Arterial Traffic Signal Optimization: A Person-based Approach

Eleni Christofa[a,*], Konstantinos Ampountolas[b], Alexander Skabardonis[c]

[a] *Department of Civil and Environmental Engineering, University of Massachusetts, Amherst, MA 01003*
[b] *School of Engineering, University of Glasgow, Glasgow G12 8LT, United Kingdom*
[c] *Institute of Transportation Studies, University of California, Berkeley, CA 94720*

Abstract

This paper presents a real-time signal control system that optimizes signal settings based on minimization of person delay on arterials. The system's underlying mixed integer linear program minimizes person delay by explicitly accounting for the passenger occupancy of autos and transit vehicles. This way it can provide signal priority to transit vehicles in an efficient way even when they travel in conflicting directions. Furthermore, it recognizes the importance of schedule adherence for reliable transit operations and accounts for it by assigning an additional weighting factor on transit delays. This introduces another criterion for resolving the issue of assigning priority to conflicting transit routes. At the same time, the system maintains auto vehicle progression by introducing the appropriate delays associated with interruptions of platoons. In addition to the fact that it utilizes readily available technologies to obtain the input for the optimization, the system's feasibility in real-world settings is enhanced by its low computation time. The proposed signal control system is tested on a four-intersection segment of San Pablo Avenue arterial located in Berkeley, California. The findings show the system's capability to outperform pretimed (i.e., fixed-time) optimal signal settings by reducing total person delay. They have also demonstrated its success in reducing bus person delay by efficiently providing priority to transit vehicles even when they are traveling in conflicting directions.

Keywords: Person-based traffic signal control, Transit signal priority, Person delay, Mixed-Integer Linear Programming, Conflicting Transit Routes, Pairwise signal optimization

*Corresponding author. Tel.: +1 (413) 577-3016
Email addresses: christofa@ecs.umass.edu (Eleni Christofa), konstantinos.ampountolas@glasgow.ac.uk (Konstantinos Ampountolas), skabardonis@ce.berkeley.edu (Alexander Skabardonis)

1. Introduction

With the continuous growth of population and car ownership and the limited funds available, there is an imperative need to design and manage multimodal transportation systems more efficiently while improving the use of existing infrastructure. With traffic signal control systems already widely deployed in urban street networks, one of the most cost-effective ways to improve efficiency and sustainability of urban transportation systems is to develop signal control strategies that enhance person mobility. This can be achieved with the development of signal control strategies that in addition to resolving conflicts between vehicles, give preferential treatment to high occupancy transit vehicles while accounting for the overall traffic conditions in the network.

Several advanced real-time signal control systems have incorporated transit signal priority strategies in their algorithms in order to manage multimodal systems more efficiently. However, very few systems are optimizing signal settings by explicitly minimizing person delay in a network. On the contrary, they usually minimize vehicle delays (Cornwell et al., 1986; Hunt et al., 1982; Bretherton et al., 2002) and provide priority based on rules that are not directly included in the optimization process (Conrad et al., 1998; Diakaki et al., 2003) or pre-select a subset of transit vehicles to apply their priority strategies (Mauro and Di Taranto, 1989; Henry and Farges, 1994). Some systems have optimized signal settings by minimizing some weighted combination of passenger delay, car delay, bus delay, and bus schedule delay (Chang et al., 1996; Vasudevan, 2005; Li et al., 2008; Stevanovic et al., 2008; Ma et al., 2013a) and others by reducing bus travel time or passenger waiting time at the downstream bus stop while minimizing the impact these priority strategies have on the rest of the traffic (Ma et al., 2013b; Lin et al., 2013; Zeng et al., 2014).

Recently, several real-time signal control systems that take advantage of data from Connected Vehicles (CV) (i.e., vehicle-to-vehicle and vehicle-to-infrastructure communications) have been developed. These systems have attempted to optimize signal timings for signalized arterials based on weighted functions of delays for all users, accounting for priority requests of buses and/or pedestrians and maintaining coordination of traffic signals. However, they either require high penetration of probe vehicles (for different modes) to be successful and in some cases have high computation times that constrain their applicability in real-world settings (He et al., 2012) or assume a background offline optimized plan (He et al., 2014).

A person-based traffic-responsive signal control system for isolated intersections was recently

proposed in Christofa and Skabardonis (2011) and Christofa et al. (2013). This system minimizes person delay by explicitly accounting for the passenger occupancy of autos and transit vehicles. This results in provision of signal priority to transit vehicles and introduces an efficient way for resolving the issue of priority assignment when transit vehicles travel in conflicting directions. The system uses real-time information that can be obtained form currently deployable surveillance and communication technologies (i.e., vehicle detectors, Automated Vehicle Location (AVL) and Automated Passenger Counter (APC) systems). A few other systems that minimize person delay at signalized intersections and arterials have been proposed since. Sun et al. (2015) developed a real-time signal control system similar to the one (Christofa et al., 2013) developed that minimizes total person delay at isolated intersections under the assumption of CV data availability. Availability of CV data allowed them to estimate vehicle delays individually avoiding second order terms in their mathematical program. However, this system was restricted to isolated intersections, and did not account for transit schedule adherence in the objective function. Another real-time signal control system that minimizes person delay for all users at consecutive signalized intersections with the use of CV data was also recently proposed (Hu et al., 2015). The system was tested only for a two-intersection arterial segment and under the assumption of a maximum of one bus priority provision per cycle. Under its current form, the system cannot be used for cases with multiple conflicting bus lines and multiple priority requests, while the current computation time could prohibit its implementation in real-world signalized arterials with multiple intersections.

Overall, most existing systems ignore the case of multiple priority requests (i.e., multiple transit lines) basing their decisions on pre-selected priority for the buses or treating them on a first come first served basis. Therefore, they lack an efficient way of assigning priority to transit vehicles especially when they are traveling on conflicting routes. Furthermore, they often ignore the importance of transit schedule adherence in providing priority, which in some cases can cause further disruptions to the transit system. Finally, recent studies that have addressed some of those issues can either not be implemented in real-world settings given their existing data requirements and high computation times or are restricted to optimize signal settings at isolated intersections.

This paper presents an extension of the traffic responsive signal control system previously published by Christofa and Skabardonis (2011) and Christofa et al. (2013). The system is extended to arterials that are characterized by multiple transit lines traveling in conflicting directions and platooned vehicle arrivals. In addition to accounting for auto vehicle progression, by assigning

the appropriate delays for interrupting the platoons, the system recognizes the importance of schedule adherence for reliable transit operations. Therefore, it introduces an additional weighting factor that reflects how early or late a transit vehicle is when arriving at an intersection and assigns priority accordingly facilitating priority assignment decisions when transit vehicles travel in conflicting directions. Another advantage of the system is its low computation time due to its mathematical formulation and pairwise signal optimization as well as the use of readily available data both of which are promising for real-time applications.

The proposed signal control system is tested on a four-intersection segment of San Pablo Avenue arterial located in Berkeley, California, and the results are compared against the performance of optimal fixed-time signal settings obtained from TRANSYT-7F (Hale, 2009). TRANSYT-7F is a state-of-the-art offline traffic signal optimization software that is extensively used in the U.S. and Europe. It is flexible in that it allows the user to choose the objective function from a variety of available functions. In addition, the user can choose from two optimization techniques (hill-climb and genetic algorithm), and the software is able to optimize all signal settings (i.e., cycle length, phasing sequence, splits, and offsets). Finally, it can handle both pretimed and actuated control.

The rest of the paper is organized as follows: Section 2 is dedicated to the proposed person-based optimization approach and the underlying mathematical model. In Section 3, a case study for a four-intersection arterial in Berkeley, California illustrates the performance and effectiveness of the proposed person-based approach under deterministic and stochastic arrivals. Finally, Section 4 discusses the findings of this work and outlines areas for future research.

2. Mathematical Model

The optimization of signal settings for an arterial is based on a pairwise optimization strategy introduced by Newell (1964, 1967). According to this strategy signal timings are optimized for a pair of consecutive intersections. Therefore, the mathematical program is formulated to minimize the total person delay at two consecutive intersections, r and $r + 1$, for all vehicles that are present during a design cycle T. The optimization process starts by determining the critical intersection of the subject arterial, which could be defined as the one with the highest intersection flow ratio, (i.e., demand flow to saturation flow ratio for the critical lane groups of an intersection), highest flow ratio in the direction of interest or highest transit traffic. Starting with the critical intersection, some level of vehicle progression through signal coordination is maintained for the

heaviest direction of traffic on the arterial. This means that the phase that serves the heaviest direction is designated as the coordinated one on all intersections. This progression is achieved by incorporating the appropriate delays for stopping the head or the tail of the platoon of all links leading to the two intersections of interest in the objective function. While this does not guarantee a green wave (i.e., that no vehicle traveling from one intersection to another will have to stop), it allows for some level of progression. Adjusting the weighting factors for those delays caused by stopping the head or tail of the platoon can be used to achieve various levels of auto vehicle progression.

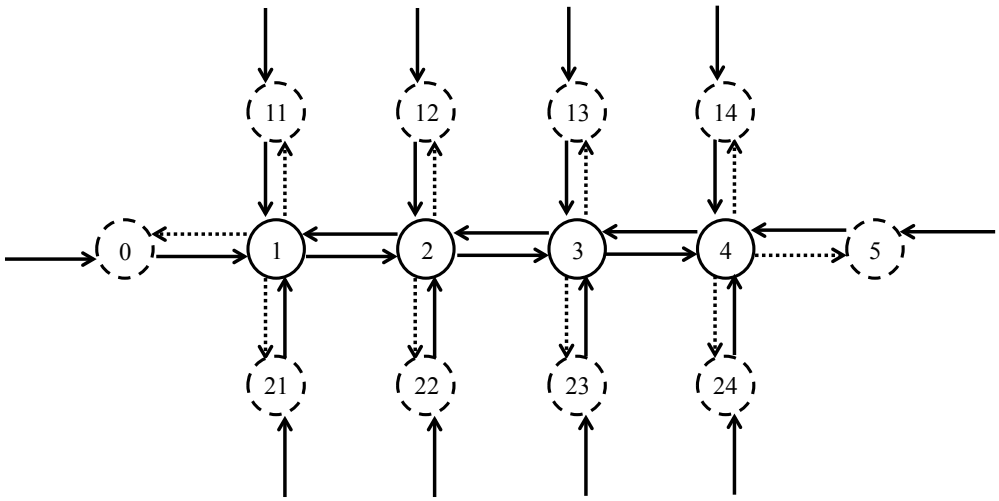
Once the signal settings for the first two intersections, r and $r+1$, are optimized, the next pair, $r+1$ and $r+2$, will be optimized. For this optimization, the beginning of green for the coordinated phase (i.e., phase that serves the heaviest direction) at $r+1$ will be constrained by the optimization outcome of r and $r+1$. This constraint ensures that the beginning of the green for that phase will be held constant when optimizing the second pair of intersections. Assuming that the yellow times (and all-red times, if any) are constant, this can be expressed as:

$$\sum_{i=1}^{c^{r+1}-1} g_{i,T}^{r+1}(p+1) = \sum_{i=1}^{c^{r+1}-1} g_{i,T}^{r+1}(p) \tag{1}$$

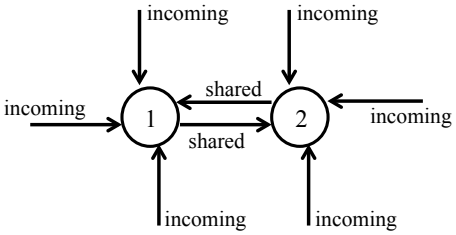
where c^{r+1} is the phase that serves the heaviest direction at intersection $r+1$, $g_{i,T}^{r+1}(p)$ is the optimal green times for phase i during cycle T at intersection $r+1$ obtained from the optimization of the pair of intersections p, and $g_{i,T}^{r+1}(p+1)$ is the corresponding green time for phase i obtained from the optimization of the second pair of intersections $p+1$ in which intersection $r+1$ belongs. The yellow time intervals preceding the coordinated phase are excluded from this equation under the assumption that they remain constant. The same pairwise optimization is repeated in the direction of interest until all intersections in the subject arterial are optimized. An illustrative example is shown in Figure 1.

In case it is not clear which direction has the heaviest traffic, the same process can be repeated in the opposing traffic direction, and the signal settings that give the lowest total person delay can be chosen. This is particularly easy to do in practice because the mathematical program can be solved very quickly (as explained at the end of Section 2.2.3). As a result, both optimizations can be performed fast enough for real-world implementations.

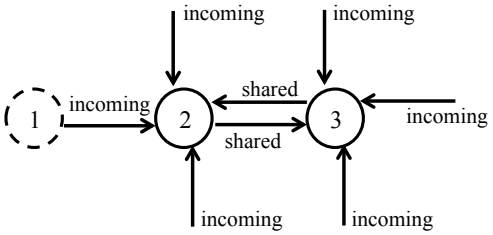
The mathematical program that minimizes total person delay at two consecutive intersections is formulated under the assumption of perfect information on traffic and transit arrival, passenger



(a) Signalized Arterial Corridor



(b) Optimization of Intersections 1 and 2



(c) Optimization of Intersections 2 and 3

Figure 1: Pairwise arterial signal optimization.

occupancies, and lane capacities, which can vary from cycle to cycle. It is also based on the assumption of fixed phase sequence. While these assumptions may seem restrictive, relaxing them does not affect the formulation of the mathematical program since it allows for the use of different values for lane capacities, traffic demands, and vehicle passenger occupancies from cycle to cycle. In fact, the stochastic arrival tests presented in Section 4.2 have been performed with the use of traffic demands that exhibit stochastic variations as we would expect in more realistic situations, so

the proposed system is demonstrated without the restriction that demand must be deterministic and constant. Furthermore, vehicles arrive at each intersection in platoons when traveling on the arterial and on the cross-street links, since the subject arterial is considered to be part of a larger arterial signalized network of arterials. In case there are reasons to assume that the cross streets experience uniform arrivals, the delay equations for those cross-street links can be updated based on the delay equations presented in Christofa et al. (2013) for isolated intersections. It is also assumed that there is negligible platoon dispersion. The cycle length is kept constant for the analysis period and it is common for all intersections along the arterial to maintain signal coordination, although may be updated in real-time by another parallel algorithm, e.g., Diakaki et al. (2003). Finally, the model is formulated assuming that transit vehicles travel on mixed-use traffic lanes along with autos. However, the formulation of the mathematical model holds even when dedicated lanes for transit vehicles exist.

To generalize the optimization process above, consider a signalized arterial, x, consisting of intersections $r \in R$, where R is the set of specified arterial intersections. The intersections r that belong to a specific arterial x form an ordered list $\Lambda_x = \{1, 2, \ldots, r-1, r, r+1, \ldots\}$. According to this, for each intersection $r \in R$ that belongs to an arterial x, an element r exists in the list as well as the elements $r-1, r+1 \in \Lambda_x$ that represent the intersections of the arterial x that are located upstream and downstream of intersection r, respectively, if any. The generalized formulation of the mathematical program that minimizes person delay for consecutive pairs of intersections $(r, r+1) \in \Lambda_x$ and for a cycle T is as follows:

$$\min \quad \mathcal{L}(r, r+1) = \sum_{(r,r+1)\in\Lambda_x} \left[\sum_{a=1}^{A_T^r} o_a d_{a,T}^r + \sum_{b=1}^{B_T^r} o_{b,T}^r (1 + \delta_{b,T}^r) d_{b,T}^r \right] \tag{2}$$

$$\text{s.t.} \quad d_{a,T}^r = d_a^r \left(g_{i,T}^r \right) \tag{3}$$

$$d_{b,T}^r = d_b^r \left(g_{i,T}^r \right) \tag{4}$$

$$g_{i,T}^r \geq g_{i,\min}^r \tag{5}$$

$$\sum_{i=1}^{I^r} g_{i,T}^r + L^r = C \tag{6}$$

where:

r: intersection index

a: auto vehicle index

b: transit vehicle index

A_T^r: total number of autos present at intersection r during cycle T

B_T^r: total number of transit vehicles present at intersection r during cycle T

o_a: passenger occupancy of auto a $[\frac{\text{pax}}{\text{veh}}]$

$o_{b,T}^r$: passenger occupancy of transit vehicle b for cycle T at intersection r $[\frac{\text{pax}}{\text{veh}}]$

$d_{a,T}^r$: delay for auto a for cycle T at intersection r [sec]

$d_{b,T}^r$: delay for transit vehicle b for cycle T at intersection r [sec]

$\delta_{b,T}^r$: factor determining the weight for schedule delay of transit vehicle b for cycle T at intersection r

$d_a^r\left(g_{i,T}^r\right)$: function relating the delay for auto a to green times

$d_b^r\left(g_{i,T}^r\right)$: function relating the delay for transit vehicle b to green times

$g_{i,T}^r$: green time allocated to phase i in cycle T at intersection r [sec]

$g_{i,\min}^r$: minimum green time for phase i at intersection r [sec]

I^r: total number of phases in a cycle for intersection r

L^r: total lost time at intersection r [sec]

C: cycle length [sec].

The mathematical program is run once for every cycle T for each pair of intersections and the objective function (2) consists of the sum of the delay for all auto and transit passengers that are present at the intersection during that cycle T. Delays for autos, $d_{a,T}^r$, and transit vehicles, $d_{b,T}^r$, depend on the green times, $g_{i,T}^r$, which are the decision variables of the mathematical program. In fact, $d_{a,T}^r$ and $d_{b,T}^r$ also depend on the green times of the previous and next cycles, the yellow times, and the traffic demand. These are either pre-specified by the user, based on historical data, or collected with the use of surveillance technologies. To simplify the notation, the delays for each platoon, lane group, and transit vehicle are included in the objective function as a variable and this variable is constrained to equal a function as shown in (3) and (4). The optimal green times also determine the beginning time of the coordinated phase, which specifies the offset between two consecutive intersections. As a result, offsets are effectively optimized within the mathematical program in those cases that the coordinated phase is not the first one in the cycle. Otherwise, offsets cannot be changed through the optimization of the green times, because the beginning of

the cycle for each of the intersections is fixed and the cycle length remains constant. In those cases, in order to maintain progression in a selected direction, the offset between successive intersections, which in this case corresponds to the elapsed time between the beginning of a cycle at the two consecutive intersections, is set equal to the average free flow travel time or is equal to the optimal offset obtained by a different optimization algorithm (e.g., TRANSYT-7F).

The delays of both autos and transit vehicles are weighted by their respective passenger occupancies in the objective function (2). Transit vehicle passenger occupancies can be obtained through currently deployable technologies such as APC systems. Auto passenger occupancies will also be available in the near future with the introduction of CVs. Such technologies can provide real-time information on passenger occupancies that can lead to a dynamic change of weights for the different delay components. Note that the mathematical program formulation is generic in that it can accommodate such dynamic passenger occupancies without any need for further modifications since passenger occupancies are inputs to the mathematical model and can be updated at every cycle. In addition, the delays for transit vehicles are weighted by a factor $(1 + \delta^r_{b,T})$ to account for the schedule delay that a transit vehicle b has when arriving at intersection r during cycle T. This factor, $\delta^r_{b,T}$, which is user-specified, can be a linear function of the schedule delay of the transit vehicle or a binary variable indicating whether a transit vehicle is ahead or behind schedule based on a predetermined threshold. In either case, the delay for a transit vehicle that is behind schedule is weighted more than a transit vehicle that is arriving early or on time at the intersection.

Three constraints are introduced for the decision variables in the mathematical program. The green times of each phase i and intersection r are constrained by their minimum green times in (5). Minimum green times, $g^r_{i\,\min}$, are necessary to ensure safe vehicle and pedestrian crossings and guarantee that no phase is skipped. The phase green times are also constrained such that the sum of the green times for all phases at each intersection plus the total lost time adds up to the cycle length (see constraint (6)). The total lost time is assumed to be the summation of the yellow times plus the all-red clearance intervals, if any per phase. The cycle length is kept constant for every cycle in the analysis period and is common among intersections in the network under consideration. Keeping the cycle length constant is not essential for the formulation. In fact, variable cycle lengths could result in lower person delays at the intersection. However, the constraint of constant and common cycle lengths simplifies the formulation of the mathematical

program and facilitates progression of auto traffic at the arterial level. In cases that the cycle length is selected by time of day, it still remains common among all intersections in the network at any time.

The mathematical program expressed with (2)–(6) is initially solved for the first pair of intersections that belongs to a specific ordered list Λ_x for an arterial x. After the first pair is optimized the additional constraint in (1) is added for each subsequent pair of intersections in Λ_x to ensure that the optimal decision from the optimization of the previous pair is accounted for in the optimization of the next one. The delay estimation for both autos and transit vehicles are explained in detail in the next sections.

2.1. Auto Delay

For each pair of intersections $(r, r+1) \in \Lambda_x$, the auto delays that contribute to the objective function of the optimization consist of three terms: 1) the delay experienced by vehicles that travel in platoons on incoming links during cycle T, 2) the delay experienced by vehicles that travel on shared links during cycle T, and 3) the delay experienced by vehicles that did not get served during the previous cycle, which constitute the residual queues at the approaches of the two intersections during cycle T. This means that during cycle T a platoon could be experiencing delay while traveling on the incoming link (approaching the first intersection of the arterial it arrives at) and a portion of that platoon that continues in the subject network could be experiencing delay while traveling on the shared link (approaching the second intersection of the arterial it arrives at)[1]. These two delay components ensure that the effect of disrupting progression is accounted for in both directions. While this does not guarantee a green wave (i.e., that no vehicle traveling from one intersection to another will have to stop), since the objective of the mathematical program is to minimize person delay, it allows for some level of progression. Also, note that by including these three components of delay the proposed signal control system can be used to optimize signal timings for any type of traffic conditions (including undersaturated and transient oversaturated conditions) as long as they do not lead to queue spillbacks.

Since the optimization is conducted using a pairwise approach, the delays are calculated for each pair of intersections $(r, r+1) \in \Lambda_x$ in order to optimize the signal settings for that pair. Note

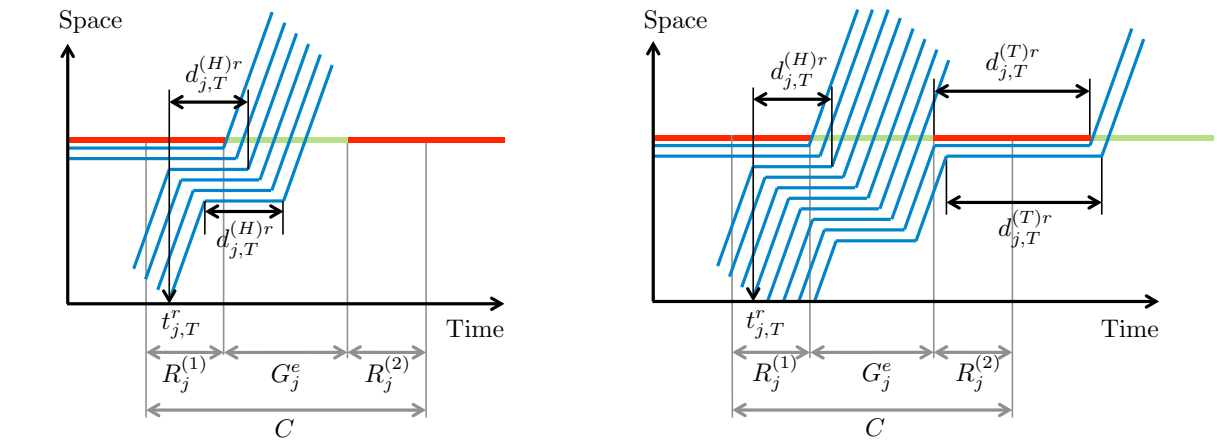
[1]For a visual explanation of the shared and incoming link for a pair of intersections you can refer to Figures 1(b) and 1(c)

that there is a symmetry in the formulas for the delays of vehicles traveling in the direction of progression and those traveling in the opposing direction. Suppose that r is the first intersection of the pair being optimized, and the optimization sequence is r and then $r+1$, which is the second intersection of the pair. For any platoon, the first intersection at which it arrives is denoted by u, and the second intersection at which a portion of it arrives is denoted by v. This means that for a platoon traveling in the direction of progression $u = r$ and $v = r+1$, while for a platoon traveling in the opposing direction $u = r+1$ and $v = r$. The same holds for transit vehicles. This notation is used for the remainder of the paper for both auto and transit delays.

The auto delays are estimated based on the assumption that vehicles arrive and are served at both intersections at capacity since they travel in platoons with no dispersion. Consequently, assuming that kinematic wave theory (Lighthill and Whitham, 1955; Richards, 1956) holds, all vehicle trajectories are parallel at all times, as shown in Figure 2. This means that the last vehicle in a platoon that is stopped will experience the same delay as the first vehicle in the same platoon that gets stopped. So, the collective delay for all vehicles can be easily estimated knowing only the arrival time of the first vehicle in a platoon at the back of its lane group's queue at intersection r, $t^r_{j,T}$, the size of that platoon, $P^r_{j,T}$, and the traffic conditions at the approach as expressed by the size of the residual queue of lane group j at the end of the previous cycle $T-1$, $N^r_{j,T-1}$.

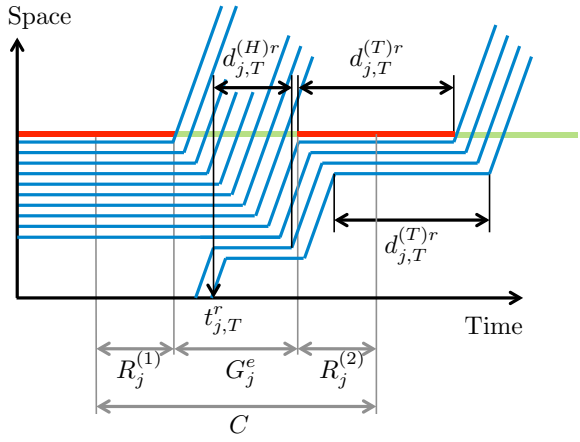
The estimation of the auto delay components of the objective function is presented next. Note that for simplicity and reduced computation time, all equations are formulated assuming that there is only one platoon per cycle per lane group. However, the algorithm can be easily extended to include multiple platoons in a cycle for the same lane group as long as the arrival times and sizes of those platoons are known. This arrival information could be available from Connected Vehicle data.

To further facilitate the notation of the mathematical program, the cycle time for each lane group can be split into three components, which are functions of the green times for each phase. The first is the component of the red time from the start of the cycle to the beginning of the green for the subject lane group, $R^{(1)r}_j(g^r_{i,T})$, the second is the duration of the effective green time itself, $G^{e,r}_j(g^r_{i,T})$, and the third is the component of the red time from the end of the green until the end of the cycle, $R^{(2)r}_j(g^r_{i,T})$. These values are illustrated in Figure 2 and are calculated as follows:

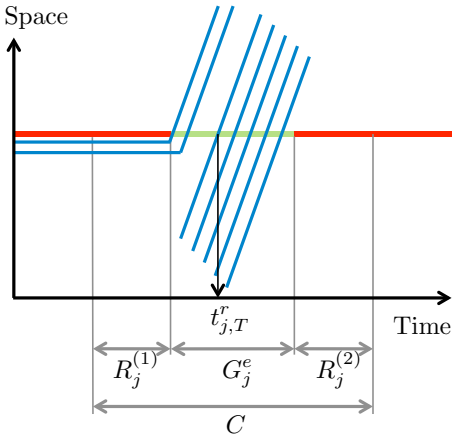


(a) Case P1

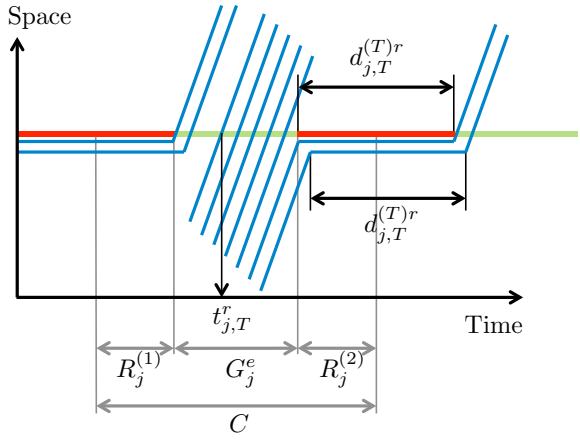
(b) Case P2



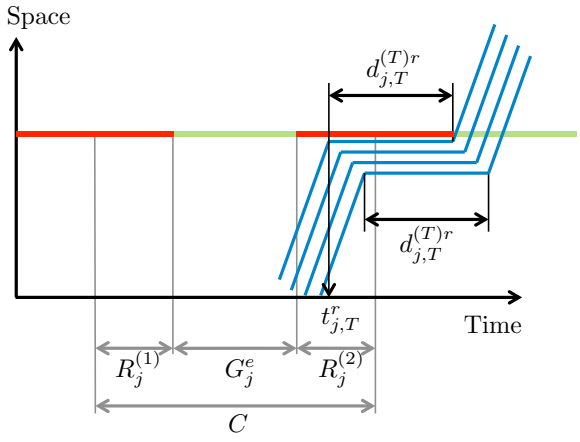
(c) Case P3



(d) Case P4



(e) Case P5



(f) Case P6

Figure 2: Auto delay estimation for platoon arrivals.

$$R_j^{(1)r}(g_{i,T}^r) = \sum_{i=1}^{k_j^r-1} g_{i,T}^r + \sum_{i=1}^{k_j^r-1} y_i^r \tag{7}$$

$$G_j^{e,r}(g_{i,T}^r) = \sum_{i=k_j^r}^{l_j^r} g_{i,T}^r + \sum_{i=k_j^r}^{l_j^r-1} y_i^r \tag{8}$$

$$R_j^{(2)r}(g_{i,T}^r) = \sum_{i=l_j^r+1}^{I^r} g_{i,T}^r + \sum_{i=l_j^r}^{I^r} y_i^r \tag{9}$$

where y_i^r is the yellow (and all-red, if any) time interval after phase i at intersection r, $g_{i,T}^r$ is the green time for phase i in cycle T at intersection r, k_j^r is the first phase in a cycle that can serve lane group j at intersection r, and l_j^r is the last phase in a cycle that can serve lane group j at intersection r. Note that in order to simplify the illustration in Figure 2 the yellow time intervals have not been marked on the time axes, but are considered to be included at the end of each phase.

2.1.1. Auto Delay for Vehicles in Platoons on Incoming Links

The auto delay for vehicles in platoons that travel on incoming links to an intersection u consists of two components: 1) the delay caused by stopping the head of the platoon, $D_{j,T}^{(H)u}$, which includes any delay incurred before the vehicles start being served and 2) the delay caused by stopping the tail of the platoon, $D_{j,T}^{(T)u}$, which corresponds to the delay experienced by vehicles after the platoon starts being served or after the end of the green phase that serves it (whichever comes first) and until the beginning of the green phase that serves it in the next cycle. Based on the arrival time of a platoon at the back of its lane group's queue, $t_{j,T}^u$, the size of the platoon, $P_{j,T}^u$, and the traffic conditions at the intersection, the delay for autos in that platoon can be estimated as in one of the following six cases:

- Case P1: Arrival before residual queue served, entire platoon served in green

  A platoon of size $P_{j,T}^u$ that belongs to lane group j of intersection u arrives at the back of its lane group's queue during cycle T at time $t_{j,T}^u$ before the time that the corresponding residual queue of j from the previous cycle $T-1$, $N_{j,T-1}^u$, would have finished being served if there was enough green time available. There is enough available green time to serve the residual queue, and spare green time to serve all $P_{j,T}^u$ vehicles in the platoon. These conditions are

summarized as:

$$t^u_{j,T} \leq t^u_T + R^{(1)u}_j(g^u_{i,T}) + \frac{N^u_{j,T-1}}{s^u_j} \tag{10}$$

$$N^u_{j,T-1} \leq G^{e,u}_j(g^u_{i,T}) s^u_j \tag{11}$$

$$P^u_{j,T} \leq G^{e,u}_j(g^u_{i,T}) s^u_j - N^u_{j,T-1} \tag{12}$$

where s^u_j is the saturation flow for lane group j at intersection u and t^u_T is the beginning of cycle T for intersection u, which is determined as:

$$t^u_T = t_T + O^u_T. \tag{13}$$

$t_T = (T-1)C$ is the beginning of cycle T at the critical intersection which is the first one to be optimized and O^u_T is the difference between the starting time of cycle T at intersection u and the critical intersection.

All vehicles in the platoon experience delay caused by stopping the head of the platoon, $D^{(H)u}_{j,T}$:

$$D^{(H)u}_{j,T} = P^u_{j,T}\left(t^u_T + R^{(1)u}_j(g^u_{i,T}) + \frac{N^u_{j,T-1}}{s^u_j} - t^u_{j,T}\right) \tag{14}$$

but no delay caused by stopping the tail of the platoon, $D^{(T)u}_{j,T}$.

The number of vehicles remaining in the residual queue at the end of cycle T, $N^u_{j,T}$, is:

$$N^u_{j,T} = 0. \tag{15}$$

The residual queue at the end of the previous cycle $T-1$ can be estimated based on the associated platoon case for the previous cycle and the subject lane group. Residual queue estimates can also be obtained with the use of detectors upstream of the stop line.

- Case P2: Arrival before residual queue served, insufficient green to serve entire platoon

  A platoon of size $P^u_{j,T}$ arrives at the back of its lane group's queue at time $t^u_{j,T}$ before the time that the corresponding residual queue, $N^u_{j,T-1}$, would have finished being served. There is enough available green time to serve the residual queue, but there is not enough spare

green time to serve all $P^u_{j,T}$ vehicles. These conditions are summarized as:

$$t^u_{j,T} \leq t^u_T + R^{(1)u}_j(g^u_{i,T}) + \frac{N^u_{j,T-1}}{s^u_j} \tag{16}$$

$$N^u_{j,T-1} \leq G^{e,u}_j(g^u_{i,T})s^u_j \tag{17}$$

$$P^u_{j,T} \geq G^{e,u}_j(g^u_{i,T})s^u_j - N^u_{j,T-1}. \tag{18}$$

All vehicles in the platoon, $P^u_{j,T}$, experience delay caused by stopping the head of the platoon, $D^{(H)u}_{j,T}$, and a portion of the vehicles equal to $P^u_{j,T} - G^{e,u}_j(g^u_{i,T})s^u_j + N^u_{j,T-1}$ experience delay caused by stopping the tail of the platoon, $D^{(T)u}_{j,T}$:

$$D^{(H)u}_{j,T} = P^u_{j,T}\left(t^u_T + R^{(1)u}_j(g^u_{i,T}) + \frac{N^u_{j,T-1}}{s^u_j} - t^u_{j,T}\right) \tag{19}$$

$$D^{(T)u}_{j,T} = \left(P^u_{j,T} - G^{e,u}_j(g^u_{i,T})s^u_j + N^u_{j,T-1}\right)\left(C - \frac{N^u_{j,T-1}}{s^u_j}\right). \tag{20}$$

The delay caused by stopping the tail of the platoon is equal to one cycle length minus the time it takes to serve the residual queue. This component is subtracted in order to avoid double counting since that delay component has already been captured in (19). However, one can adjust this delay estimate to change the penalty imposed for stopping the tail of the platoon.

The number of vehicles remaining in the residual queue at the end of cycle T, $N^u_{j,T}$, is:

$$N^u_{j,T} = P^u_{j,T} + N^u_{j,T-1} - G^{e,u}_j(g^u_{i,T})s^u_j. \tag{21}$$

- Case P3: Arrival before end of green, insufficient green to serve residual queue

A platoon of size $P^u_{j,T}$ arrives at the back of its lane group's queue at time $t^u_{j,T}$, before the end of the phase that can serve it, but there is not enough available green time to serve all $N^u_{j,T-1}$ vehicles in the residual queue. These conditions are summarized as:

$$t^u_{j,T} \leq t^u_T + R^{(1)u}_j(g^u_{i,T}) + G^{e,u}_j(g^u_{i,T}) \tag{22}$$

$$N^u_{j,T-1} \geq G^{e,u}_j(g^u_{i,T})s^u_j. \tag{23}$$

All vehicles in the platoon experience delay caused by stopping the head of the platoon, $D^{(H)u}_{j,T}$, and by stopping the tail of the platoon, $D^{(T)u}_{j,T}$:

$$D^{(H)u}_{j,T} = P^u_{j,T}\left(t^u_T + R^{(1)u}_j(g^u_{i,T}) + G^{e,u}_j(g^u_{i,T}) - t^u_{j,T}\right) \tag{24}$$

$$D^{(T)u}_{j,T} = P^u_{j,T}\left(C - G^{e,u}_j(g^u_{i\,\text{next}})\right) \tag{25}$$

where $g^u_{i\,\text{next}}$ is a pre-specified value for the green time of phase i for the next cycle $T+1$ at intersection u.

The number of vehicles remaining in the residual queue at the end of cycle T, $N^u_{j,T}$, is:

$$N^u_{j,T} = P^u_{j,T} + N^u_{j,T-1} - G^{e,u}_j(g^u_{i,T})s^u_j. \tag{26}$$

- Case P4: Arrival after residual queue served, entire platoon served in green

A platoon of size $P^u_{j,T}$ arrives at the back of its lane group's queue at time $t^u_{j,T}$ after the time that the corresponding residual queue, $N^u_{j,T-1}$, would have finished being served. There is enough available green time to serve the residual queue, and there is enough spare green time to serve all $P^u_{j,T}$ vehicles. These conditions are summarized as:

$$t^u_{j,T} \geq t^u_T + R^{(1)u}_j(g^u_{i,T}) + \frac{N^u_{j,T-1}}{s^u_j} \tag{27}$$

$$N^u_{j,T-1} \leq G^{e,u}_j(g^u_{i,T})s^u_j \tag{28}$$

$$P^u_{j,T} \leq \left(t^u_T + R^{(1)u}_j(g^u_{i,T}) + G^{e,u}_j(g^u_{i,T}) - t^u_{j,T}\right) s^u_j. \tag{29}$$

In this case, vehicles in the platoon do not experience any delay at intersection u. As a result, both the delay caused by stopping the head of that platoon, $D^{(H)u}_{j,T}$, and by stopping the tail of the platoon, $D^{(T)u}_{j,T}$, are zero.

The number of vehicles remaining in the residual queue at the end of cycle T, $N^u_{j,T}$, is:

$$N^u_{j,T} = 0. \tag{30}$$

- Case P5: Arrival after residual queue served, insufficient green to serve entire platoon

A platoon of size $P^u_{j,T}$ arrives at the back of its lane group's queue at time $t^u_{j,T}$ after the time that the corresponding residual queue, $N^u_{j,T-1}$, would have finished being served and before the end of the phase that can serve it. There is enough available green time to serve the residual queue, but there is not enough spare green time to serve all $P^u_{j,T}$ vehicles. These conditions are summarized as:

$$t^u_{j,T} \geq t^u_T + R^{(1)u}_j(g^u_{i,T}) + \frac{N^u_{j,T-1}}{s^u_j} \tag{31}$$

$$t^u_{j,T} \leq t^u_T + R^{(1)u}_j(g^u_{i,T}) + G^{e,u}_j(g^u_{i,T}) \tag{32}$$

$$N^u_{j,T-1} \leq G^{e,u}_j(g^u_{i,T})s^u_j \tag{33}$$

$$P^u_{j,T} \geq \left(t^u_T + R^{(1)u}_j(g^u_{i,T}) + G^{e,u}_j(g^u_{i,T}) - t^u_{j,T}\right) s^u_j. \tag{34}$$

A portion of vehicles in the platoon equal to $P^u_{j,T} - \left(t^u_T + R^{(1)u}_j(g^u_{i,T}) + G^{e,u}_j(g^u_{i,T}) - t^u_{j,T}\right) s^u_j$ experience delay caused by stopping the tail of the platoon, $D^{(T)u}_{j,T}$:

$$D^{(T)u}_{j,T} = \left[P^u_{j,T} - \left(t^u_T + R^{(1)u}_j(g^u_{i,T}) + G^{e,u}_j(g^u_{i,T}) - t^u_{j,T}\right) s^u_j\right] \left(t^u_{T+1} - t^u_{j,T} + R^{(1)u}_j(g^u_{i\text{ next}})\right) \quad (35)$$

but no delay caused by stopping the head of the platoon, $D^{(H)u}_{j,T}$.

The number of vehicles remaining in the residual queue at the end of cycle T, $N^u_{j,T}$, is:

$$N^u_{j,T} = P^u_{j,T} - (t^u_T + R^{(1)u}_j(g^u_{i,T}) + G^{e,u}_j(g_{i,T}) - t^u_{j,T}) s^u_j. \quad (36)$$

- Case P6: Arrival after the green

  A platoon of size $P^u_{j,T}$ arrives at the back of its lane group's queue at time $t^u_{j,T}$ after the end of the phase that can serve it. This case captures all arrivals not satisfying the conditions of cases P1 through P5, and it can also be expressed as:

$$t^u_{j,T} > t^u_T + R^{(1)u}_j(g^u_{i,T}) + G^{e,u}_j(g^u_{i,T}). \quad (37)$$

  All vehicles in the platoon experience delay caused by stopping the tail of the platoon, $D^{(T)u}_{j,T}$:

$$D^{(T)u}_{j,T} = P^u_{j,T} \left(t^u_{T+1} - t^u_{j,T} + R^{(1)u}_j(g^u_{i\text{ next}})\right) \quad (38)$$

  but no delay caused by stopping the head of the platoon, $D^{(H)u}_{j,T}$.

The number of vehicles remaining in the residual queue at the end of cycle T, $N^u_{j,T}$, is:

$$N^u_{j,T} = P^u_{j,T}. \quad (39)$$

2.1.2. Auto Delay for Vehicles in Platoons on Shared Links

The auto delay for vehicles in platoons that travel on shared links on the main arterial from one intersection u to another v can be estimated with the use of the equations for the six cases described above. The only differences are the values for the platoon size that need to be adjusted based on the portion of the platoon continuing downstream and its arrival time at the second intersection, v, which is a function of its arrival and therefore service time at the first intersection u. In addition, some of the delay estimates for stopping the tail of the platoon are modified.

An estimate of the size of the platoon in lane group j at the second intersection, v, during cycle T, denoted by $\hat{P}^v_{j,T}$, is used in the optimization instead of the actual incoming platoon size.

The estimate can be obtained with data from detectors located at the upstream end of the shared link between the two intersections and information on the percentage that is expected to join the subject lane group j. This information can be obtained by stop line detectors. The platoon size estimate is used instead of a direct calculation of platoon size from the signal settings and arrivals at the upstream intersection, because it reduces the number of bilinear and trilinear terms in the objective function, and as a result, it decreases the computation time of the optimization process. In addition, in cases that the number of lanes changes throughout the arterial, the saturation flow and delays at intersection v are normalized to represent saturation flow and delays on a per lane basis.

The arrival time of a platoon at the back of its lane group's queue at the second intersection, $t^v_{j,T}$, is estimated based on the arrival case for the first intersection as follows:

$$t^v_{j,T} = \begin{cases} t^u_T + R^{(1)u}_j(g^u_{i,T}) + tt^v_{j,u} & \text{for cases P1, P2, and P3} \\ t^u_{j,T} + tt^v_{j,u} & \text{for cases P4, P5} \\ t^u_{T+1} + R^{(1)u}_j(g^u_{i\text{ next}}) + tt^v_{j,u} & \text{for case P6} \end{cases} \tag{40}$$

where $tt^v_{j,u}$ is the average free flow travel time to traverse the shared links between intersections u and v. For cases P1, P2, P3, and P6 the estimate of the platoon's arrival time at v is based on the assumption that vehicles from the incoming platoon join the vehicles in the residual queue at u and travel together as one platoon. As a result, this new platoon is assumed to arrive at the downstream intersection $tt^v_{j,u}$ seconds after the beginning of green at u when the residual queue starts being served at the current (for Cases P1, P2, and P3) or next cycle (for Case P6). For cases P4 and P5, the platoons are served by u as soon as they arrive. This implies that residual queues are short, and as a result, the majority of vehicles at the downstream intersection, v, can be assumed to be mainly vehicles that have just been served by the upstream intersection u. Therefore, it is assumed that the arrival time of the platoon at v depends only on the service time of the platoon at u.

Finally, the delay estimates for stopping the tail of the platoon for cases P5 and P6 (see (35) and (38)) are adjusted as follows to avoid introducing nonlinear terms in the objective function:

$$D^{(T)v}_{j,T} = \begin{cases} \left[\hat{P}^v_{j,T} - \left(t^v_T + R^{(1)v}_j(g^v_{i,T}) + G^{e,v}_j(g^v_{i,T}) - t^v_{j,T}\right) s^v_j\right]\left(C - G^{e,v}_j(g^v_{i\text{ next}})\right) & \text{for case P5} \\ \hat{P}^v_{j,T}\left(C - G^{e,v}_j(g^v_{i\text{ next}})\right) & \text{for case P6.} \end{cases} \tag{41}$$

2.1.3. Auto Delay for Vehicles in Residual Queues

This delay component is added to account for the delays of the residual queues that are already present when a platoon arrives at an intersection during the design cycle. These are the vehicles that had arrived in a previous cycle and were not able to be served. The equations presented for Cases P1-P6, include only the delays of the vehicles that arrive in platoons during the design cycle, so the delays of these residual queues would otherwise not be captured. The equations presented in this section hold for the residual queues of any lane group regardless of the platoon case the vehicles that created these residual queues came from.

The auto delay for vehicles in residual queues at both intersections of the pair $(r, r+1) \in \Lambda_x$ are estimated based on the size of the residual queue and whether or not it can be entirely served during cycle T. Two cases arise which are described next along with the corresponding delay equations for an intersection r (but also hold for $r+1$):

- Case R1: Residual queue served in green

  The residual queue of a lane group j, $N^r_{j,T-1}$, can be entirely served during cycle T:

  $$N^r_{j,T-1} \leq G^{e,r}_j(g^r_{i,T}) s^r_j. \tag{42}$$

  So, the total delay experienced by all vehicles in the residual queue, $D^{(Q)r}_{j,T}$, is:

  $$D^{(Q)r}_{j,T} = N^r_{j,T-1} R^{(1)r}_j(g^r_{i,T}). \tag{43}$$

- Case R2: Residual queue not entirely served in green

  The residual queue of a lane group j, $N^r_{j,T-1}$, cannot be entirely served during cycle T:

  $$N^r_{j,T-1} \geq G^{e,r}_j(g^r_{i,T}) s^r_j. \tag{44}$$

  So, the total delay experienced by all vehicles in the residual queue, $D^{(Q)r}_{j,T}$, is:

  $$D^{(Q)r}_{j,T} = N^r_{j,T-1} R^{(1)r}_j(g^r_{i,T}) + \left(N^r_{j,T-1} - G^{e,r}_j(g^r_{i,T}) s^r_j\right) C \tag{45}$$

  because the vehicles that do not get served will have to wait for an extra cycle before they start being served.

Note that estimates of the delays incurred during the design cycle and the following cycle are considered for residual queues. If a residual queue persists for many cycles, only the delays

in the design cycle and following cycle are used to optimize the signal settings. The remaining unserved vehicles will be accounted for in the optimization of the next cycle. This is not expected to affect the results much since the delay included for the next cycle is just an estimate, and it will ultimately be determined when the signal settings of that next cycle are optimized.

2.1.4. Total Auto Person Delay

The auto delay component of the objective function that minimizes person delay at two consecutive intersections $(r, r+1) \in \Lambda_x$ is as follows:

$$\sum_{(r,r+1)\in\Lambda_x} \sum_{a=1}^{A_T^r} o_a d_a^r = \sum_{u=1}^{2} \sum_{j\in J_{IN}^u} \bar{o}_a \left(D_{j,T}^{(H)u} + D_{j,T}^{(T)u} \right) + \sum_{v=1}^{2} \sum_{j\in J_{SH}^v} \bar{o}_a \left(D_{j,T}^{(H)v} + D_{j,T}^{(T)v} \right) + \sum_{r=1}^{2} \sum_{j=1}^{J^r} \bar{o}_a D_{j,T}^{(Q)r} \tag{46}$$

where $\bar{o}_a$ is the average auto passenger occupancy, J_{IN}^u is the set of lane groups for the incoming links at intersection u, J_{SH}^v is the set of lane groups for the shared links at intersection v, and J^r is the total number of lane groups at intersection r. The summation over u represents the delays experienced by all vehicles on the incoming links of the two intersections being optimized. The summation over v represents all delays experienced by vehicles on the shared links of those two intersections (see Figures 1(b) and 1(c)). The delay components can be estimated with the use of the corresponding case equation for each of the three delay types as presented above. For each platoon and lane group, each case introduces a binary decision variable and the related constraints.

More specifically, the sum of the delay components for the lane groups for the incoming links at intersection u is given by:

$$\begin{aligned} D_{j,T}^{(H)u} + D_{j,T}^{(T)u} = {} & \left(h_{j,T}^1 + h_{j,T}^2\right) P_{j,T}^u \left(t_T^u + R_j^{(1)u}(g_{i,T}^u) + \frac{N_{j,T-1}^u}{s_j^u} - t_{j,T}^u \right) \\ & + h_{j,T}^2 \left(P_{j,T}^u - G_j^{e,u}(g_{i,T}^u) s_j^u + N_{j,T-1}^u \right) \left(C - \frac{N_{j,T-1}^u}{s_j^u} \right) \\ & + h_{j,T}^3 P_{j,T}^u \left[\left(t_T^u + R_j^{(1)u}(g_{i,T}^u) + G_j^{e,u}(g_{i,T}^u) - t_{j,T}^u \right) + \left(C - G_j^{e,u}(g_{i\,\text{next}}^u) \right) \right] \\ & + \left[h_{j,T}^5 \left(P_{j,T}^u - \left(t_T^u + R_j^{(1)u}(g_{i,T}^u) + G_j^{e,u}(g_{i,T}^u) - t_{j,T}^u \right) s_j^u \right) + h_{j,T}^6 P_{j,T}^u \right] \\ & \times \left(t_{T+1}^u - t_{j,T}^u + R_j^{(1)u}(g_{i\,\text{next}}^u) \right) \end{aligned} \tag{47}$$

where $h_{j,T}^1$, $h_{j,T}^2$, $h_{j,T}^3$, $h_{j,T}^4$, $h_{j,T}^5$, $h_{j,T}^6$ are binary variables and for the two platoons traveling on the arterial (i.e., on shared links) and arriving at the second intersection, v, that also belongs to the

pair currently being optimized, is as follows:

$$
\begin{aligned}
D_{j,T}^{(H)v} + D_{j,T}^{(T)v} = {} & \left(z_{j,T}^{1} + z_{j,T}^{2}\right) \hat{P}_{j,T}^{v} \left[t_T^v + R_j^{(1)v}(g_{i,T}^v) + \frac{N_{j,T-1}^v}{s_j^v} - t_{j,T}^v \right] \\
& + z_{j,T}^{2} \left(\hat{P}_{j,T}^{v} - G_j^{e,v}(g_{i,T}^v) s_j^v + N_{j,T-1}^v \right) \left(C - \frac{N_{j,T-1}^v}{s_j^v} \right) \\
& + z_{j,T}^{3} \hat{P}_{j,T}^{v} \left[\left(t_T^v + R_j^{(1)v}(g_{i,T}^v) + G_j^{e,v}(g_{i,T}^v) - t_{j,T}^v \right) + \left(C - G_j^{e,v}(g_{i\text{ next}}^v) \right) \right] \\
& + z_{j,T}^{5} \left[\hat{P}_{j,T}^{v} - \left(t_T^v + R_j^{(1)v}(g_{i,T}^v) + G_j^{e,v}(g_{i,T}^v) - t_{j,T}^v \right) s_j^v \right] \\
& \times \left(C - G_j^{e,v}(g_{i\text{ next}}^v) \right) \\
& + z_{j,T}^{6} \hat{P}_{j,T}^{v} \left(C - G_j^{e,v}(g_{i\text{ next}}^v) \right). \qquad (48)
\end{aligned}
$$

where $z_{j,T}^1$, $z_{j,T}^2$, $z_{j,T}^3$, $z_{j,T}^4$, $z_{j,T}^5$, $z_{j,T}^6$ are binary variables.

In order to estimate this component of the objective function, estimates of the arrival times at the back of the queue of the corresponding lane group at the downstream intersection are needed, $t_{j,T}^v$, which are based on the cases shown in (40) can be expressed as follows:

$$
\begin{aligned}
t_{j,T}^v = {} & \left(h_{j,T}^1 + h_{j,T}^2 + h_{j,T}^3\right) \left(t_T^u + R_j^{(1)u}(g_{i,T}^u) + tt_{j,u}^v \right) + \left(h_{j,T}^4 + h_{j,T}^5\right) \left(t_{j,T}^u + tt_{j,u}^v \right) \qquad (49) \\
& + h_{j,T}^6 \left(t_{T+1}^u + R_j^{(1)u}(g_{i\text{ next}}^u) + tt_{j,u}^v \right)
\end{aligned}
$$

Finally, the person delay for the autos that are already in the residual queue of an intersection u is estimated as follows:

$$
\bar{o}_a \sum_{j=1}^{J^r} D_{j,T}^{(Q)r} = \bar{o}_a \left(x_{j,T}^1 + x_{j,T}^2\right) N_{j,T-1}^r R_j^{(1)r}(g_{i,T}^r) + \bar{o}_a x_{j,T}^2 \left(N_{j,T-1}^r - G_j^{e,r}(g_{i,T}^r) s_j^r \right) C \qquad (50)
$$

where J^r is the total number of lane groups at intersection r and $x_{j,T}^1$, $x_{j,T}^2$ are binary variables.

The summation of binary variables for a platoon or lane group for a specific type of delay (i.e., at intersection u, v or when in residual queue) is constrained to be one to ensure that only one of the delay equations will be added to the objective function for that delay type. Additional constraints to determine each of the cases are also included. Constraints for case 1 of the lane groups at intersection u are presented here as an example.

$$
t_{j,T}^u \le t_T^u + R_j^{(1)u}(g_{i,T}^u) + \frac{N_{j,T-1}^u}{s_j^u} + (1 - h_{j,T}^1) M \qquad (51)
$$

$$
N_{j,T-1}^u \le G_j^{e,u}(g_{i,T}^u) s_j^u + (1 - h_{j,T}^1) M \qquad (52)
$$

$$
P_{j,T}^u \le G_j^{e,u}(g_{i,T}^u) s_j^u - N_{j,T-1}^u + (1 - h_{j,T}^1) M \qquad (53)
$$

A detailed presentation of all constraints included in the mathematical program can be found in Christofa (2012). A big value constant M is used with the binary variables to determine which constraints to activate for the relevant cases. For this mathematical program, M is set equal to TC for a cycle indexed by T with length C.

2.2. Transit Delay

In addition to auto person delay, the objective function includes the person delay for transit vehicles present at the two intersections during cycle T, which consists of two terms: 1) the delay transit passengers experience when traveling on incoming links during cycle T, and 2) the delay they experience when traveling on shared links during cycle T. In addition, users of transit vehicles that do not get served during the cycle in which they arrive experience an extra component of delay equal to $R_j^{(1)r}(g_{i\,\text{next}}^r)$, which is the time a transit vehicle would experience if it was the first one in the queue to be served in the next cycle. Since transit vehicles travel in mixed-use traffic lanes and perfect information about their arrivals is assumed, their delays can be estimated as in the cases of platoons, further assuming that a transit vehicle behaves similarly to a platoon of size one.

2.2.1. Transit Delay for Vehicles on Incoming Links

The transit delay for vehicles traveling on incoming links to an intersection u, depends on the actual arrival time at the back of its lane group's queue at that intersection, $t_{b,T}^u$, as well as whether the vehicle is served during cycle T or not, which also depends on traffic conditions on the subject approach. Note that the delay equations are formulated based on the assumption that a transit vehicle arrives at the back of the queue before or after the arrival of the platoon due to its dwell time at bus stops. This means that when arriving at the intersection, the bus observes only the residual queue in front of it. This assumption is justified because most local transit lines in urban settings have bus stops at almost every signalized intersection. In cases that this is not true (i.e., transit vehicles do not stop at stops), their delay is equal to the delay of a vehicle in the platoon within which they are traveling. Such delay equations can be found in Farid et al. (2014) and can be used to update the mathematical program accordingly. The four delay estimation cases for transit arrivals are presented next.

- Case T1: Arrival before residual queue served, transit vehicle served in green

  A transit vehicle arrives at the back of its lane group's queue at time $t^u_{b,T}$ before the time the corresponding residual queue, $N^u_{j,T-1}$, would have finished being served and there is enough available green time to serve the residual queue in front of it. These conditions are summarized as:

$$t^u_{b,T} \leq t^u_T + R^{(1)u}_j(g^u_{i,T}) + \frac{N^u_{j,T-1}}{s^u_j} \tag{54}$$

$$N^u_{j,T-1} \leq G^{e,u}_j(g^u_{i,T}) s^u_j. \tag{55}$$

  The transit vehicle is served during cycle T and its delay, $d^u_{b,T}$, can be expressed as:

$$d^u_{b,T} = t^u_T + R^{(1)u}_j(g^u_{i,T}) + \frac{N^u_{j,T-1}}{s^u_j} - t^u_{b,T}. \tag{56}$$

- Case T2: Arrival before residual queue served, transit vehicle not served in green

  A transit vehicle arrives at the back of its lane group's queue at time $t^u_{b,T}$ before the time the corresponding residual queue, $N^u_{j,T-1}$, would have finished being served, but there is not enough available green time to serve the residual queue in front of it. These conditions are summarized as:

$$t^u_{b,T} \leq t^u_T + R^{(1)u}_j(g^u_{i,T}) + \frac{N^u_{j,T-1}}{s^u_j} \tag{57}$$

$$N^u_{j,T-1} \geq G^{e,u}_j(g^u_{i,T}) s^u_j. \tag{58}$$

  The transit vehicle is not served during cycle T and its delay, $d^u_{b,T}$, can be expressed as:

$$d^u_{b,T} = t^u_{T+1} - t^u_{b,T} + R^{(1)u}_j(g^u_{i\,\text{next}}). \tag{59}$$

- Case T3: Arrival after residual queue served and before the end of green

  A transit vehicle arrives at the back of its lane group's queue at time $t^u_{b,T}$ after the time the corresponding residual queue, $N^u_{j,T-1}$, would have finished being served and before the end of the green time for the phase that can serve it. There is enough available green time to serve the residual queue in front of it. These conditions are summarized as follows:

$$t^u_{b,T} \geq t^u_T + R^{(1)u}_j(g^u_{i,T}) + \frac{N^u_{j,T-1}}{s^u_j} \tag{60}$$

$$t^u_{b,T} \leq t^u_T + R^{(1)u}_j(g^u_{i,T}) + G^{e,u}_j(g^u_{i,T}) \tag{61}$$

$$N^u_{j,T-1} \leq G^{e,u}_j(g^u_{i,T}) s^u_j. \tag{62}$$

The transit vehicle is served as soon as it arrives at the intersection and as a result its delay, $d^u_{b,T}$, is zero.

- Case T4: Arrival after the end of green

  A transit vehicle arrives at the back of its lane group's queue during cycle T at time $t^u_{b,T}$ after the end of the green time for the phase that can serve it, which is expressed as follows:

$$t^u_{b,T} > t^u_T + R^{(1)u}_j(g^u_{i,T}) + G^{e,u}_j(g^u_{i,T}). \tag{63}$$

  The transit vehicle is not served during cycle T, and its delay, $d^u_{b,T}$, can be expressed as:

$$d^u_{b,T} = t^u_{T+1} - t^u_{b,T} + R^{(1)u}_j(g^u_{i\,\text{next}}). \tag{64}$$

2.2.2. Transit Delay for Vehicles on Shared Links

The delay for a transit vehicle that arrives at its second intersection, v, after it is served by the first intersection, u, is also included in the objective function to account for the impact that the signal timings at one intersection have on the other. This means that some level of progression for the transit vehicles between adjacent intersections is taken into account as well. As with autos, the delay for transit vehicles traveling on shared links on the main arterial from one intersection, u, to another, v, can be estimated with the use of the equations for the four cases presented above given that the transit arrival time at the second intersection, v, $t^v_{b,T}$, can be estimated. $t^v_{b,T}$ depends on the arrival case at the first intersection, u, and can be estimated as follows:

$$t^v_{b,T} = \begin{cases} t^u_T + R^{(1)u}_j(g^u_{i,T}) + \frac{N^u_{j,T-1}}{s^u_j} + tt^v_{b,u} & \text{for case T1} \\ t^u_{b,T} + tt^v_{b,u} & \text{for case T3} \end{cases} \tag{65}$$

where $tt^v_{b,u}$ is the expected travel time for the shared link between intersections u and v for a transit vehicle b, and it includes the lost time due to transit stops. For cases T2 and T4, the transit vehicle is not served during cycle T at u, and no delay at the downstream intersection, v, is included in the objective function for cycle T.

2.2.3. Total Transit Person Delay

The transit delay component of the objective function that minimizes person delay at two consecutive intersections is as follows:

$$\sum_{b=1}^{B^u_T} o^u_{b,T}(1+\delta^u_{b,T})d^u_{b,T} + \sum_{b=1}^{B^v_T} o^v_{b,T}(1+\delta^v_{b,T})d^v_{b,T} \tag{66}$$

where $d^u_{b,T}$ and $d^v_{b,T}$ can be estimated per one of the cases presented above. As for the auto delays, each case introduces a binary decision variable and the related constraints. The summation of binary variables for a transit vehicle at intersection u or v is constrained to add up to one to ensure that only one of the delay equations will be added to the objective function per transit vehicle for that intersection.

The total delay for the transit vehicles consists of two terms: 1) the delay experienced by the transit vehicles at the first intersection of the pair they arrive, u, $d^u_{b,T}$, and 2) the delay experienced by the transit vehicles that travel on the arterial at the second intersection they arrive, v, $d^v_{b,T}$.

The transit person delay component of the objective function for the transit vehicles at the first intersection at which they arrive during cycle T, $d^u_{b,T}$, is expressed as follows:

$$\sum_{b=1}^{B^u_T} o^u_{b,T}(1+\delta^u_{b,T})d^u_{b,T} = \zeta^1_{b,T}\left(t^u_T + R^{(1)u}_j(g^u_{i,T}) + \frac{N^u_{j,T-1}}{s^u_j} - t^u_{b,T}\right) + \left(\zeta^2_{b,T}+\zeta^4_{b,T}\right)\left(t^u_{T+1} - t^u_{b,T} + R^{(1)u}_j(g^u_{i\,\text{next}})\right) \tag{67}$$

and the person delay for those that continue in the network to the other intersection of the pair being optimized and arrive at their second intersection, v, during cycle T, experience delay, $d^v_{b,T}$, which is expressed as follows:

$$\sum_{b=1}^{B^v_T} o^v_{b,T}(1+\delta^v_{b,T})d^v_{b,T} = \eta^1_{b,T}\left(t^v_T + R^{(1)v}_j(g^v_{i,T}) + \frac{N^v_{j,T-1}}{s^v_j} - t^v_{b,T}\right) \tag{68}$$

$$+ \left(\eta^2_{b,T}+\eta^4_{b,T}\right)\left(t^v_{T+1} - t^v_{b,T} + R^{(1)v}_j(g^v_{i\,\text{next}})\right). \tag{69}$$

In order to estimate this component of the objective function, estimates of the arrival times at the back of the queue of the corresponding lane group at the downstream intersection, $t^v_{b,T}$, are needed. These estimates are based on the cases shown in (65) and can be expressed as follows:

$$t^v_{b,T} = \zeta^1_{b,T}\left(t^u_T + R^{(1)u}_j(g^u_{i,T}) + \frac{N^u_{j,T-1}}{s^u_j} + tt^v_{b,u}\right) + \zeta^3_{b,T}(t^u_{b,T} + tt^v_{b,u}). \tag{70}$$

2.3. Mathematical Program Formulation

The objective function of the mathematical program that minimizes person delays for two intersections for cycle T, is the summation of the total auto person delays and transit person

delays as follows:

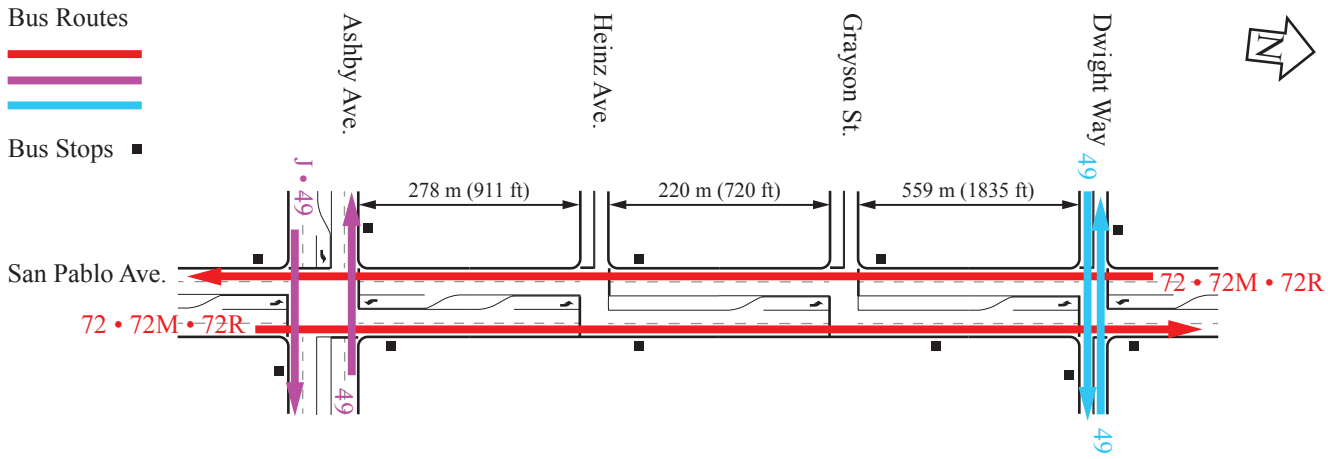
$$\sum_{u=1}^{2}\left(\sum_{j\in J_{IN}^{u}}\bar{o}_a\left(D_{j,T}^{(H)u}+D_{j,T}^{(T)u}\right)+\sum_{b=1}^{B_T^u}o_{b,T}^{u}(1+\delta_{b,T}^{u})d_{b,T}^{u}\right)$$
$$+\sum_{v=1}^{2}\left(\sum_{j\in J_{SH}^{u}}\bar{o}_a\left(D_{j,T}^{(H)v}+D_{j,T}^{(T)v}\right)+\sum_{b=1}^{B_T^v}o_{b,T}^{v}(1+\delta_{b,T}^{v})d_{b,T}^{v}\right)+\sum_{u=1}^{2}\sum_{j=1}^{J^r}\bar{o}_a D_{j,T}^{(Q)r}. \quad (71)$$

This objective function includes bilinear and trilinear terms caused by multiplication between continuous and binary decision variables, which introduces nonlinear terms in the objective function and the constraints. To avoid this problem, convex relaxations for bilinear and trilinear terms as described in Meyer and Floudas (2004) are used. More specifically, first all bilinear and trilinear terms are replaced with new linear variables. Then additional linear constraints are introduced to the mathematical program to relate the new linear variables with the original nonlinear terms and accordingly determine the bounds of the new variables to make the relaxation as tight as possible. After performing the convex relaxations, the mathematical program consists of an objective function that is linear in its continuous and binary variables and has linear constraints. This leads to a Mixed-Integer Linear Programming (MILP) problem that may be readily solved with the use of broadly available codes or commercial software (e.g., CPLEX (IBM, 2011)) within few CPU-seconds. This can be scaled to larger arterial networks because the optimization is performed at two intersections at a time, so the computation time increases linearly with the number of intersections. A detailed description of the complete mathematical program after the implementation of the convex relaxations can be found in Christofa (2012).

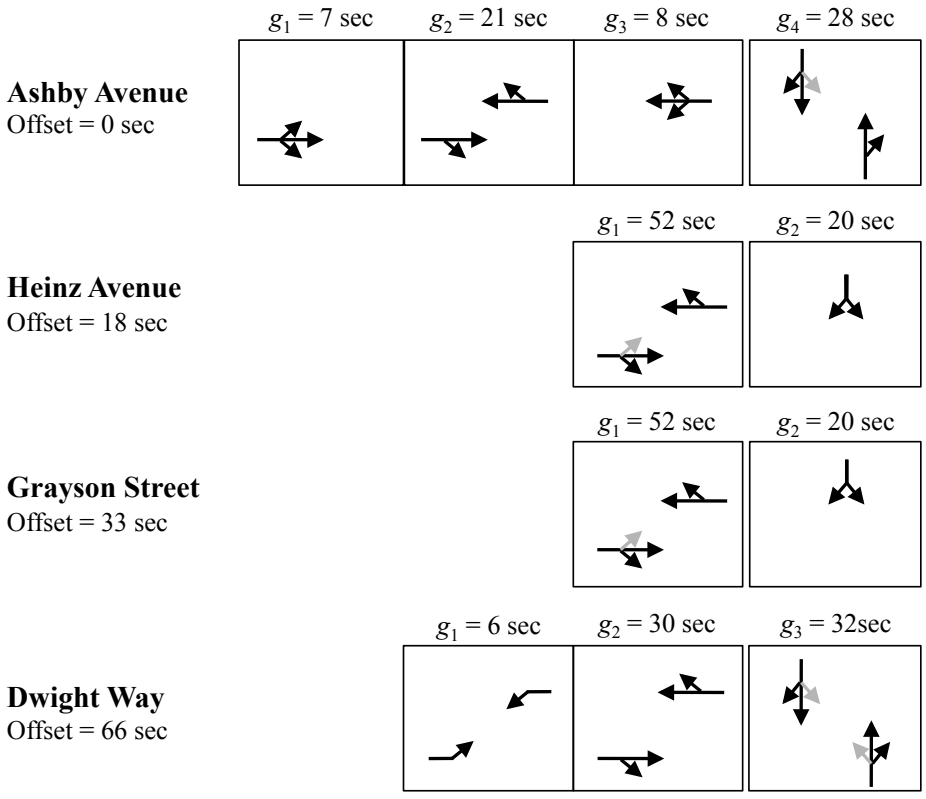
3. Test Site

The performance of the person-based traffic responsive signal control system is tested at a four-intersection segment of a real-world signalized arterial. In particular, the test site consists of the intersections along San Pablo Avenue at Ashby Avenue, Heinz Avenue, Grayson Street, and Dwight Way located in Berkeley, California. This segment has been selected due to the variety in phasing schemes utilized on the four intersections and the existence of conflicting bus routes at two intersections. In particular, the four intersections that have been chosen vary in their car demands as expressed through their intersection flow ratios, number of conflicting transit routes, number of phases (including 2, 3, and 4-phase signalized intersection), and varying link lengths. Note also

that the auto demands have been obtained through previous field studies while the bus frequencies are based on the published transit schedule, therefore, representing real-world conditions.



(a) Layout and bus routes (not to scale)



(b) Signal phasing and TRANSYT-7F optimal green times and offsets

Figure 3: San Pablo Avenue test site.

The arterial segment's layout and bus routes are shown in Figure 3(a). Figure 3(b) presents the phasing and TRANSYT-7F optimal offsets and green times for all intersections during the

evening peak hour (4–5pm). The optimal splits and offsets were obtained by minimizing the disutility index in TRANSYT-7F, which is a linear combination of delay and number of stops, while the cycle length was fixed at 80 seconds, and the phase sequence was kept constant. The hill climb optimization technique is used to optimize the signal settings in TRANSYT-7F. As can be seen, the intersections consist of a variety of signal phasing schemes that cover all of the basic possible phasing schemes. All intersections operate under a common cycle length of 80 seconds and the demand used for the tests corresponds to the evening peak hour. During that time period, all four intersections operate in undersaturated traffic conditions with intersection flow ratios varying from 0.4 to 0.8 (0.4 corresponds to the intersection of San Pablo Avenue and Grayson Street and 0.8 to the intersection of San Pablo Avenue and Dwight Way). The intersection of Ashby and San Pablo Avenues has been chosen as the critical one because it presents a higher transit demand and the northbound is the heaviest traffic direction. The link lengths between the intersections of the selected segment vary from 220 to 560 meters and the existing signal control is fixed-time coordinated.

Five bus routes travel through the segment under consideration in mixed-use traffic lanes with headways that vary from 12 to 30 minutes on each route. This corresponds to an average of 24 buses per hour for the analysis period. The numbers next to the directional arrows in Figure 3(a) correspond to the different bus routes. Of the buses that travel in the corridor, 60% travel on San Pablo Avenue and 40% on the two cross streets: Ashby Avenue and Dwight Way. At these cross streets, buses travel in two conflicting directions. The location of the bus stops varies with some of them being located nearside and some others farside (Figure 3(a)). The bus schedule is available at the Alameda-Contra Costa Transit District's website (AC Transit, 2011).

4. Application

Data from the arterial segment of San Pablo Avenue are used to test the proposed signal control system. Since the heaviest direction is the northbound, the pairwise optimization is performed in that direction. First, tests are performed for a few cycles assuming that perfect information exists on the platoon sizes and arrival times of platoons and transit vehicles at intersections (deterministic arrival tests). These give an idea of the maximum benefit that can be achieved by the proposed system. Deterministic arrival tests provide the maximum benefit that can be achieved by a real-time signal control system regardless of the traffic conditions that prevail on

the arterial of interest. Next, tests are performed with Emulation-In-the-Loop Simulation (EILS) to evaluate the system when perfect information is not available and predictions of inputs are based on measured quantities from the simulated network as it would be done in reality (stochastic arrival tests). A warm-up period equal to the common cycle length of all intersections is used. The computation times for both types of tests are on the order of 5-10 seconds[2] for optimization of the signal settings at all four intersections. Given that the optimization can be performed sequentially, and the computation time for optimizing a single pair of intersections is less than 5 seconds, the proposed mathematical program can be easily used for real-world implementations and is scalable to large arterial networks. In particular, given that phases usually have a minimum green time of 7-10 seconds, and the optimization is run once per cycle, the proposed system can easily be run during the last phase of the previous cycle. Therefore the system is applicable for controlling real-world signals.

The evening peak average flows are used as the input for auto demand and the bus arrivals are based on the actual schedule. The average auto occupancy, $\bar{o}_a$, is assumed to be 1.25 passengers per vehicle and the average bus occupancy, $\bar{o}_b$, is fixed to 40 passengers per vehicle for all buses. Similar values have been used by other studies, for example Hu et al. (2015). However, any bus passenger or auto passenger occupancies can be utilized with the proposed system. For the intersections where the coordinated phase is the first of the cycle, the offset has been set equal to the optimal offsets obtained from TRANSYT-7F. The green times for the next cycle, $g^r_{i\,\text{next}}$, are set to be the same as the fixed optimal signal timings provided by TRANSYT-7F for the specific traffic conditions under evaluation (Figure 3(b)). In addition, the upper bounds for the green times of the phases, $g^r_{i\,\max}$, are set equal to $C - \sum_{i=1}^{I^r} y^r_i$ for each intersection r. Non-zero lower bounds for the green times of each phase, $g^r_{i\,\min}$, are also introduced for the green times for each intersection r to ensure that no phase is skipped and all phases are allocated some minimum green time to guarantee sufficient time for safe vehicle and pedestrian crossings. A total minimum green time of 7 seconds is assigned to each of the left-turn phases and 12 seconds to each of the through phases.

[2]All tests, both for deterministic and stochastic arrivals, were performed on an Intel Core i5 2.4 GHz processor with a memory of 4 GB.

4.1. Deterministic Arrival Tests

Deterministic arrival tests are performed for the four-intersection segment for five signal cycles and two optimization scenarios: 1) TRANSYT-7F fixed-time optimal signal settings and 2) person-based optimization signal settings (i.e., when total person delay for both auto and bus passengers is minimized for each pair of intersections). Table 1 presents the auto, bus, and total person delay obtained from these scenarios. A comparison of the person-based optimization with TRANSYT-7F indicates that the proposed system can achieve a reduction in total person delay of 25% by reducing the delay of the bus users by about 18% and auto users by 29%. This results in an average bus delay saving of about 2.5 seconds per intersection. The values presented here are the maximum benefits that can be achieved for the specific traffic, transit, signal setting and geometric conditions since they are based on the assumption of perfect information for bus and platoons arrivals and passenger occupancies. In reality, these inputs can be estimated with some errors, so the benefits are expected to be lower. The next section shows the results of tests performed when the assumption of perfect information is relaxed.

Table 1: Person delays on the arterial segment for $\bar{o}_b/\bar{o}_a = 40/1.25$ and five signal cycles of traffic operations (deterministic arrivals).

	Auto Person Delay (pax-hrs)	Bus Person Delay (pax-hrs)	Overall Person Delay (pax-hrs)
TRANSYT-7F	5.91	2.70	8.61
Person-based Optimization	4.20	2.22	6.42
% Change	–28.93%	–17.78%	–25.43%

4.2. Stochastic Arrival Tests (Simulation Experiments)

EILS tests are performed with the AIMSUN microscopic simulation model (TSS, 2010) for one hour of operations. Each scenario is evaluated ten times and the average person delays of these replications for auto, bus, and all passengers are compared. The auto inter-arrival times on the incoming links are simulated to follow an exponential distribution. However, these vehicles stop at upstream fixed-time signalized intersections before they arrive at the intersections on the

arterial. This ensures that vehicles arrive in platoons at the intersections being optimized. As a result, the auto demand for an approach is estimated based on measurements from detectors located at the upstream end of each incoming link under consideration. Exponential smoothing is used on the measured counts during the previous cycle in order to estimate the demand of the respective lane group for the next cycle. Predictions of auto arrival times at the intersections are based on an average free flow speed of 45 km/hr. Although the input traffic demands for the subject arterial segment imply undersaturated conditions, the stochastic arrivals in the simulation exhibit random fluctuations. These fluctuations could lead to the existence of residual queues and transient oversaturated conditions, which can certainly be handled by the proposed optimization approach and last only a few cycles.

The timetable of the bus arrivals at the entry links of the network is fixed and based on headways obtained from the actual schedule. The arrival time of the buses at the intersections is predicted based on their location on a link at the end of the previous cycle. Information on the location of vehicles and bus stops determine the estimated arrival time of a bus at the intersection. For simplicity, dwell times for all buses and stops are set to 30 seconds. For each bus that stops, an additional 6 seconds are added to its estimated travel time to reach the intersection in order to account for lost time due to acceleration and deceleration. The average speed assumed for buses is 36 km/hr which is slower than the free-flow speed for autos.

Table 2 shows the auto, bus, and overall average vehicle delays per intersection and total person delays for each arterial direction, the cross streets and the whole arterial segment obtained when TRANSYT-7F and person-based optimal signal settings are implemented. Standard deviations are presented in parentheses. Figure 4 shows the percent changes in person delay from TRANSYT-7F to person-based optimization for autos, buses, and the overall four-intersection arterial segment per direction and for the whole system. The figure also presents the 95% confidence intervals of those changes.

The results indicate that the proposed person-based optimization can reduce the overall person delay by 5.1% compared to TRANSYT-7F by reducing the bus person delay by 3.9% and the auto person delay by 5.8%. These correspond to average vehicle delay reductions of about 1.2 seconds per auto and 1.8 seconds per bus per intersection. These reductions are statistically significant as shown by the 95% confidence intervals presented in Figure 4. There are directions whose autos and buses receive large reductions of delay, thereby benefiting multiple users and significantly reducing

Table 2: Person and vehicle delays for $\bar{o}_b/\bar{o}_a = 40/1.25$ and 1 hour of traffic operations (simulation).

	Auto		Bus		Overall	
	Total (pax-hrs)	Average (sec/veh)	Total (pax-hrs)	Average (sec/veh)	Total (pax-hrs)	Average (sec/veh)
Main Arterial Northbound						
TRANSYT-7F	23.31 (1.88)	14.22 (0.94)	22.62 (0.34)	57.99 (0.84)	45.93 (1.67)	22.64 (0.61)
Person-based	22.86 (2.98)	13.94 (1.66)	22.01 (0.76)	56.28 (2.08)	44.87 (3.33)	22.09 (1.46)
% Change	-1.9%	-2.0%	-2.7%	-3.0%	-2.3%	-2.4%
Main Arterial Southbound						
TRANSYT-7F	20.58 (2.14)	17.77 (1.37)	14.03 (0.48)	39.46 (1.36)	34.62 (1.88)	22.81 (0.78)
Person-based	18.09 (1.41)	15.59 (0.82)	13.73 (0.38)	38.60 (1.07)	31.81 (1.29)	21.00 (0.48)
% Change	-12.1%	-12.2%	-2.2%	-2.2%	-8.1%	-7.9%
Cross Streets						
TRANSYT-7F	34.97 (3.76)	34.10 (3.33)	6.98 (0.60)	32.58 (3.19)	41.95 (4.30)	34.06 (3.16)
Person-based	33.35 (1.84)	32.56 (1.47)	6.18 (0.46)	28.85 (2.32)	39.53 (1.95)	31.91 (1.27)
% Change	-4.6%	-4.5%	-11.4%	-11.4%	-5.8%	-6.3%
System						
TRANSYT-7F	78.87 (4.35)	20.64 (0.99)	43.62 (0.72)	45.44 (0.71)	122.49 (4.41)	25.63 (0.80)
Person-based	74.30 (3.64)	19.44 (0.90)	41.92 (0.86)	43.62 (1.06)	116.22 (3.88)	24.30 (0.77)
% Change	-5.8%	-5.8%	-3.9%	-4.0%	-5.1%	-5.2%

total person delay. In particular, cross streets and the southbound approaches are the biggest beneficiaries of a person-based optimization. For cross streets person-based optimization results in about 5.8% reduction in overall person delay achieved by reducing the auto person delay by 4.6% and the bus person delay by 11.4% compared to TRANSYT-7F. The provision of priority to buses on cross streets leads to longer green time for them which also substantially benefits auto users. The cross streets for this specific study site have low auto demands, therefore, auto passengers also often benefit when the delays of those links are weighed more due to the presence of buses. These reductions, that are statistically significant with 95% confidence, correspond to average vehicle delay reductions of about 1.5 seconds and 3.7 seconds for autos and buses, respectively. Note that these numbers are average values of vehicle delays at each link. Therefore, this proposed system

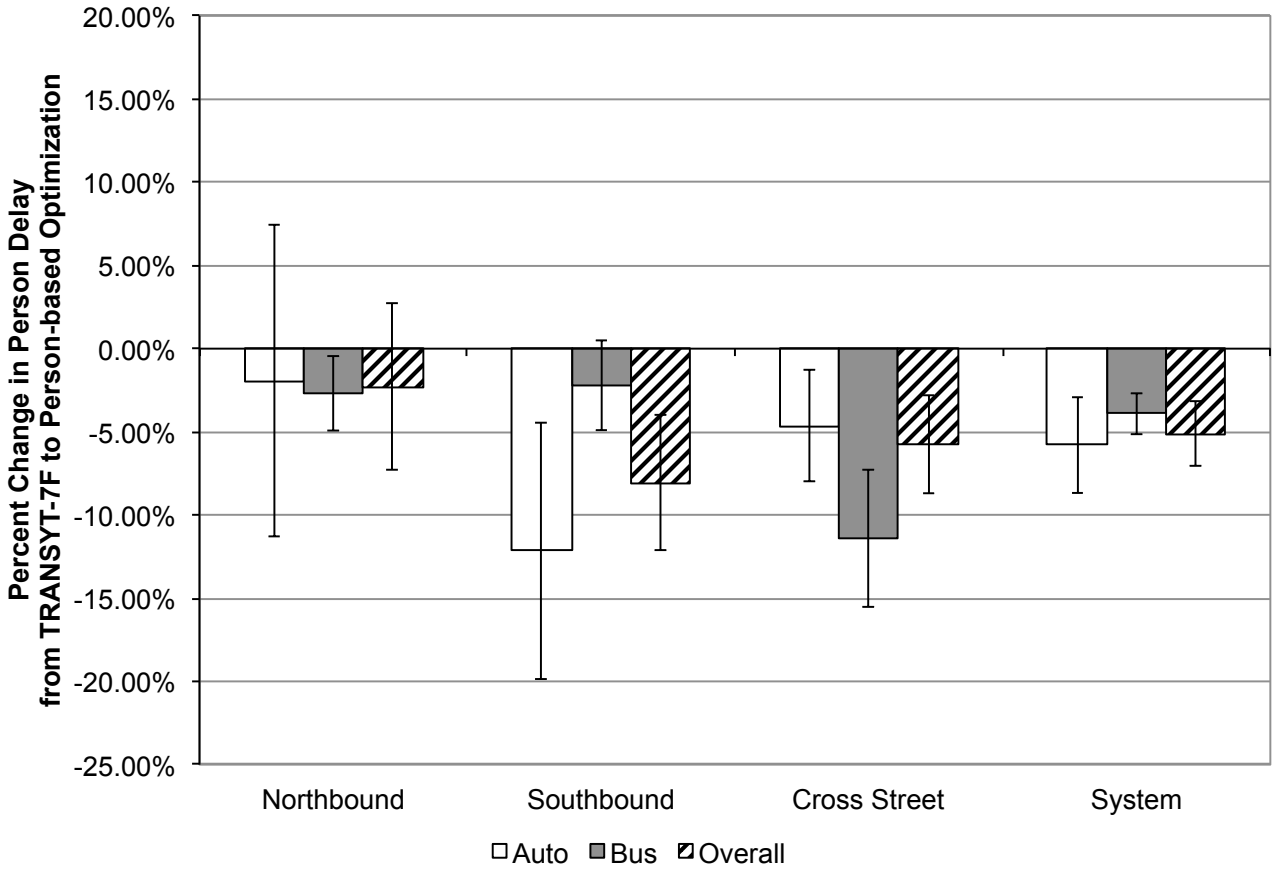


Figure 4: Percent Change in Person Delays from TRANSYT-7F to Person-based Optimization (simulation).

can result in bigger reductions in bus delay when implemented on long bus arterial corridors.

Southbound passengers see an overall reduction in their delay of about 8%, which corresponds to 1.8 seconds of delay savings for all users per vehicle per intersection. The overall and auto person delay reductions are statistically significant with 95% confidence. Southbound auto passengers experience significant reductions of 12.1% in their delay when person-based optimization is implemented because the TRANSYT-7F signal settings have been optimized for the northbound direction. Therefore, the person-based optimization provides additional green to the southbound direction compared to TRANSYT-7F in order to minimize person delay since the demands at two of the intersections for the southbound direction are very similar. On the other hand, passengers on the northbound direction experience smaller delay reductions when person-based optimization is implemented compared to TRANSYT-7F signal settings. However, the overall and auto delay reductions for the northbound passengers are not statistically significant with 95% confidence. This is due to the fact that the northbound passengers are already experiencing relatively low

delays when TRANSYT-7F signal settings are in place, and TRANSYT-7F signal settings have been optimized to provide progression to the northbound direction.

A comparison of person-based with vehicle-based optimization reveals that when very few buses appear in the arterial segment under consideration, the proposed person-based optimization results in signal settings and delays that are similar to the corresponding vehicle-based optimization since the underlying formulation of the two algorithms is the same and only the weighting factors differ. The same results would be expected in a scenario when buses carry very few passengers.

The specific outcomes of a person-based optimization and their magnitude is a function of relative auto and bus demands as well as passenger occupancies and signal phase design settings. Most importantly, in stochastic environments, the inability to accurately predict primarily the bus arrival times at the intersections, as well as the platoon size and arrival times, can be an obstacle in observing the potential benefits of the proposed person-based optimization, which requires accurate predictions especially for buses in order to perform well. Despite the fact that the bus arrival predictions are based on average speeds that do not change over time or from one link to another, the promising results of the proposed person-based optimization outperforming the TRANSYT-7F signal settings prove that the proposed algorithm is robust with respect to the bus arrival time inputs. Note also that platoon dispersion or the presence of multiple platoons from turning vehicles, driveways, etc. has not been accounted for. The algorithm is robust with regards to the bus passenger occupancy estimates as well. Additional tests have been performed considering the passenger occupancy of each bus for each cycle as a random variable following a Gaussian distribution with a mean value of 40 and a standard deviation of 10. The results indicate that the differences between the overall, auto, and bus person delays of the randomly varying bus passenger occupancy tests and the constant bus passenger occupancy tests are not statistically significant at the 95% level of confidence. Overall, the results show that the proposed system is robust enough to provide some benefit even when the arrivals and passenger occupancies cannot be predicted accurately since it can still achieve significant person delay reductions for the system even though they might be smaller in magnitude compared to when perfect information is available.

Figure 5 shows the average speed and 95% confidence interval per bus line direction for both the TRANSYT-7F and person-based optimization for the ten runs that were performed under the assumption of stochastic vehicle arrivals. This average speed includes the speeds of all buses

that belong to a specific direction while traversing the links directly affected by the four chosen signalized intersections. The comparison of the average speeds reveals that all bus line directions experience higher average speeds under the proposed person-based signal control system compared to TRANSYT-7F and these increases vary between 1.4% for the southbound direction and 15.3%for the westbound direction at Ashby Avenue. The biggest benefits are achieved for the cross-street buses and in particular for the westbound buses on Ashby Avenue and eastbound buses on Dwight Way, which experience about 15% and 8% increases in their speeds, respectively. Note that these increases are statistically significant at the 95% level of confidence. This is another indication of the success of the person-based real-time signal control system in improving travel speed, and therefore bus schedule adherence and reliability.

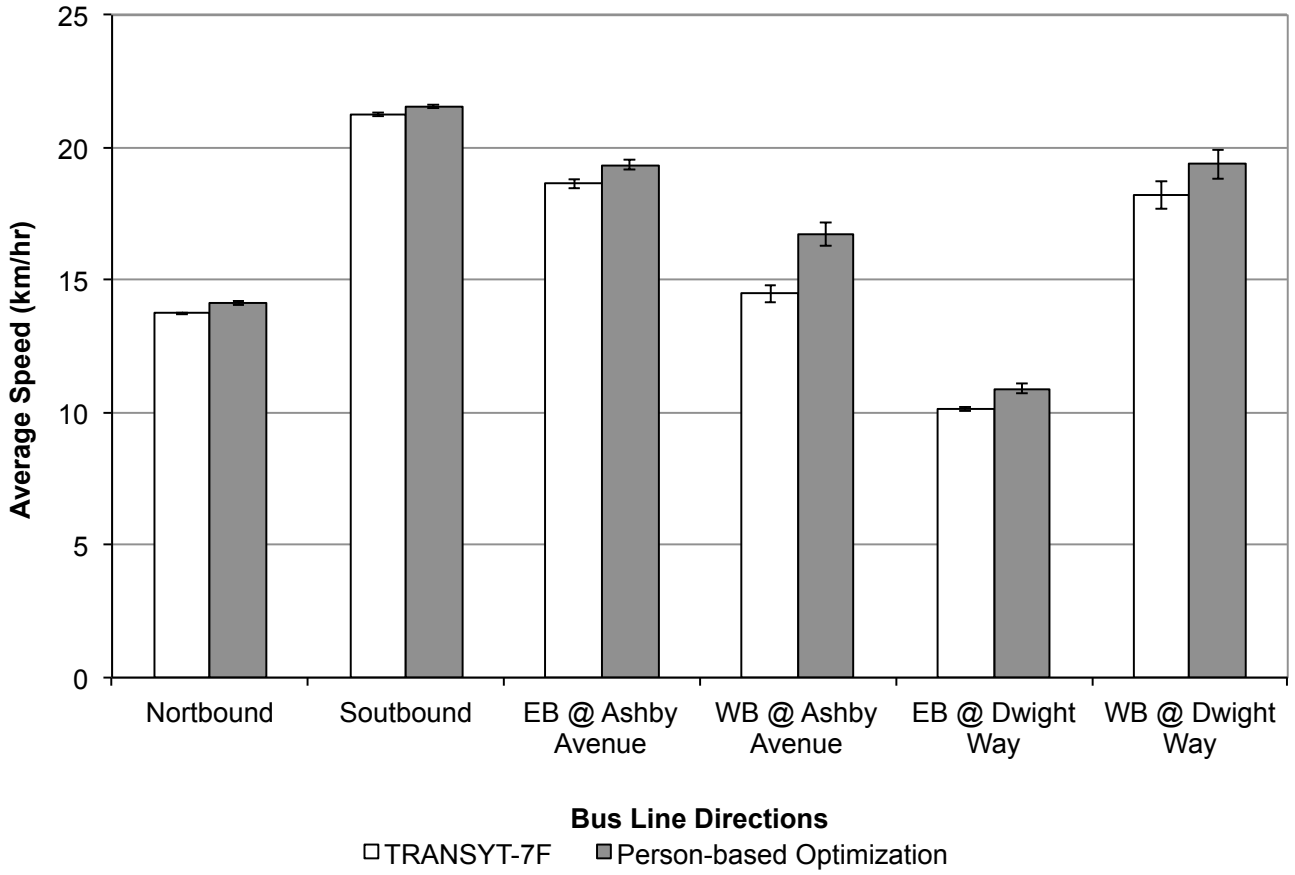


Figure 5: Average Speed Per Bus Line Direction.

Finally, tests are performed under the assumption that all transit vehicles arrive late at the intersections in the network. For these tests, accounting for bus schedule delay translates into weighting the delay of buses by a factor of 2 (i.e., $\delta^r_{b,T} = 1$). The results indicate that the benefit

to bus users improves very slightly to a 4% reduction in their delay with person-based optimization compared to TRANSYT-7F when schedule delay is considered (Table 3). The small magnitude of benefit to buses in this case is due to the fact that despite the higher weight assigned to buses, there is no more flexibility in the signal settings to assign additional priority to transit vehicles. Greater benefits are expected with better bus arrival time prediction accuracy because the proposed system can allocate green times more efficiently to minimize total person delay while accounting for schedule adherence. The ratio of average passenger occupancy of buses over autos will also affect the level of priority provided.

Table 3: System person and vehicle delays for $\bar{o}_b/\bar{o}_a = 40/1.25$, $\delta^r_{b,T} = 1$ and 1 hour of traffic operations (simulation).

	Auto		Bus		Overall	
	Total (pax-hrs)	Average (sec/veh)	Total (pax-hrs)	Average (sec/veh)	Total (pax-hrs)	Average (sec/veh)
TRANSYT-7F	78.87 (4.35)	20.64 (0.99)	43.62 (0.72)	45.44 (0.71)	122.49 (4.41)	25.63 (0.80)
Person-based	74.36 (3.67)	19.46 (0.90)	41.86 (0.85)	43.56 (1.07)	116.22 (3.91)	24.30 (0.77)
% Change	-5.7%	-5.7%	-4.0%	-4.1%	-5.1%	-5.2%

5. Conclusions

The paper has presented an arterial-level person-based traffic responsive signal control system. The proposed pairwise optimization method explicitly accounts for passenger occupancy to provide priority to transit vehicles by minimizing person delay at two consecutive intersections at a time. In addition to accounting for auto vehicle progression, by assigning the appropriate delays for interrupting the platoons, the system recognizes the importance of schedule adherence for reliable transit operations. Therefore, it introduces an additional weighting factor that reflects how early or late a transit vehicle is or whether it is early or late. The passenger occupancy and schedule adherence weighting factors facilitate priority assignment decisions for conflicting transit routes.

The proposed system is flexible because the user can choose to weigh transit delay and delays caused by interrupting the head or tail of the platoon differently to reach different objectives. In addition, one can choose different weights for the delays of different modes. The system relies on currently deployable technologies that can provide real-time information for estimating platoon

sizes and arrival times at intersections (e.g., demand, travel times, and turning ratios from detectors with appropriate communication systems), transit vehicle arrival times (e.g., speed and location from AVL systems), and transit vehicle passenger occupancies (e.g., APC systems). In addition, the proposed system has low computation time. This is due to the fact that the optimization is performed for two consecutive intersections simultaneously, which substantially reduces the computation effort compared to systems that optimize multiple intersections at a time. Furthermore, since the mathematical program has been formulated as a MILP, a solution can be obtained in a few seconds. This contributes to the feasibility and economic viability of its implementation in real-world settings and large networks. Therefore, the proposed system is advantageous over existing adaptive signal control systems, which are complex and expensive to implement in real networks (Stevanovic, 2010).

The system incorporates the delays for all turning movements that are served by protected phases by including them in the objective function. However, there are three types of delays from turning vehicles that are not accounted for: 1) the delay associated with permissive right turns on red, 2) the extra delay that permissive left turning vehicles experience while waiting to make their turn, and 3) the delay that vehicles turning onto the arterial experience at downstream intersections on the main arterial. While the first two types of delay are not expected to be substantial compared to all the other vehicles' delays, the third type can result in changes in the optimal solution. These vehicle delays can be accounted for by grouping all vehicles for each approach into smaller platoons as it was done in He et al. (2012). In addition, slight modifications to the mathematical program formulation should be made to appropriately capture delays for more than one platoon per lane group. Overall, any levels of turning percentages can be accommodated by the proposed real-time signal control system as long as there is no queue spillover on the turning pockets affecting the capacity of the through lanes. In cases that queue spillovers occur from turning pockets, the system could still be used by adjusting the capacity of the through lanes in real-time.

The system has been evaluated with deterministic arrival tests under the assumption of perfect information about inputs for demand and arrival times and through simulation to test its performance under more realistic traffic and transit operations. Deterministic and stochastic arrival tests have shown that the person-based arterial traffic responsive signal control system outperforms static signal settings provided by TRANSYT-7F, even for auto person delays. However, the

specific outcomes of a person-based optimization and their magnitude depends on the direction of travel as well as other features such as bus frequencies, auto demands, passenger occupancies, and signal phase design settings. In addition, the success of the proposed system and the magnitude of the benefit that can be achieved is dependent on the accuracy of real-time arrival predictions when implemented in stochastic simulated and real-world environments. Additional tests should be performed to investigate the impact of bus and platoon arrival accuracies on the benefits achieved by the person-based signal control system. In addition, improved arrival prediction algorithms and adjustments in the delay equations to account for input inaccuracies are expected to improve its performance. The benefits of the proposed person-based signal control system are also expected to be higher under varying auto demand conditions since the system can respond to changes from cycle to cycle.

The tests have also shown that buses traveling on cross streets with low auto demand experience very high benefits when transit priority is provided. This is due to the much higher weight the cross street vehicle delays get when a bus is present and person-based optimization is used compared to when TRANSYT-7F signal settings are used. In addition, passengers traveling against the heaviest direction of the main arterial also benefit from a person-based optimization compared to TRANSYT-7F whose optimal signal settings benefit primarily the heaviest direction of traffic. Bus speeds have been found to increase by up to 15.3% when person-based optimization is in place compared to TRANSYT-7F. Finally, accounting for schedule delay provides additional benefit to transit users as long as there is still flexibility in the signal settings to achieve higher reductions in person delay by providing more priority to transit vehicles. Overall, the proposed system does not distinguish between different approaches, providing priority to directions when appropriate and being able to address the issue of conflicting transit lines in an efficient way.

The proposed system is promising for reducing person delay for an arterial segment and providing priority to transit vehicles even when multiple transit routes run in conflicting directions at intersections. It can be effectively combined with the recently proposed three-dimensional macroscopic fundamental diagram for bi-modal traffic (Geroliminis et al., 2014) to improve the aggregate performance of autos and buses in a region. For example, in a hierarchical two-level control approach, the first level could control a region within a network so that the aggregate performance of autos and buses can be improved while the second level could be utilizing the proposed system to improve traffic operations at the local level and provide priority to transit vehicles. In addi-

tion, due to its generic formulation, this person-based optimization approach can be applied to control operations on complete streets where multiple modes exist (i.e., autos, transit vehicles, pedestrians, bicyclists).

Ongoing and future work includes comparing the system performance with other dynamic coordination strategies and adaptive signal control systems, incorporating pedestrian delays in the system, performing additional sensitivity analysis tests for a variety of bus to auto passenger occupancy ratios, schedule delay weighting factors, geometric configurations, and demand levels. It is expected that the benefit to transit vehicles will be smaller with lower bus to auto passenger occupancy ratios as it was shown in Christofa and Skabardonis (2011) and Christofa et al. (2013) for isolated intersections as well as with higher auto vehicle demands. In addition, we plan to improve the arrival prediction algorithms and adjust the delay equations to account for inaccuracies in the input estimates and loss of capacity due to queue spillbacks. Future work will also relax the assumption of no platoon dispersion and incorporate the delays of turning vehicles. Finally, we plan to expand the system to arterial signalized networks by using the proposed pairwise optimization method along multiple arterials.

References

AC Transit, 2011. Alameda-Contra Costa Transit District. http://www.actransit.org. Accessed August 2011.

Bretherton, D., Bowen, G., Wood, K., 2002. Effective urban traffic management and control: SCOOT Version 4.4. European Transport Conference .

Chang, G.L., Vasudevan, M., Su, C.C., 1996. Modelling and evaluation of adaptive bus-preemption control with and without automatic vehicle location systems. Transportation Research Part A: Policy and Practice 30, 251–268.

Christofa, E., 2012. Traffic signal optimization with transit priority: A person-based approach. Ph.D. thesis. University of California, Berkeley.

Christofa, E., Papamichail, I., Skabardonis, A., 2013. Person-based traffic responsive signal control optimization. IEEE Transactions on Intelligent Transportation Systems 14(3), 1278–1289.

Christofa, E., Skabardonis, A., 2011. Traffic signal optimization with application of transit signal priority to an isolated intersection. Transportation Research Record 2259, 192–201.

Conrad, M., Dion, F., Yagar, S., 1998. Real-time traffic signal optimization with transit priority: Recent advances in the signal priority procedure for optimization in real-time model. Transportation Research Record 1634, 100–109.

Cornwell, P., Luk, J., Negus, B., 1986. Tram priority in SCATS. Traffic Engineering and Control 27, 561–565.

Diakaki, C., Dinopoulou, V., Aboudolas, K., Papageorgiou, M., Ben-Shabat, E., Seider, E., Leibov, A., 2003. Extensions and new applications of the traffic-responsive urban control strategy: Coordinated signal control for urban networks. Transportation Research Record: Journal of the Transportation Research Board 1856, 202–211.

Farid, Y.Z., Christofa, E., Collura, J., 2014. Person-based evaluation of transit preferential treatments on signalized arterials, in: Transportation Research Board 93rd Annual Meeting.

Geroliminis, N., Zheng, N., Ampountolas, K., 2014. A three-dimensional macroscopic fundamental diagram for mixed bi-modal urban networks. Transportation Research Part C: Emerging Technologies 42, 168–181.

Hale, D., 2009. Transyt-7f users guide, mctrans center. University of Florida, Gainesville, FL .

He, Q., Head, K., Ding, J., 2012. Pamscod: Platoon-based arterial multi-modal signal control with online data. Transportation Research Part C: Emerging Technologies 20, 164–184.

He, Q., Head, K.L., Ding, J., 2014. Multi-modal traffic signal control with priority, signal actuation and coordination. Transportation Research Part C: Emerging Technologies 46, 65–82.

Henry, J., Farges, J., 1994. P.T. priority and Prodyn. Proceedings of the 1st World Congress on Application of Transport Telematics and Intelligent Vehicle-Highway Systems 6, 3086–3093.

Hu, J., Park, B., Leeb, Y.J., 2015. Coordinated transit signal priority supporting transit progression under connected vehicle technology. Transportation Research Part C: Emerging Technologies. http://dx.doi.org/10/1016/j.trc.2014.12.05.

Hunt, P., Bretherton, R., Robertson, D., Royal, M., 1982. SCOOT on-line traffic signal optimisation technique. Traffic Engineering and Control 23, 190–192.

IBM, 2011. IBM ILOG CPLEX, Version 12.1: High performance mathematical programming engine. http://www.ilog.com/products/cplex.

Li, Y., Koonce, P., Li, M., Zhou, K., Li, Y., Beaird, S., Zhang, W., Hegen, L., Hu, K., Skabardonis, A., et al., 2008. Transit signal priority research tools. PATH Research Report UCB-ITS-PRR-2008-4. California Partners for Advanced Transit and Highways, University of California, Berkeley.

Lighthill, M., Whitham, G., 1955. On kinematic waves II. A theory of traffic flow on long crowded roads. Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences 229, 317–345.

Lin, Y., Yang, X., Chang, G.L., Zou, N., 2013. Transit priority strategies for multiple routes under headway-based operations. Transportation Research Record: Journal of the Transportation Research Board 2356, 34–43.

Ma, W., Liu, Y., Yang, X., 2013a. A dynamic programming approach for optimal signal priority control upon multiple high-frequency bus requests. Journal of Intelligent Transportation Systems 17, 282–293.

Ma, W., Ni, W., Head, L., Zhao, J., 2013b. Effective coordinated optimization model for transit priority control under arterial progression. Transportation Research Record: Journal of the Transportation Research Board 2356, 71–83.

Mauro, V., Di Taranto, C., 1989. UTOPIA. Proceedings of the 6th IFAC-IFIP-IFORS Symposium on Control, Computers, and Communications in Transportation , 245–252.

Meyer, C., Floudas, C., 2004. Trilinear monomials with mixed sign domains: Facets of the convex and concave envelopes. Journal of Global Optimization 29, 125–155.

Newell, G., 1964. Synchronization of traffic lights for high flow. Quarterly of Applied Mathematics 21, 315–324.

Newell, G., 1967. Traffic signal synchronization for high flows on a two-way street. Research Report. Institute of Transportation and Traffic Engineering, University of California.

Richards, P., 1956. Shock waves on the highway. Operations Research 4, 42–51.

Stevanovic, A., 2010. Adaptive traffic control systems: Domestic and foreign state of practice, NCHRP Synthesis 403, Transportation Research Board. National Academy of Sciences, Washington DC .

Stevanovic, J., Stevanovic, A., Martin, P.T., Bauer, T., 2008. Stochastic optimization of traffic control and transit priority settings in VISSIM. Transportation Research Part C: Emerging Technologies 16, 332–349.

Sun, X., Zeng, X., Zhang, Y., Quadrifoglio, L., 2015. Person-based adaptive priority signal control with connected vehicle information, 15-4743. 94th Annual Meeting of the Transportation Research Board .

TSS, 2010. Aimsun users manual v6.1. Transport Simulation Systems, Barcelona, Spain.

Vasudevan, M., 2005. Robust optimization model for bus priority under arterial progression. Ph.D. thesis. Deapartment of Civil Engineering, University of Maryland, College Park.

Zeng, X., Zhang, Y., Balke, K.N., Yin, K., 2014. A real-time transit signal priority control model considering stochastic bus arrival time. IEEE Transactions on Intelligent Transportation Systems 15.