7	18
s-1 = 4	1	1	1.3	2.4	22	560
s-1 = 8	1	1.1	1.8	9.6	1100	$>> 10^3$
s-1 = 16	1.1	1.4	4.4	250	$>> 10^3$	$>> 10^3$
s-1 = 32	1.2	2.2	47	$>> 10^3$	$>> 10^3$	$>> 10^3$

2.1.2 Conventional schedule control: To avoid bunching problems and keep the system running on time, transit agencies introduce enough slack into their schedules to guarantee that buses can meet the target travel times for each segment despite the variability of the uncontrolled travel times.

We recommend introducing at least four standard deviations of the noise term as slack, i.e. $p_s \geq c_s + 4\sigma_s$, because the noise terms tend to be positively skewed. With this form of control thus, reliable service is achieved at the cost of slowing bus service by approximately $4\sigma_s$ time units for the segment from s to s+1, and a like amount for all other segments.

Since the total amount of bus delay is the sum of the delays at the control points, transit agencies usually combine segments to reduce the number of control points. This is not a panacea, however. For example, if β was 0.1 for a single-stop segment and the transit agency were to combine 16 extra stops, the RMSE in arrival times at the control point would be $4.5\times(16)^{1/2}$ = 18 times larger than after a single-stop segment; see Table I. Thus, the combination would have gained nothing. Of course, combining segments is more effective for smaller β's but separating control points lowers the system's resiliency. This brings us to the new model.

2.2 The Proposed Strategy and its Dynamic Equations

In view of the destabilizing forces associated with (5) we propose introducing a compensating force that would attract buses when they are too far and repel them when they are too close. The simplest policy of this type would act only on the following bus of each pair, speeding it when it lags and retarding it when it closes. We propose adding a headway-dependent delay $D_{n,s}$ to the time that the bus on run n would otherwise spend traveling uncontrolled from s to s+1 so that the law of motion becomes: $a_{n,s+1} = a_{n,s} + U_{n,s} + D_{n,s}$. To compensate for the attraction force and then reverse it, the added delay is chosen to be

$$D_{n,s}= d_s + (\alpha+\beta_s)(H - h_{n,s}) \quad \text{for some } d_s \geq 0 \text{ and } \alpha \in (0, 1). \qquad (6)$$

The constants d_s and α characterize the policy and represent the average bus delay at equilibrium and the sensitivity to control, respectively. At equilibrium, the arrival times would satisfy (1) with $p_s = c_s + d_s$.

Every constant and variable on the RHS of (6) is known by the time the bus on run n departs s. So they are available when needed. The constants should be chosen to ensure that added delays are rarely negative. In other words, if we use $\sigma_{hs}{}^2$ for the variance of the headway at s, which is an endogenous quantity to be determined, the constants should satisfy: $d_s \geq 3(\alpha+\beta_s)\sigma_{hs}$. We use 3 standard deviations because as we shall see in the next section $h_{n,s}$ is approximately Gaussian.

Let us now use the buses' stochastic law of motion, $a_{n,s+1} = a_{n,s} + U_{n,s} + D_{n,s}$, to derive a recursive set of dynamic equations for the deviations from the schedule. Inserting (4) and (6) in the law of motion we find: $a_{n,s+1} = a_{n,s} + c_s + \beta_s(h_{n,s} - H) + \nu_{n,s+1} + d_s + (\alpha+\beta_s)(H - h_{n,s})$. In terms of arrival times, this is:

$$a_{n,s+1} = a_{n,s} + c_s + d_s + \alpha\,(H - a_{n,s} + a_{n-1,s}) + \nu_{n,s+1}\,. \qquad (7)$$

Now focus on the deviations $\varepsilon_{n,s}$ of the $a_{n,s}$ from the equilibrium arrival times (1). The latter satisfy $t_{n,s+1} = t_{n,s} + c_s + d_s$ and $H - t_{n,s} + t_{n-1,s} = 0$. So, subtracting these relations from (7) we find our dynamic equation:

$$\varepsilon_{n,s+1} = (1-\alpha)\varepsilon_{n,s} + \alpha\,\varepsilon_{n-1,s} + \nu_{n,s+1} \qquad \text{for} \quad n = 1, 2\ldots;\ s = 0, 1, 2,\ldots \qquad (8)$$

The boundary conditions are $\varepsilon_{0,s} = 0$ for $s = 0, 1, 2\ldots$ and $\varepsilon_{n,0} = 0$ for $n = 0, 1, 2\ldots$

It is convenient to introduce the constants $f_0 = (1-\alpha)$, $f_1 = \alpha$, and $f_j = 0$ for all other integer j, and at the same time define $\varepsilon_{n,s} = 0$ and $\nu_{n,s} = 0$ for all $n < 0$, because this convention allows us to rewrite (8) as:

$$\varepsilon_{n,s+1} = \sum_{j=-\infty}^{\infty} f_{n-j}\,\varepsilon_{j,s} + \nu_{n,s+1} \qquad \text{for} \quad n = 1, 2\ldots;\ s = 0, 1, 2,\ldots \qquad (9a)$$

and then further simplify it on recognizing that the first term on the RHS is a convolution. Thus, using boldface for vectors and "∗" for the convolution operation, we rewrite (9a) as:

$$\boldsymbol{\varepsilon}_{s+1} = \boldsymbol{f} * \boldsymbol{\varepsilon}_s + \boldsymbol{\nu}_{s+1} \qquad \text{for} \quad s = 0, 1, 2\ldots \tag{9b}$$

where $\boldsymbol{f}$ is the kernel of the convolution. Note that $\boldsymbol{f}$ is the p.m.f. of a binary Bernoulli random variable, and therefore it will be called the Bernoulli kernel.

This paper will also examine (9) where $\boldsymbol{f}$ is more generally allowed to have the form of the p.m.f. of a non-negative random variable with mean $\mu > 0$ and variance $v^2 \in (0, \infty)$. Consideration shows that this general case arises when instead of (6) we use the following for the added delay:

$$D_{n,s} = d_s + (F_0 + \beta_s)(H - h_{n,s}) + \sum_{j=1}^{\infty} F_j \, (H - h_{n-j,s}). \tag{10}$$

where F_j is the complementary c.d.f. of $\boldsymbol{f}$ (i.e., $F_j = \sum_{m=j+1}^{\infty} f_m$) and $h_{n,s} = H + \varepsilon_{n,s} - \varepsilon_{n-1,s}$, which is defined for $n \in (-\infty, \infty)$. Note that (10) is a weighted sum of past headway deviations so the calculation can be done by the time it is needed, and that earlier headways carry less weight.

2.3 Time-dependence and other practical considerations

The proposed strategy can be extended to practical situations where the transit agency wishes to provide an irregular schedule with different headways H_n for different runs while recognizing that the expected demand and traffic conditions change with time, and not just space; i.e. that the parameters c_s , β_s and σ_s depend on the run number and should be labeled: $c_{n,s}$, $\beta_{n,s}$ and $\sigma_{n,s}$. The schedule is now of the form:

$$t_{n,s+1} = t_{0,0} + \sum_1^n H_j + \sum_0^s p_i \qquad \text{for} \quad n, s = 0, 1, 2\ldots$$

where it is understood that $\sum_1^0 H_j = 0$. The only quirk in this case is that the scheduled travel times p_s should be set to accommodate the slowest bus, i.e., that which requires $max_n\{c_{n,s}\}$ time units on average when running on schedule. Therefore, if this maximum travel time is denoted c_s, the scheduled travel times should be defined to be of the form $p_s = c_s + d_s$, where d_s continues to be the systematic slack built into the schedule by the control rule.

If we now repeat the derivations of (3)-(10) recognizing these changes in the model we find that all the equations continue to hold if we simply replace H by H_n and β_s by $\beta_{n,s}$. Qualitatively, the new control rules continue to stipulate added delays that compensate for the (now run-dependent) attraction forces acting on each bus; e.g., (6) becomes $D_{n,s} = d_s + (\alpha + \beta_{n,s})(H_n - h_{n,s})$. But more importantly, (8) and (9), which do not include run-dependent parameters, continue to hold. So the analysis of these equations in Sec. 3 applies to the general time-dependent problem with an irregular schedule.

To complete the discussion of practical matters we must describe what to do in the rare occasions when the calculated $D_{n,s}$ would require a bus to travel faster than it can; i.e., be negative in the case of a regular schedule or smaller than $c_{n,s} – c_s$ in the general case. We propose asking bus drivers to speed up by not picking up passengers until their $D_{n,s}$ becomes feasible again. This is not so drastic a form of intervention as one may think because for systems operated with small headways and short segments it would delay only a little the passengers waiting at a just a few stops.[1]

An appealing feature of the proposed strategy is that it increases bus speed at the first hint of an unduly long headway, before the problem grows to unmanageable proportions. In so doing, it acts as a robust servomechanism that compensates for the type of recurrent disruptions that wreak havoc with schedule-based control.[2] The properties of this servomechanism are analyzed below.

3. STABILITY RESULTS

This section examines the performance of the strategy from the perspective of reliability. Subsection 3.1 analyzes its on-time performance, i.e. the deviations (8), and subsection 3.2 its ability to maintain even headways. The results are encouraging. Subsection 3.1 will show that although the deviations from the schedule grow without limit as a bus run progresses, they do so at a declining rate and they turn out to be small for runs of practical length (with fewer than 100 segments). More importantly, subsection 3.2 will show that the deviations in headway do not grow without limit; they are in fact uniformly bounded and quite small for any number of segments and bus runs.

3.1 Deviations from the schedule

Our first question is determining whether the deviations in (8) stay bounded (and small) or grow to infinity as (8) is iterated for increasing values of n and s. We first answer this question assuming that the $\nu_{n,s}$ are bounded, i.e. $|\nu_{n,s}| \le M$, and then examine in more detail the case where they are u.i.v.

If we replace $\boldsymbol{\varepsilon}_s$ in the RHS of (9b) by its expression according to (9b) we obtain: $\boldsymbol{\varepsilon}_{s+1} = \boldsymbol{f}{*}(\boldsymbol{f}{*}\boldsymbol{\varepsilon}_{s-1}+\boldsymbol{\nu}_s) + \boldsymbol{\nu}_{s+1}$. Since $\boldsymbol{f}*$ is a linear operator this is: $\boldsymbol{\varepsilon}_{s+1} = \boldsymbol{f}{*}\boldsymbol{f}{*}\boldsymbol{\varepsilon}_{s-1}+ \boldsymbol{f}{*}\boldsymbol{\nu}_s + \boldsymbol{\nu}_{s+1}$. And if we use $\boldsymbol{f}_{|j}$ for the p.m.f. that arises by convolving $\boldsymbol{f}$ with itself j times we can write: $\boldsymbol{\varepsilon}_{s+1} = \boldsymbol{f}_{|2}{*}\boldsymbol{\varepsilon}_{s-1} + \boldsymbol{f}_{|1}{*}\boldsymbol{\nu}_s + \boldsymbol{f}_{|0}{*}\boldsymbol{\nu}_{s+1}$. If we now replace

---

[1] Pickups are already refused in real-world busy routes whenever buses reach capacity.

[2] Although pickups could also be refused with schedule-based systems, interventions of this type would usually be late because drivers can only know their headways at the few control points on their routes. And if drivers were to be given the discretion to stop picking up passengers at any location only on the basis of their schedule delay (ignoring the bus they follow) there would be false alarms.

$\boldsymbol{\varepsilon}_{s-1}$ by its corresponding instance of (9b) and repeat this s times we obtain: $\boldsymbol{\varepsilon}_{s+1} = \boldsymbol{f}_{|s+1} * \boldsymbol{\varepsilon}_0 + \boldsymbol{f}_{|s} * \boldsymbol{\nu}_1 + \boldsymbol{f}_{|s-1} * \boldsymbol{\nu}_2 + \ldots + \boldsymbol{f}_{|1} * \boldsymbol{\nu}_s + \boldsymbol{f}_{|0} * \boldsymbol{\nu}_{s+1}$. And since $\boldsymbol{\varepsilon}_0 = \mathbf{0}$, we finally have:

$$\boldsymbol{\varepsilon}_{s+1} = \sum_{j=0}^{s} \boldsymbol{f}_{|j} * \boldsymbol{\nu}_{s+1-j} \qquad \text{for} \quad s = 0, 1, 2\ldots \qquad (11a)$$

In scalar notation, using $f_{m|j}$ for the m^{th} term of $\boldsymbol{f}_{|j}$ the expression is:

$$\varepsilon_{n,s+1} = \sum_{j=0}^{s} \sum_m f_{m\,|\,j}\, \nu_{n-m,s+1-j} \qquad \text{for} \quad n = 1, 2\ldots;\ s = 0, 1, 2,\ldots \qquad (11b)$$

Equations (11) are useful because they express our unknowns (the errors $\varepsilon_{n,s}$) as a linear combination of known random variables (the noise terms).

The coefficients of (11) can be calculated numerically, and can also be expressed analytically with transform methods. However, since the repeated convolution of a p.m.f expresses the p.m.f. of the sum of a corresponding number of i.i.d. random variables, closed forms for the Bernoulli, Poisson and negative binomial kernels can be readily written. More generally, however, the coefficients for each j should approach the normal distribution as j increases. So, if we write ϕ for the standard normal density function, we always have:

$$f_{m\,|\,j} \approx v_j^{-1}\, \phi(z_j/v_j) \qquad \text{for large } j, \quad \text{where } z_j = m - j\mu \ \text{ and } \ v_j^2 = jv^2. \qquad (12)$$

We are now ready to present the results.

PROPOSITION 1 (Stability)**:** *If* $|\,\nu_{n,s}\,| \le M \ \ \forall n, s$ *in the solution domain, then* $|\,\varepsilon_{n,s}\,| \le Ms \ \ \forall n, s.$

<u>*Proof*</u>*:* Taking absolute values in (11b) and using the triangle inequality we find that: $|\,\varepsilon_{n,s+1}\,| \le \left|\sum_{j=0}^{s} \sum_m f_{m\,|\,j}\, \nu_{n-j,s-j+1}\right| \le \sum_{j=0}^{s} \sum_m f_{m\,|\,j} \left|\nu_{n-j,s-j+1}\right| \le M \sum_{j=0}^{s} \sum_m f_{m\,|\,j} = M(s+1)$. □

Proposition 1 shows that the proposed control policy is stable and robust; i.e., bounded noise cannot produce unbounded errors no matter how many bus runs are introduced. But if the noise has a known covariance structure exact formulas for the variance of the errors can also be developed because (11) links linearly the errors and the noise.

If the segments are short so that the attraction forces do not have the opportunity to alter a bus' path noise terms should be uncorrelated. Of interest is the case where the noise terms are u.i.v. with $\sigma_s^2 = \sigma^2$. In this case (11) yields $\text{var}(\varepsilon_{n,s+1}) = \sum_{j=0}^{s} \sum_m f_{m\,|\,j}^2 \sigma^2$ and we see that the proposed policy amplifies the variance of the noise by a factor $k_{\varepsilon,s}^2 \equiv \text{var}(\varepsilon_{n,s+1})/\sigma^2$, which is

$$k_{\varepsilon,s}^2 = \sum_{j=0}^{s} \sum_m f_{m\,|\,j}^2 \ . \qquad (13)$$

Because (13) gives little insight and is tedious to calculate even in the simplest cases, a simplification is given below.

RESULT 1 (Variance of Arrival Deviations): *If the noise terms are u.i.v., then:*

$$k_{\varepsilon,s}^2 \approx \sqrt{\frac{s}{\pi v^2}} \qquad \text{if} \quad sv^2 >> 1. \tag{14}$$

Proof: For large s the contribution to (13) by terms with small j is small and the remaining terms can be approximated with (12). Therefore we shall use (12). If j is so large that $v_j >> 1$ we can also replace the inner sum $P_j \equiv \sum_m f_{m|j}^2$ by the integral $P_j \approx \int_{-\infty}^{+\infty} v_j^{-2} \phi^2\left(\frac{z_j}{v_j}\right) dz_j = (2\sqrt{\pi} v_j)^{-1} = (2\sqrt{\pi j v^2})^{-1}$. (The integral was solved using the substitution: $\phi(x)^2 = \phi(\sqrt{2}x)/\sqrt{2\pi}$). This approximation for P_j applies if $jv^2 >> 1$. It cannot be used for j=1 but improves with increasing j. Thus, if we use the approximation for j >1 only and recognize that P_0=1 we can write: $k_{\varepsilon,s}^2 = 1 + \sum_{j=1}^{s} P_j \approx 1 + (2\sqrt{\pi v^2})^{-1} \sum_{j=1}^{s} j^{-1/2} \approx 1 + (2\sqrt{\pi v^2})^{-1} \int_{0.5}^{s+0.5} j^{-1/2} dj \approx (\pi v^2/s)^{-1/2}$ for sufficiently large s, which matches (13). The result holds for $sv^2 >> 1$ and improves with increasing s because then the bulk of the contribution to the sum of the P_j's comes from the terms with large j which satisfy $jv^2 >> 1$ and are well approximated by the integral. □

A simulation of the Bernoulli model, i.e. where $v^2 = \alpha(1-\alpha)$, with s = 1, 2… 150 control points and several thousand consecutive bus runs shows that if $\alpha \in (0.1, 0.9)$ then (14) predicts $k_{\varepsilon,s}$ with errors below 7% for $s > 10$ and below 2% for $s > 30$. So, (14) can be used as a rough recipe to predict expected deviations from the schedule in practical applications. We now turn our attention to the headways.

3.2 Deviations from the ideal headway

We now show that if the noise is u.i.v. then the variances of the deviations in headway at every control point are bounded by a common quantity that is independent of s. So, even though according to (14) the deviations from the schedule can theoretically grow arbitrarily large for very long imaginary routes, we shall see that the headways cannot. In other words, the proposed policy keeps near-constant headways for all buses indefinitely.

To verify this idea we shall work with the headway deviations $\xi_{n,s} = \varepsilon_{n,s} - \varepsilon_{n-1,s}$, expressing them by subtracting two instances of (11b). After collecting terms with the same $\nu_{n,s}$'s we find:

$$\xi_{n,s+1} = \sum_{j=0}^{s} \sum_m (f_{m|j} - f_{m-1|j}) \nu_{n-m,s+1-j} \,. \tag{15}$$

Taking variances in (15) we see that the variance amplification is:

$$k_{h,s}^2 = \sum_{j=0}^{s} Q_j \quad \text{where} \quad Q_j = \sum_m \left(f_{m\,|\,j} - f_{m-1\,|\,j}\right)^2. \qquad (16)$$

This is an exact result. We used the subscript h instead of ξ because the variance of these deviations equals the variance of the headways. We now show that these variances have a common upper bound for all s.

PROPOSITION 2 (Headway Variance Bound): *Series* (16) *converges to a quantity* k_h^2 *that bounds the headway variance amplification at all stages s for all buses n.*

Proof: Note that (16) is monotonic; therefore to prove the theorem it suffices to show that $Q_j = O(j^{-c})$ for some $c > 1$. We shall show that $c = 3/2$. In view of (12), we can approximate ($p_{m|j}$ - $p_{m-1|j}$) for large j by the derivative of $v_j^{-1}\ \phi(z_j/v_j)$ with respect to z_j, which is: $-z_j v_j^{-3}\ \phi(z_j/v_j)$. Therefore, Q_j can also be approximated by the integral of the square of this quantity. This integrand can again be simplified with the substitution: $\phi(x)^2 = \phi(\sqrt{2}x)/\sqrt{2\pi}$. It then reduces to $(-z_j v_j^{-3}\ \phi(z_j/v_j))^2 = \frac{z_j^2}{v_j^6\sqrt{2\pi}}\phi(\sqrt{2}z_j/v_j)$. The resulting integral has the form for the variance of a zero-mean normal variable, so the final result turns out to be $Q_j \approx (4\sqrt{\pi}v_j^3)^{-1}$. Since this approximation improves for increasing j and $v_j^2 = jv^2$, we conclude that $Q_j \approx (4\sqrt{\pi})^{-1}(jv^2)^{-3/2} = O(j^{-3/2})$. □

An approximate expression for k_h^2 can be obtained by adding the asymptotic expression for Q_j derived in the proof of this theorem, $Q_j \approx (4\sqrt{\pi})^{-1}(jv^2)^{-3/2}$, but this result turns out to be quite poor because the main contribution to $k_{h,s}^2$ comes from the first few terms of (16) which are not well approximated by the asymptotic expression. The expression does suggest, however, that kernels with large v (e.g., involving several headways) are more effective in smoothing bus flow than those with small v. Simulations bear this out.

We examined with simulation the Bernoulli kernel in detail because it is the simplest to implement in practice and because it can be used as a point of reference. We fitted to the data power functions of $v = \sqrt{\alpha(1-\alpha)}$ and found the following:

RESULT 2 (Headways of the Bernoulli Kernel): *The headways are approximately Gaussian with:*

$$k_h \approx 0.95[\alpha(1-\alpha)]^{-1/2}. \qquad (17a)$$

The error is below 1% for $\alpha \in (0.1, 0.9)$ *and below 2% for* $\alpha \in (0.03, 0.97)$. Furthermore:

$$k_h < [\alpha(1-\alpha)]^{-1/2} \quad \textit{for } \alpha \in (0.01, 0.99). \qquad (17a)$$

Table II summarizes selected results from the simulation, including three non-Bernoulli kernels. In addition to the variance of the headways, it displays at how many segments away from the boundary equilibrium is reached, and the required slack per segment assuming that β is small enough to be neglected. (The slack d_s was calculated by making sure that the fluctuations in (6) and (10) rarely gave rise to negative values.) Note how the simulated headway variances for the three Bernoulli cases match (17a), and how these variances are reduced by nearly 40% by non-Bernoulli kernels that require equal or less slack.

Table II: Selected simulation results for different kernels; the unit of time is σ.

Kernel: $f_0, f_1, f_2 \ldots$	Headway variance	Number of segments to equilibrium	Slack per segment
.5, .5 (Bernoulli)	3.8	7	2.9
.8, .2 (Bernoulli)	5.6	9	1.4
.9, .1 (Bernoulli)	10.5	30	1
.4, .2, .2, .2	2.35	2	2.4
.7, .1, .1, .1	3.5	3	1.4
.85, .05, .05, .05	6.4	7	.9

4. DISCUSSION

We present below an example that illustrates both, how to apply the results of Sec. 3, and the type of practical benefits that can be expected. We then briefly discuss some implementation issues and future work.

4.1 Example

Assume that we want to provide frequent service with H = 5 min on a homogeneous bus line. In the interest of resiliency all its segments are to be 1 km long. On each of these segments the attraction parameter is β = 0.03 and the uncontrolled travel time averages c_s = 3 min with a standard deviation, σ = 0.25 min. These times are uncorrelated.

With the Bernoulli kernel, the proposed strategy produces Gaussian headways with standard deviation $\sigma_h \approx 0.95\sigma[\alpha(1-\alpha)]^{-1/2}$ and requires $d = 3(\alpha+0.03)\sigma_h$ as slack for each segment. For $\alpha = 0.2$ we find $\sigma_h \approx$ 36s and $d \approx$ 25s; and for $\alpha = 0.1$, $\sigma_h \approx$ 47s and $d \approx$ 19s. Since the fluctuations in headway are small compared with the headway, buses would not pair up.

Let us now look at this from the perspective of a randomly arriving passenger. It is well known that the headway fluctuations add $\frac{1}{2}\sigma_h^2/H$ time units to the passenger's average out-of vehicle delay; i.e., only about 2s if $\alpha = 0.2$ and 4s if $\alpha = 0.1$. The passenger's in-vehicle delay due to the slack is more severe. A 5km trip, which would take 15min in a perfect deterministic world, would take about 17min if $\alpha = 0.2$ and 16.5min if $\alpha = 0.1$. These trip times could be slightly reduced by using non-Bernoulli kernels, and also by spacing the control points more widely. The latter is not recommended, however, for the resiliency reasons mentioned at the outset of this paper.

Now, compare this performance with the schedule-based approach. Since noise disturbances (unlike headway disturbances) are skewed, the schedule-based approach would at least require a slack of $4\sigma = 1$ min. Although it would produce headways with $\sigma_h^2 \approx \sigma^2$ away from the control points, resulting in negligible out-of-vehicle delays, our 5 km/15 min trip would take 20 min instead of 16.5 or 17.

Even if we were to combine segments in order to reduce the number of control points (giving up resiliency) the schedule-based approach falls short. For example combining 10 segments, which would increase the noise variance σ^2 by a factor of about 14 (since the variance is amplified by a factor of 10 due to the effect of length and by a factor of about 1.2^2 due to the effect of β, see Table I), would result in passenger delays averaging about 4s outside the vehicles and about 23s/km inside. This performance is still inferior to that of the Bernoulli model with $\alpha = 0.1$ and could be improved further by more sophisticated kernels such as the one on the last row of Table II.

The comparative advantage of headway control declines but is still significant for systems with strong attraction forces. Table III, below, summarizes the results one obtains if the above calculations are repeated for systems with $\beta = 0.1$ (busy) and $\beta = 0.3$ (very busy). The values of s on the fist column indicate the number of segments combined for schedule-based control. Note that buses travel faster with headway control than headway control if we do not combine segments: by 28 s/km when $\beta = 0.1$ and 6s/km when $\beta = 0.3$. And more could be gained with non-Bernoulli kernels.

If we combine segments the schedule-based approach improves, losing by 7s/km to headway control when $\beta = 0.1$ and winning by 5s/km when $\beta = 0.3$. This, of course, comes at the cost of resiliency. But if

Table III: Performance of different control methods under different demand scenarios. Added wait is the extra out-of-vehicle waiting time resulting from the variability in the headways. Added pace is the bus delay per kilometer caused by the slack. Results apply for σ =1/4 min. To obtain results for different σ's, scale the values of the table up or down by the same factor as σ.

Control method		Attraction parameter β = 0.1			Attraction parameter β = 0.3	
		Added wait (sec)	Added pace (sec/km)		Added wait (sec)	Added pace (sec/km)
.5, .5 (Bernoulli)		1	50		1	68
.8, .2 (Bernoulli)		2	32		2	54
.9, .1 (Bernoulli)		4	28		4	58
s–1 = 0 (schedule-based)		0	60		0	60
s–1 = 1 (schedule-based)		0	46		0	51
s–1 = 2 (schedule-based)		1	38		1	49
s–1 = 4 (schedule-based)		1	35		1	64
s–1 = 8 (schedule-based)		3	36		3	192
s–1 = 16 (schedule-based)		6	65		buses would bunch	> 10^3

resiliency is not an issue, segments can also be combined for headway control. Pairing segments for example would reduce the bus delay to about 12s/km for β = 0.1 and about 40s/km for β = 0.3. These outcomes allow buses to travel faster than with schedule control. Bus speed is important because faster buses are more productive buses and this can benefit the transit agency; not just its customers.

4.2 Conclusion

The above example illustrates that transit agencies can retain reasonably fast bus speeds while closely tracking and controlling their buses. This is very beneficial.

By closely tracking buses over short segments the control system can automatically trigger corrective measures before problems grow large. For example if the bus on run n were to malfunction and go out of service, the transit agency could reassign its run to the following bus, which from then on would follow the bus on run n–1. The target headway of the reassigned bus, and those of a few others on succeeding runs, would then be increased to spread their loads; e.g., by 25% if four buses were involved. Some of these buses would have to be temporarily accelerated but they can do this by skipping pickups. The nice thing about the proposed strategy and its short segments is that this correction would take effect quickly, upon the buses arrivals to the problematic control point. With 1 km segments this would take just a few minutes -- not a large fraction of an hour – so that the remedial measure would have an excellent chance to contain the damage. Although it is ideal for systems with frequent service, headway control can also be helpful for scheduled systems with long headways as a fail-safe operating mode when large jams or storms disrupt the complete system.

Headway control can be implemented easily. In its most rudimentary form, perhaps appropriate for a pilot study, one person would be stationed at each control point with the responsibility to calculate (6) or (10) upon each bus arrival and then postpone the bus' departure accordingly. This system can be easily automated with computer-controlled signals. Alternatively, one could use on board computers equipped with GPS and wireless communication devices. This would also allow the transit agency to monitor its bus routes even more closely, improve communication and guidance to drivers, and reduce cost.

This type of on-board architecture would also enable the use of more advanced bus cooperation strategies involving leaders and followers, which for example would allow a bus to slow down when it is running ahead of the bus behind. This flexibility has the potential for speeding up bus service even more. The architecture can also be used to control buses on closed loop routes where buses may be introduced and taken out of service during the course of a day. In this case, the goal is maximizing the commercial speed of the buses in circulation while maintaining regular headways. This problem is mathematically more difficult because the headway is now an endogenous variable: the problem's dynamic equations are non-linear and have to be approximated. Further research to build and demonstrate on-board system architectures including applications of the types just described is under way.

ACKNOWLEDGEMENT

This research was supported by the U.C. Berkeley Center of Excellence on Future Urban Transport. The comments of my colleagues M. Cassidy and Y. Li are gratefully acknowledged.

REFERENCES

Newell, G.F. (1977). Unstable Brownian motion of a bus trip. Statistical Mechanics and Statistical Methods in Theory and Applications (ed. U. Landman}, Plenum Press, 645-667.