https://doi.org/10.1016/j.trc.2021.103211

An integrated intersection design for promoting bus and car traffic

Weihua Gu*[a], Yu Mei[a], Haoyu Chen[b], Yiguang Xuan[b], Xiaochun Luo[c]

[a] Department of Electrical Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China
[b] Institute of Transportation Studies, University of California, Berkeley, CA 94720, United States
[c] Department of Building and Real Estate, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China

Abstract

Signal priority and exclusive bus lanes are common measures used to promote bus travel at signalized intersections. However, these bus priority measures create significant damage to the car traffic especially at intersections with heavy bus and car traffic. For this reason, bus priority is not welcome at busy intersections. We examine a novel intersection approach design that can solve the above dilemma. The design integrates a bus lane, bus signal priority, and a midblock pre-signal for sorting different car traffic streams in tandem in the approach. Car capacity gains from the use of pre-signal can potentially recover the car capacity lost to bus lane and signal priority schemes. This paper first presents how the pre-signal and main signal can be timed to realize bus signal priority, and where the pre-signal should be placed. Models are then formulated for estimating the expected bus delay and car capacity under the integrated design. They are compared against three alternative designs, including a conventional intersection design without bus priority or pre-signal. Numerical results unveil that the integrated design produces not only significant bus delay savings, but also higher car capacities in most instances. Even greater car capacity gains are observed with higher bus frequencies. Moreover, the benefits are fairly robust when real-world operating features, such as bus arrival time prediction error, are considered. Thus, the integrated design can potentially promote both bus and car traffic at congested intersections.

Keywords: public transport; pre-signal; bus signal priority; bus lane; tandem intersection design

1. Introduction

Bus signal priority (BSP) schemes were often used to reduce bus delays at signalized intersections (e.g., Garrow and Machemehl, 1997; Ma et al., 2014). Their effectiveness has been well proved by real-world case studies (Dion et al., 2004). However, BSP often diminishes the green times received by the car traffic (Dion et al., 2004). Worse still, BSP was commonly deployed jointly with a dedicated bus lane (DBL); see Zhou and Gan (2005) and Farid et al. (2018)[1]. Excluding cars from using the DBL would further reduce the car discharge capacity at an intersection. As a result, cars will suffer greater delays and may form queues that spill back to an upstream intersection. Therefore, BSP was unwelcome at intersections with heavy car traffic (Abdy and Hellinga, 2011), and, even more so, at those where BSP would be frequently activated due to a high bus flow.

Solutions to the above problem were proposed, most of which aim to alleviate the car capacity loss by allowing cars to use the bus lane when no bus is present. They are termed the "intermittent bus lane (IBL)" strategies (Wu and Hounsell, 1998; Viegas and Lu, 2004; Guler and Menendez, 2014; He et al., 2016). A mid-block pre-signal was often used in those strategies to control the entry of cars into the IBL. Apparently, those strategies cannot fully recover the car capacity loss caused by the bus lane. Moreover, their benefits would vanish when the bus flow is high (Qiu et al., 2014).

* Corresponding author. Email: weihua.gu@polyu.edu.hk.
[1] Although there exist BSP schemes for prioritizing buses traveling in mixed-traffic lanes (e.g., Bie et al., 2011), those schemes only produced limited benefits since buses were often blocked by car queues.

On the other hand, research efforts have also targeted at increasing a signalized intersection's vehicle discharge capacity via unconventional intersection designs. Examples include the continuous flow intersection design (Yang and Cheng, 2017), the tandem intersection design (Xuan et al., 2011), the left-turn waiting area (Ma et al., 2017), and their variants (e.g., Wu et al., 2019). These unconventional designs can reduce or eliminate conflicts between left-turning and through-moving traffic streams at the intersection (assuming right-hand traffic). The two streams can thus discharge in the same phase (the case of the continuous flow intersection design), or discharge via all lanes of an approach in separate phases (the case of the tandem intersection design). In all these designs, the number of lanes that are simultaneously used to discharge vehicles is increased, resulting in significant capacity improvements. Integrating the above designs with the bus priority measures can potentially compensate the capacity loss created by the latter.

In light of the above, this paper examines a novel intersection design that integrates a DBL, a BSP scheme, and the tandem design as described in Xuan et al. (2011). We expect that this integrated intersection design can prioritize buses at a busy approach without undermining its capacity for serving the car traffic.[2] The tandem design is selected because it only involves limited change of layout in one intersection approach, and its resulting capacity gain is relatively large. Its benefit has also been verified by real-world applications (e.g., Luo, 2011). For simplicity, we focus on a single approach with three or more lanes; see Figure 1a for an example with three mixed-traffic lanes. Under the integrated design, one of the mixed-traffic lanes is converted into a bus lane, and a pre-signal is installed in the remaining lanes to sort the left-turning (henceforth *L*-) and through-moving (henceforth *T*-) traffic streams; see the layout in Figure 1b[3]. The pre-signal will alternately admit L- and T-vehicles into the lane area between the pre-signal and the intersection (termed the *sorting area*). The two types of vehicles queue up in tandem in the sorting area. Each queue will then use all lanes in the sorting area to discharge during a separate signal phase; see Figure 2.

The integrated design was first presented in an earlier version of this paper (Gu et al., 2015) and Chen (2016). A recent study (Bie et al., 2020) examined selected instances of this design via simulation. Partly due to the limited number of instances they examined, the latter work failed to unveil how a number of operating factors affect the performance of this design, and general insights into the trade-off between car discharge capacity and bus delay savings. For a fuller assessment of the integrated design under various operating conditions, models formulated for the general case are desired.

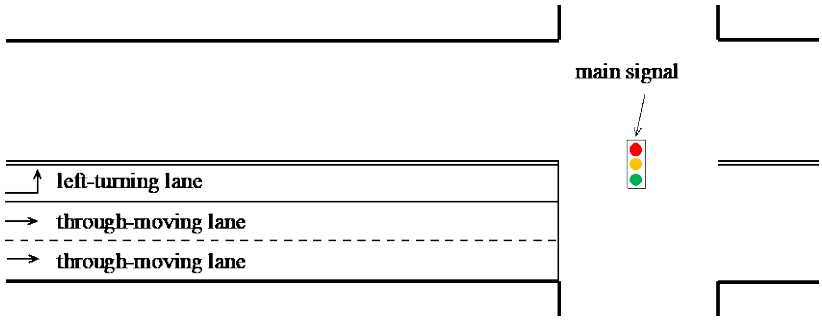
We propose adaptive, coordinated timing plans for the pre-signal and the signal at the intersection (termed the *main signal* henceforth) when a BSP scheme is incorporated. A formula for estimating the sorting area length is also developed. Models for estimating two primary performance metrics, the expected bus delay and the car discharge capacity[4], are then formulated. By comparing against alternative designs that do not include the pre-signal and bus priority measures simultaneously, we show that the integrated design can significantly reduce bus delays while still achieving some car capacity gains. This facilitates the deployment of bus priority measures at busy intersections where both car and bus flows are high, which was believed to be unprofitable by conventional wisdom (Abdy and Hellinga, 2011; Lin et al., 2019). We also show that the advantages of this design are still significant

[2] By "capacity", we mean the maximum rate that cars can discharge into the intersection over the long run, which is attained when a persistent car queue is present.

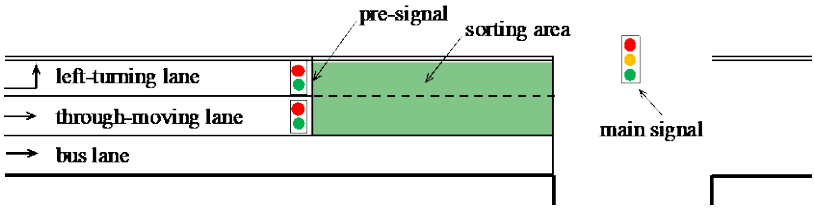
[3] In the rest of this paper, the word "pre-signal" refers to the one used to sort L- and T-traffic of the subject approach (see Figure 1b); i.e., the pre-signal does not control buses.

[4] Here we choose to estimate the car discharge capacity instead of car delays at an intersection because the former is generally easier to formulate analytically (Gu et al., 2014; Gayah et al., 2016).

78 when real-world operational issues, e.g., uncertainties in actual bus arrival times and car discharge flows,
79 are accounted for.

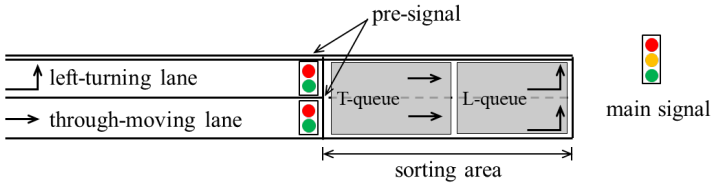


80
81 **(a) The original intersection approach**



82
83 **(b) The integrated design**

84 **Figure 1. A 3-lane intersection approach**



85
86 **Figure 2. Car lanes controlled by the pre-signal**

87 The rest of the paper is organized as follows. Section 2 describes the integrated design in detail.
88 Section 3 formulates models for estimating the car discharge capacity and expected bus delay under the
89 integrated design and three alternatives. Numerical case studies and parametric analyses are furnished
90 in Section 4. Section 5 discusses how the models can be extended to incorporate some real-world
91 concerns, including the joint design of two opposing approaches of the intersection, and operational
92 uncertainties. Findings and insights are summarized in Section 6.

93 2. The Integrated Intersection Design

94 Section 2.1 introduces the major assumptions used in this paper. Sections 2.2 and 2.3 describe the phase
95 plans of the main signal and the pre-signal, respectively. Section 2.4 presents the BSP scheme. Section
96 2.5 develops the minimum length of the sorting area. The notations used in this paper are summarized
97 in Appendix A.

98 2.1 Assumptions

99 The following assumptions are made for the simplicity of analysis. Similar assumptions have been
100 commonly used in studies on bimodal (i.e., buses and cars) traffic operations at signalized intersections
101 (e.g., Gu et al., 2013; 2014; Guler and Menendez, 2014).

i) Only through-moving buses are considered.
ii) Bus arrivals to the approach follow a stationary Poisson process with rate λ_b, ranging from 0 to around 120 bus/h. This assumption ensures that the number of bus arrivals in each signal cycle is small, and that they are randomly distributed among the car traffic.
iii) Bus arrival times to the intersection can be predicted accurately using loop detectors, on-board GPS, or connected vehicle technology.
iv) Only the BSP requests in the subject approach are considered (and we do not impose a limit on the number of requests entertained). No conflicting requests (e.g., those from the cross-street approaches) will be handled.
v) The car traffic in the approach can be described using a triangular fundamental diagram, which is determined using historical traffic information.
vi) Right-turning vehicles are ignored.[5]

2.2 Main signal phase plan

The main signal has a fixed cycle length denoted by T. The signal phase plan is illustrated in Figure 3, where G_L and G_T denote the green phase durations for the L- and T-traffic in the subject approach, respectively; and R_L and R_T the phase durations for the cross-street traffic. An amber duration of $t_y = 4$ s is inserted between any two consecutive phases. Thus, $T = G_L + G_T + R_L + R_T + 4t_y$. We assume this signal phase plan because it entails a shorter sorting area (Xuan, 2011). Tandem designs with long sorting areas are difficult to implement in short city blocks. The phase durations will be optimized by models furnished in Section 3.

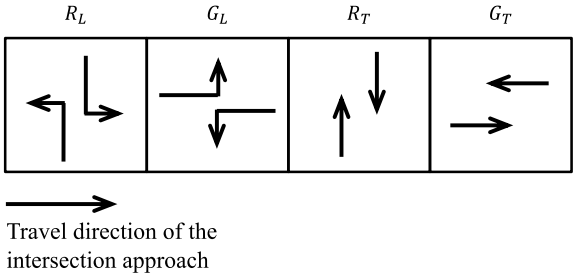


Figure 3. Main signal phase plan

2.3 Pre-signal phase plan

The pre-signal's cycle length is equal to that of the main signal. It contains two green phases: a g_L phase that serves L-vehicles and a g_T phase that serves T-vehicles. Between g_L and g_T is an amber phase of duration t_y (thus $g_L + g_T + 2t_y \leq T$), and a red phase to fill up any gap in a cycle. To ensure all vehicles that pass a pre-signal green phase can always discharge into the intersection during a corresponding main signal phase, the phase durations must satisfy $g_T n_T \leq G_T M$ and $g_L n_L \leq G_L M$, where n_T and n_L denote the numbers of lanes allocated to T- and L-vehicles upstream of the pre-signal, respectively; and M the number of lanes in the sorting area. The pre-signal phase durations are also optimized by models furnished in Section 3.

The coordination between the pre-signal and main signal timing plans is illustrated in Figures 4a-c. Each figure is a time-space diagram that describes a distinct case of the traffic states in the sorting area during a cycle. These states are: the no-traffic state denoted by ***O***; the jam (queueing) state denoted

[5] In reality, right-turning cars can use a curbside bus lane if their maneuvers do not impede the buses; e.g., see the policy enacted by the City of New York Government (2014).

by $\boldsymbol{J}$; the saturated state as queued vehicles are discharging into the intersection, denoted by $\boldsymbol{S}$; and the states for L- and T-vehicles discharging from the pre-signal into the sorting area, denoted by $\boldsymbol{A_L}$ and $\boldsymbol{A_T}$, respectively. These time-space diagrams are created using a triangular fundamental diagram (Lighthill and Whitham, 1955; Richards, 1956; Newell, 1993) for the sorting area, which is shown in Figure 4d. In the fundamental diagram, v_f denotes the free-flow travel speed; w the backward wave speed; q_S the saturation flow per travel lane; q_T and q_L the saturation flows of T- and L-vehicles discharging from the pre-signal, respectively; and w_T and w_L the speeds of formation shockwaves for T- and L-vehicle queues in the sorting area, respectively. We have $q_T = q_S n_T$ and $q_L = q_S n_L$. The w_T and w_L are formulated using the kinematic wave theory as follows:

$$w_T = \frac{q_T}{Mq_S\left(v_f^{-1}+w^{-1}\right)-\frac{q_T}{v_f}} = \frac{n_T}{(M-n_T)v_f^{-1}+Mw^{-1}}, \tag{1a}$$

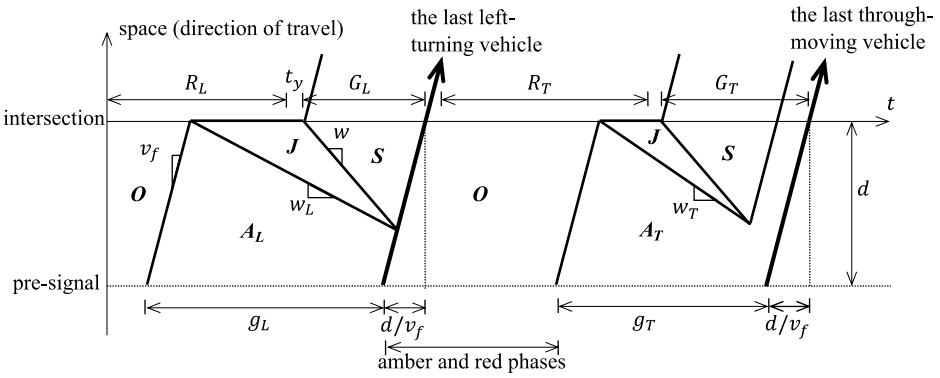
$$w_L = \frac{n_L}{(M-n_L)v_f^{-1}+Mw^{-1}}. \tag{1b}$$

Specifically, Figures 4a illustrates the case where the L- and T-queues in the sorting area (marked by state $\boldsymbol{J}$) are temporally separated. This case, numbered case i), occurs where $g_L \le R_L + G_L + t_y$ and $g_T \le R_T + G_T + t_y$ are both satisfied, as can be learned from the figure. In this case, g_L and g_T are timed such that the last vehicle in each phase can discharge into the intersection without any delay in the sorting area; see the bolded vehicle trajectories in the figure. Specifically, g_L and g_T end a duration $\frac{d}{v_f}$ earlier than G_L and G_T, respectively, where d denotes the sorting area length.

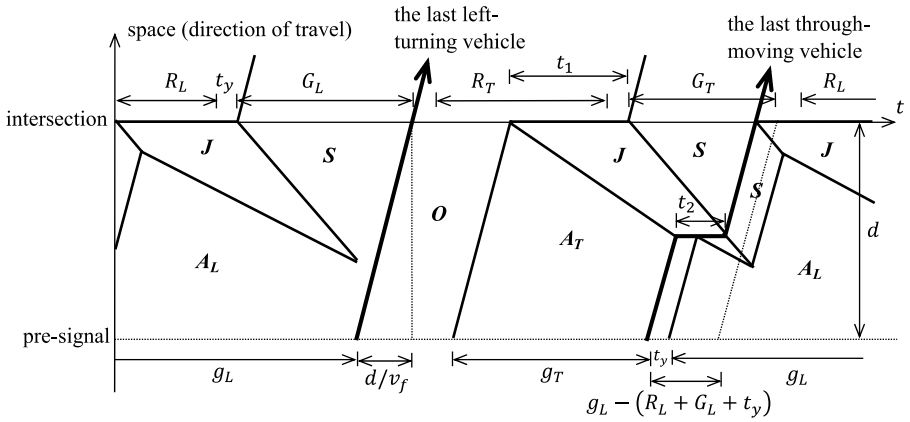
Figure 4b illustrates case ii), where a small L-queue may be superimposed on top of a T-queue. This case occurs where $g_L > R_L + G_L + t_y$ but $g_T < R_T + G_T + t_y$. In this case, the g_L phase still ends $\frac{d}{v_f}$ earlier than the corresponding G_L. However, the long g_L phase will "push" the preceding g_T phase toward left in the time axis by $g_L - \left(R_L + G_L + t_y\right)$. As a result, the last T-vehicle passing in g_T may experience a delay in the sorting area.

Finally, Figure 4c shows case iii) where a T-queue may be superimposed on top of a L-queue. This occurs where $g_T > R_T + G_T + t_y$ but $g_L < R_L + G_L + t_y$. Contrary to case ii), now the long g_T is timed to end $\frac{d}{v_f}$ earlier than the corresponding G_T, while it "pushes" the preceding g_L toward left by $g_T - \left(R_T + G_T + t_y\right)$.

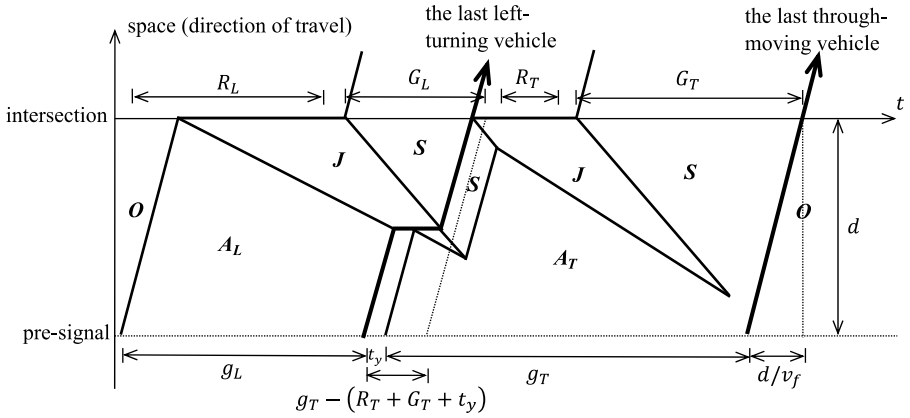
Note that the above three cases cover all the possibilities, since $g_L + g_T \le T - 2t_y = \left(R_L + G_L + R_T + G_T + 4t_y\right) - 2t_y = \left(R_L + G_L + t_y\right) + \left(R_T + G_T + t_y\right)$, which means $g_L > R_L + G_L + t_y$ and $g_T > R_T + G_T + t_y$ cannot be both true.



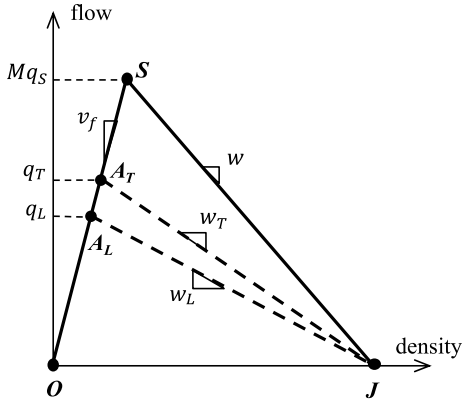
(a) Case i) (when $g_T \le R_T + G_T + t_y$ and $g_L \le R_L + G_L + t_y$)



(b) Case ii) (when $g_L > R_L + G_L + t_y$)



(c) Case iii) (when $g_T > R_T + G_T + t_y$)



(d) Fundamental diagram of the sorting area

Figure 4. Coordination between main signal and pre-signal timing plans without BSP

2.4 Bus signal priority

For simplicity, this paper only examines a single type of BSP scheme: green extension. We denote t_m as the maximum allowable extension of a green phase.[6] For all buses that are predicted to arrive at the intersection within t_m after the end of the last G_T phase, that G_T phase will be extended to let those buses pass without delay. For other bus arrivals, no priority is granted. The actual duration of extension,

[6] The choice of t_m depends on a number of operating factors and concerns. A discussion on this issue is provided in Section 6.1.

denoted by t_e ($t_e \leq t_m$), is determined by the arrival time of the last prioritized bus. We choose green extension because this scheme benefits the buses that would otherwise just miss the green signal and experience long delays. Nevertheless, other BSP schemes (e.g., red truncation and green insertion) can be incorporated into our modeling framework in similar ways.

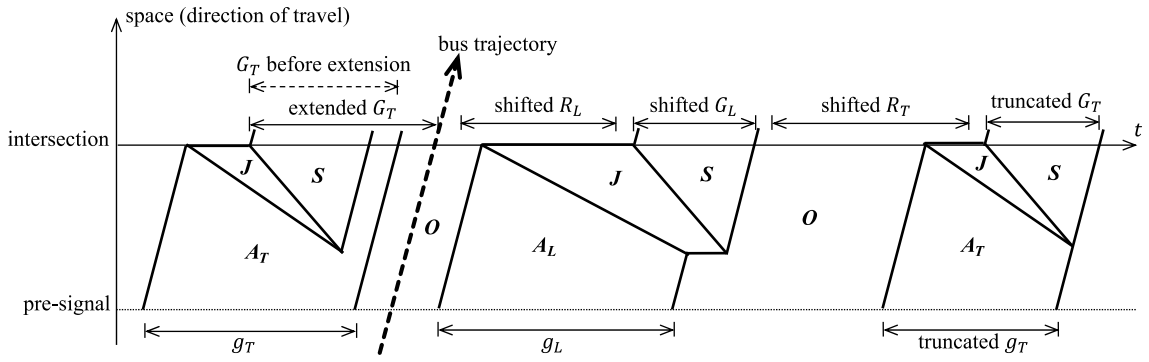
We further specify that, when a G_T phase is extended, the following phases R_L, G_L, and R_T are postponed accordingly but not shortened. This ensures that the discharge capacities for the other traffic streams at the intersection are not compromised. The next G_T phase is trimmed to ensure all future cycles are not affected by the green extension. Hence, the total green time allocated to T-traffic (including the buses) is unchanged. This also indicates that $t_m \leq G_T$.

The pre-signal phase plan is also altered adaptively in response to a green extension. These alterations are illustrated in Figures 5a-c for the three cases described in Section 2.3. Bus trajectories are plotted as bolded, dashed arrows in these figures.

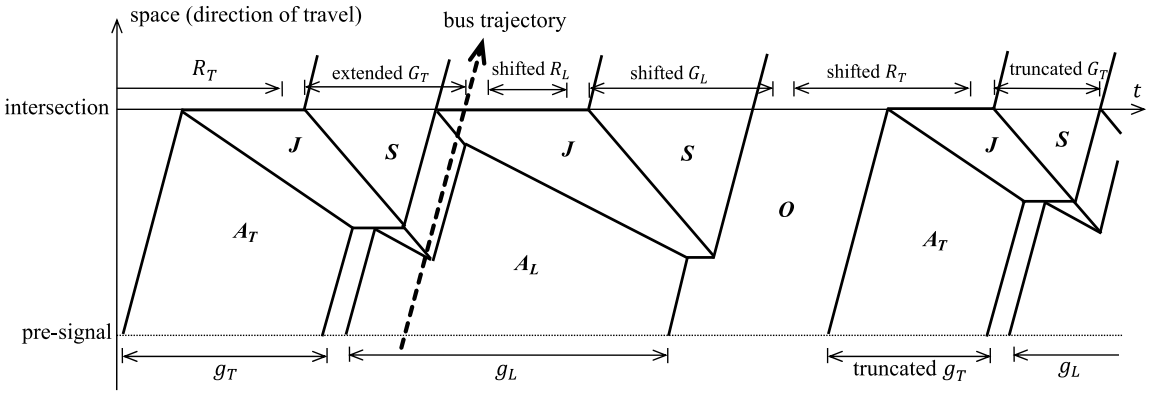
Figure 5a describes the simple case with temporally separated T- and L-queues. As shown in the figure, only the g_T phase corresponding to the truncated G_T is shortened to ensure every vehicle that enters the sorting area can discharge into the intersection without being delayed to the next cycle. Mathematically, the truncated g_T phase duration can be calculated by:

$$\tilde{g}_T = \min\left\{g_T, \tilde{G}_T \cdot \frac{M}{n_T}\right\} \tag{2}$$

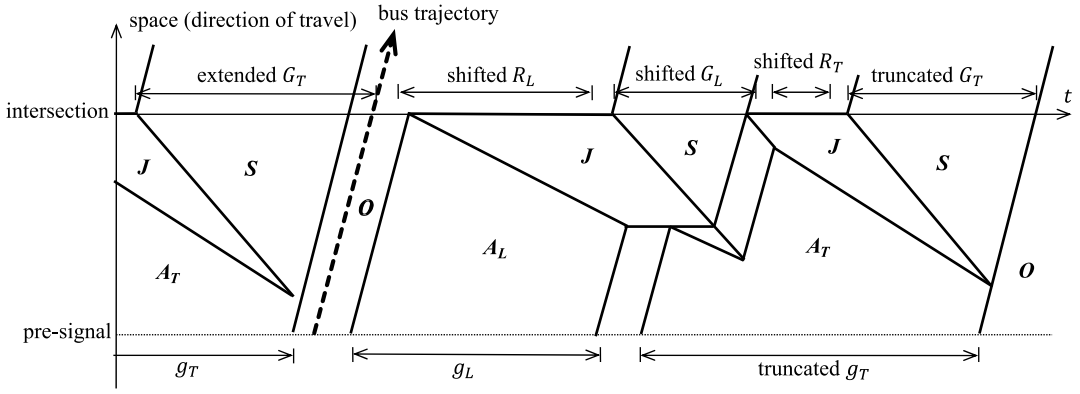
where $\tilde{g}_T$ and $\tilde{G}_T$ denote the durations of truncated g_T and G_T phases, respectively. All the other pre-signal phases stay unchanged. Although the first g_T phase in the figure can be possibly extended toward left to better utilize the extended G_T, this requires earlier detection of bus arrivals, and is ignored here for brevity and conservativeness.



(a) Case i) in Section 2.3 (separated T- and L-queues)



(b) Case ii) in Section 2.3 (when $g_L > R_L + G_L + t_y$)



(c) Case iii) in Section 2.3 (when $g_T > R_T + G_T + t_y$)

Figure 5. Pre-signal's adaptive coordination with green extension

Figures 5b and c illustrate the two tandem queue cases in the sorting area. Similarly, only the g_T phase corresponding to the truncated G_T is truncated.

2.5 Length of sorting area

To contain the vehicle queues, the sorting area's length, d, should be no less than a threshold value d_{min}. An approximation of d_{min} is presented as follows:

$$d_{min} \approx \begin{cases} \max\left\{g_L\left(w_L^{-1}+v_f^{-1}\right)^{-1}, g_T\left(w_T^{-1}+v_f^{-1}\right)^{-1}\right\}, & \text{if } g_T \le R_T+G_T+t_y \text{ and } g_L \le R_L+G_L+t_y \\ \max\left\{g_L\left(w_L^{-1}+v_f^{-1}\right)^{-1}, g_T\left(w_T^{-1}+v_f^{-1}\right)^{-1} + \max\left\{\frac{g_T n_T}{M}+g_L-\left(G_T+R_L+G_L+2t_y\right),0\right\}\cdot\left(w_L^{-1}-w^{-1}\right)^{-1}\right\} & \\ \qquad \text{if } g_L > R_L+G_L+t_y & \\ \max\left\{g_T\left(w_T^{-1}+v_f^{-1}\right)^{-1}, g_L\left(w_L^{-1}+v_f^{-1}\right)^{-1} + \max\left\{t_m+\frac{g_L n_L}{M}+g_T-\left(G_L+R_T+G_T+2t_y\right),0\right\}\cdot\left(w_T^{-1}-w^{-1}\right)^{-1}\right\} & \\ \qquad \text{if } g_T > R_T+G_T+t_y & \end{cases} \quad (3)$$

The derivation of (3) is relegated to Appendix B.

We next formulate models for assessing the car discharge capacity and expected bus delay under the above design.

3. Models for expected bus delays and car capacities

To better understand the integrated design's performance, we compare it against the following three alternatives:

i) A conventional design where none of the pre-signal, the DBL, and the BSP is used;
ii) A pre-signal-only design where a pre-signal is installed to sort L- and T-vehicles (including both cars and buses), but neither the DBL nor the BSP is implemented; and
iii) A BSP-only design where no pre-signal is used, but buses enjoy a DBL and green extension.

The differences between the four alternatives are highlighted in Table 1.

Table 1. Definition of alternative designs

	No BSP, no DBL	**BSP and DBL**
No pre-signal	Conventional	BSP-only
Pre-signal	Pre-signal-only	Integrated

Sections 3.1-3.4 present models for estimating the expected bus delay and the T-cars' discharge capacity under the four alternatives, respectively. We estimate the latter metric because only the capacity of T-cars is reduced by the BSP scheme; see again Section 2.4.

Prior to presenting the bus delay and car capacity models, the signals' phase durations and the numbers of lanes allocated to L- and T-traffic in the approach are optimized. To this end, we use models that are similar to those furnished in Xuan et al. (2011). The models maximize the approach's overall discharge capacity for both L- and T-traffic for a given number of lanes in the approach, N, left-turning ratio, l, and effective green time allocated to the approach, $G = G_L + G_T$. The effect of BSP is ignored (at this stage only) for simplicity. Solutions to those models consist of the optimal G_L, G_T, g_L, g_T, n_L and n_T under the integrated design and the pre-signal-only alternative, and G_L, G_T, N_L and N_T under the conventional and BSP-only alternatives, where N_L and N_T denote the numbers of L- and T-lanes upstream of the intersection, respectively. Note that for simplicity of the notations, each variable above is used in more than one alternative designs. They are however optimized under each alternative separately. For example, all the four designs use G_T to indicate the green duration for T-traffic, but the optimal value of G_T is different under each design. Detailed formulation and solution method of those optimization models are furnished in Appendix C.

3.1 Models under the conventional design

Since Poisson bus arrivals are uniformly distributed over time (Ross, 2014), and they are sparse in the car traffic (see Assumption ii) in Section 2.1), the expected bus delay, denoted by $E[W_{B,1}]$, is equal to the average delay of a T-car. The latter can be obtained using the following equation (Newell, 1965; Daganzo, 1997):

$$E[W_{B,1}] = \frac{q_S}{2T(q_S - q_A)}(T - G_T)^2, \tag{4}$$

where q_A denotes the T-vehicle inflow per lane. The q_A is calculated by:

$$q_A = \frac{Q_{T,1} + \lambda_b \delta}{N_T}, \tag{5}$$

where $Q_{T,1}$ denotes the (undersaturated) T-car inflow; and δ the passenger car equivalent (PCE) of a bus. Combining (4) and (5), we have the following equation that relates $E[W_{B,1}]$ to $Q_{T,1}$ for undersaturated traffic:

$$E[W_{B,1}] = \frac{q_S}{2T(q_S - (Q_{T,1} + \lambda_b \delta)N_T^{-1})}(T - G_T)^2. \tag{6}$$

Equation (6) shows that the expected bus delay varies with the T-car inflow. The T-car capacity under this alternative is $Q_{T,1}^{max} = \frac{q_S G_T N_T}{T} - \lambda_b \delta$, which can be derived from (5) by letting q_A equal the saturation flow, $q_S \cdot \frac{G_T}{T}$. When the T-car inflow reaches the above capacity, the expected bus delay attains its maximum, which is $E[W_{B,1}]^{max} = \frac{1}{2}(T - G_T)$.

3.2 Models under the pre-signal-only alternative

The expected bus delay under this alternative, denoted by $E[W_{B,2}]$, is again equal to the average delay of a T-car. The following equation is an approximation of the mathematical relation between $E[W_{B,2}]$ and the undersaturated T-car inflow, denoted by $Q_{T,2}$. The derivation is relegated to Appendix D.

$$E[W_{B,2}] = \begin{cases} \frac{q_A n_T T}{2 q_S N} + \frac{T}{2} + g_L - \left(G_T + R_L + G_L + t_y\right), \text{if } q_S N\left(G_T + R_L + G_L + t_y - g_L\right) < q_A n_T T; \\ \frac{q_S (T-g_T)^2}{2T(q_S - q_A)} + \frac{q_S N \left(g_T - G_T + \max\{0, g_L - (R_L + G_L + t_y)\}\right)^2}{2 q_A T (N - n_T)}, \\ \qquad \text{if } \left(\frac{q_A (T-g_T)}{q_S - q_A} - \left(g_T - G_T + \max\{0, g_L - \left(R_L + G_L + t_y\right)\}\right)\right) N \geq n_T \frac{q_A (T-g_T)}{q_S - q_A}; \\ \frac{q_S N \left(T - G_T + \max\{0, g_L - (R_L + G_L + t_y)\}\right)^2}{2T(q_S N - q_A n_T)}, \text{otherwise.} \end{cases} \tag{7}$$

where $q_A = \frac{Q_{T,2} + \lambda_b \delta}{n_T}$.

The T-car capacity under this alternative is $Q_{T,2}^{max} = \frac{q_S g_T n_T}{T} - \lambda_b \delta$. When the car inflow equals this capacity, the expected bus delay attains its maximum, which is:

$$E[W_{B,2}]^{max} = \begin{cases} \frac{g_T n_T}{2N} + \frac{T}{2} + g_L - \left(G_T + R_L + G_L + t_y\right), \text{if } N\left(G_T + R_L + G_L + t_y - g_L\right) < g_T n_T; \\ \frac{T - g_T}{2} + \frac{N \left(g_T - G_T + \max\{0, g_L - (R_L + G_L + t_y)\}\right)^2}{2 g_T (N - n_T)}, \text{ otherwise.} \end{cases} \tag{8}$$

3.3 Models under the BSP-only alternative

Other than the first two alternatives, the expected bus delay under this alternative, denoted by $E[W_{B,3}]$, is independent of the car inflow. Instead, it is a function of parameter t_m:

$$E[W_{B,3}] = \frac{1}{\lambda_b T}\left(\left(\frac{\lambda_b}{2}(T - G_T)^2 - (T - G_T) + \frac{1}{\lambda_b}\right) + e^{-\lambda_b t_m}\left(T - G_T - t_m - \frac{1}{\lambda_b}\right)\right) \tag{9}$$

The derivation of (9) is relegated to Appendix E.

The T-car capacity remains the same as if there is no BSP. This is because any extended green period will be fully utilized by discharging T-cars when a queue of these cars is always present upstream of the stop line. Note that capacity is defined as the maximum rate that cars can discharge into the intersection over the long run. The T-car capacity under the present alternative is thus:

$$Q_{T,3}^{max} = q_S N_T \frac{G_T}{T}. \tag{10}$$

3.4 Models under the integrated design

The expected bus delay, denoted by $E[W_{B,4}]$, has the same formula as (9):

$$E[W_{B,4}] = \frac{1}{\lambda_b T}\left(\left(\frac{\lambda_b}{2}(T - G_T)^2 - (T - G_T) + \frac{1}{\lambda_b}\right) + e^{-\lambda_b t_m}\left(T - G_T - t_m - \frac{1}{\lambda_b}\right)\right) \tag{11}$$

Note that it does not mean the integrated design produces the same expected bus delay as the BSP-only alternative, since the optimal G_T is different between the two alternatives.

The T-car capacity under this design, denoted by $Q_{T,4}^{max}$, is diminished by green extensions. Let t_L denote the time reduction in a g_T phase.[7] We have:

$$Q_{T,4}^{max} = q_S n_T \frac{g_T - E[t_L]}{T} \tag{12}$$

where $E[t_L]$ is given by:

[7] Note in our design that the g_T phase becomes saturated when the T-car inflow reaches its maximum rate, while the G_T phase may still be undersaturated. Thus, the capacity loss is proportional to the lost time in g_T due to BSP.

$$E[t_L] = \max\left\{0, (t_m - G_T)\frac{N-1}{n_T} + g_T\right\} - \frac{N-1}{\lambda_b n_T}\left(1 - e^{-\lambda_b \max\left\{0, t_m - G_T + \frac{g_T n_T}{N-1}\right\}}\right) \quad (13)$$

The derivation of (13) is also furnished in Appendix E. Since both the expected bus delay and the T-car capacity are functions of t_m, one can study the trade-off between the two metrics by varying t_m. This trade-off is examined next via numerical examples.

4. Numerical Case Studies

The above models are programmed in Matlab R2015a to conduct numerical experiments. Section 4.1 compares the integrated design against the other three alternatives. Section 4.2 explores how the benefits of the integrated design vary with key operating factors. Section 4.3 examines the minimum required length of sorting area.

4.1 Comparison between alternative designs

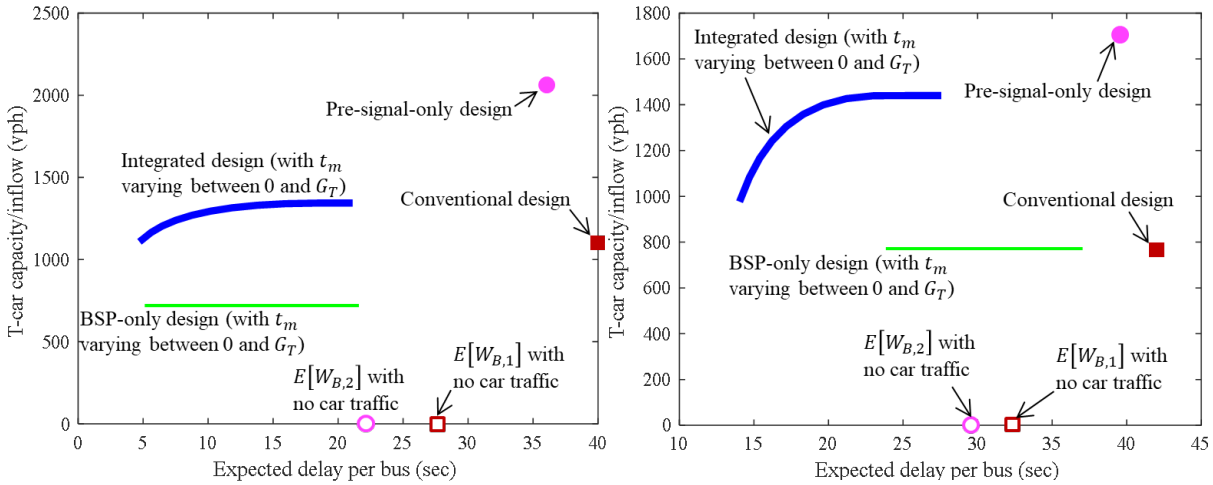
For the first batch of numerical instances, we assume: $N = 3$, $l = 0.2$, $T = 120$ s, $G = 60$ s, $R_L = 20$ s, $R_T = 40$ s, $q_S = 1800$ vehicle/h, and $\lambda_b = 30$ bus/h. The PCE of a bus, δ, is set to 3.5 (Ahuja et al., 2003). The T-car capacity under the integrated design, $Q_{T,4}^{max}$, is plotted against the expected bus delay, $E[W_{B,4}]$, for t_m varying in $[0, G_T]$ ($G_T = 48$ s in this batch of instances); see the thick blue curve in Figure 6a. This curve shows the trade-off between bus priority and car capacity: as t_m increases from 0 (the curve's right end) to 48 s (the left end), the expected bus delay decreases from 21 s to 5 s, and the T-car capacity diminishes from 1344 vph to 1122 vph. Note that the right half of this curve is nearly flat; i.e., when t_m increases from 0 to 22 s, the expected bus delay is reduced by 48%, while the T-car capacity diminishes by only 2.9%. This implies that the car capacity is insensitive to t_m when the latter takes small to medium values.

For comparison, Figure 6a also plots: i) $Q_{T,1}^{max}$ against $E[W_{B,1}]^{max}$ under the conventional design (this is a single point marked by the solid red square); ii) $Q_{T,2}^{max}$ against $E[W_{B,2}]^{max}$ under the pre-signal-only alternative (marked by the solid pink circle); and iii) $Q_{T,3}^{max}$ against $E[W_{B,3}]$ for $t_m \in [0, G_T]$ under the BSP-only alternative (this is again a curve, i.e., the thin green one). For a broader comparison, Figure 6a further plots $E[W_{B,1}]$ and $E[W_{B,2}]$ when the car inflow is equal to zero as the hollow red square and hollow pink circle, respectively. They illustrate the minimum expected bus delays that can be achieved under the two alternatives.

Comparison between the integrated design and the conventional alternative unveils that the former yields a 2-22% higher T-car discharge capacity and a 47-88% lower bus delay. Compared against the pre-signal-only alternative, the integrated design produces a 34-45% lower T-car discharge capacity, however it reduces the average bus delay by 41-86%. In addition, the integrated design produces lower bus delays than the above two alternatives even when the car inflow is near zero (see the hollow square and circle markers). This is owing to the incorporation of bus lane and BSP. While compared against the BSP-only alternative, the integrated design increases the T-car capacity by 56-87% while producing similar bus delays. These results manifest that the integrated design attains a nice balance between prioritizing bus travel and maintaining the intersection's capacity for serving cars.

The above findings hold in general as parameter values change. For example, Figure 6b shows another batch of instances where $N = 4$, $l = 0.4$, and $\lambda_b = 90$ bus/h (heavier bus traffic). All other parameters take the same values as in Figure 6a. The figure unveils that under a heavier bus traffic the integrated design produces greater car capacity gains. For example, as compared to the conventional design, the integrated design increases the T-car capacity by 30-88%, and meanwhile it still reduces the

average bus delay by 34-66%. Note that the integrated design curve (the thick blue one)'s right half is still flat, meaning that a small to medium level of BSP would significantly reduce bus delays while the car discharge capacity is only slightly compromised. On the other hand, that curve's left half becomes steeper as compared to Figure 6a. This is because a higher bus frequency renders more frequent green extensions, and thus the car discharge capacity diminishes faster as t_m grows. Finally, the expected bus delay under the integrated design is also significantly smaller than that under the BSP-only alternative. This is because the optimal G_T under the integrated design (39 s) is larger than that under the BSP-only alternative (26 s).



(a) $N = 3, l = 0.2, \lambda_b = 30$ bus/h **(b) $N = 4, l = 0.4, \lambda_b = 90$ bus/h**

Figure 6. T-car capacity and expected bus delay for the integrated design and the three alternatives

4.2 Parametric analysis

This section examines how the benefits of integrated design vary with key operating factors, including: bus frequency, λ_b; left-turning ratio, l; the overall green ratio for the approach, $\frac{G}{T}$; and cycle length, T. Specifically, we present the percentages of T-car capacity increase and bus delay saving of the integrated design, as compared against the conventional design.

Figure 7a plots the two percentages for a batch of instances with varying $\lambda_b \in [5,120]$ bus/h and $t_m \in [0, G_T]$ for $N = 3$. The other parameter values are the same as in Figure 6a. The solid contour lines in the figure represent the percentage T-car capacity increase, and the dashed ones represent the percentage bus delay saving. The horizontal axis expresses t_m as a percentage of G_T. The figure shows that even setting $t_m = 0$ can already reduce the expected bus delay by 50% (thanks to the DBL), while increasing t_m to G_T can reduce another 20-40% (thanks to the BSP). Meanwhile, T-car capacity gain up to 60% can often be achieved. Modest capacity loss (<10%) is observed only when t_m is very close to G_T and bus flow is high; see the top-right corner in the figure.

Further examination shows that the bus delay saving is sensitive to λ_b (specifically, it decreases as λ_b grows) only for large t_m's. This is because as λ_b grows, more buses are likely to arrive during a single signal cycle. While some buses may benefit from a green extension, other buses arriving later in the same cycle may suffer additional delays due to the resulting truncation of the following G_T phase. This undermines the bus delay savings, especially when t_m is large.

We also observe for large t_m's that the T-car capacity gain diminishes as λ_b increases. This is as expected. For small to medium t_m values, however, the above trend is reversed. This is explained as follows. First, when t_m is small, the car capacity under the integrated design only decreases moderately

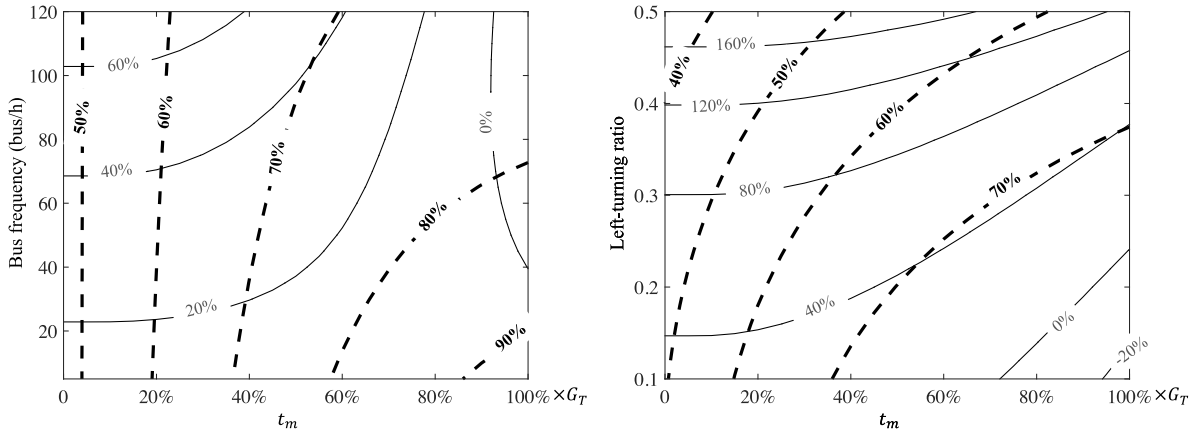
as λ_b increases. This is because the optimal G_T in this case has some redundancy, and thus the car capacity is not compromised if G_T is not truncated by much. On the other hand, under the conventional design, the car capacity decreases with λ_b at a faster rate since buses and cars share the lanes; see Section 3.1. Hence, the percentage gain in T-car capacity under the integrated design exhibits an increasing trend as λ_b grows. This implies that the integrated design is more beneficial under higher bus flows.

Figure 7b plots the same contour lines as in Figure 7a, but for $l \in [0.1,0.5]$ and $t_m \in [0, G_T]$. The λ_b is set to 90 bus/h, and all other parameter values are the same as in Figure 7a. The figure shows that as l increases, the bus delay saving diminishes while the car capacity gain increases. This can be explained as follows. First, when l increases, G_T would diminish. This in turn renders smaller t_m's (recall that $t_m \leq G_T$) and lower bus delay savings. On the other hand, under the conventional design, a larger l indicates that the T-traffic capacity (for both buses and cars) is smaller. Thus, the same bus flow λ_b would take a larger proportion of T-traffic. This means segregating buses from T-cars under the integrated design would bring a greater T-car capacity gain.

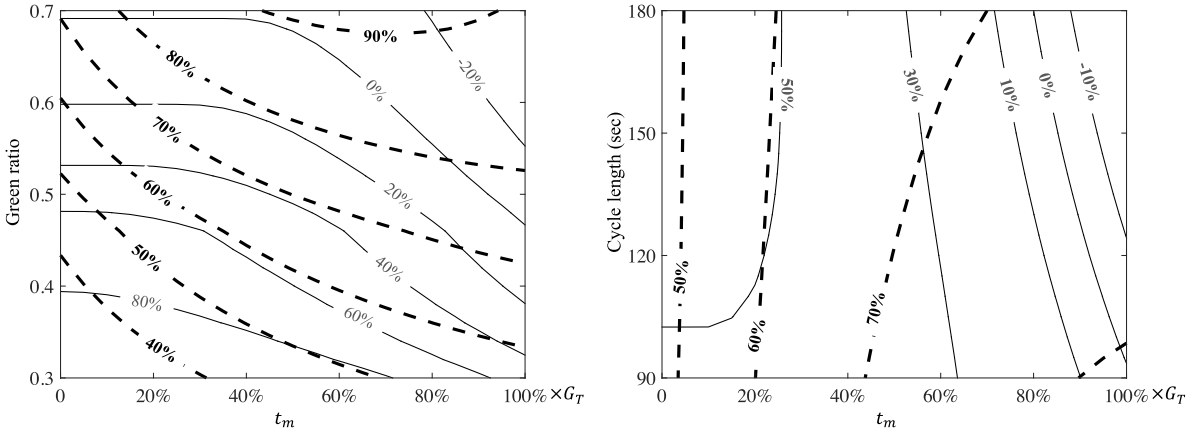
Figure 7c again plots the same contours, this time for $\frac{G}{T} \in [0.3,0.7]$ and $t_m \in [0, G_T]$. We set $\lambda_b = 90$ bus/h and the other parameter values are the same as in Figure 7a. The figure shows that when $\frac{G}{T}$ increases, the bus delay saving grows while the car capacity gain diminishes. This is mainly because a higher $\frac{G}{T}$ leads to a larger G_T, and in turn larger t_m's. The latter renders greater bus delay savings and lower car capacity gains due to more disruptions to the car discharge traffic. In addition, when $\frac{G}{T}$ continues to increase, the T-car capacity under the integrated design may reach the maximum that the pre-signal can offer. On the other hand, the T-car capacity under the conventional design will keep growing with $\frac{G}{T}$. This also leads to decreasing gains in T-car capacity as $\frac{G}{T}$ increases.

The last batch of contour lines are shown in Figure 7d for $T \in [90,180]$ s and $t_m \in [0, G_T]$. We again set λ_b to 90 bus/h and the other parameters to the same values as in Figure 7a. The figure shows that the car capacity gain and bus delay saving are not very insensitive to the cycle length T.

It is worth nothing in the above figures that the integrated design always reduces bus delays significantly, and that it produces greater car capacities than the conventional design for most of the cases. Similar results were also observed when N take different values. They are omitted here in the interest of brevity.



(a) Effect of λ_b **(b) Effect of l**



(c) Effect of G/T **(d) Effect of T**

Figure 7. Parametric analysis ($N = 3$)

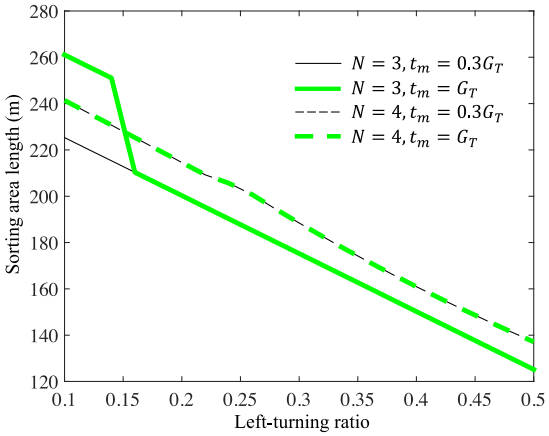
4.3 Minimum length of sorting area

In this section we examine how d_{min} varies with l, N, t_m, $\frac{G}{T}$, and T. Figure 8a plots d_{min} versus l for $N = 3$ and 4, and $t_m = 0.3G_T$ and G_T. The free-flow speed and backward wave speed are set as $w =$ 6.26 m/s and $v_f = 15.64$ m/s, respectively (Skabardonis and Geroliminis, 2005). All other parameter values are the same as in Figure 6a. The figure shows that d_{min} decreases significantly as l grows from 0.1 to 0.5. This is because a left-turning ratio close to 0.5 renders the T-queue and L-queue formed in the sorting area having similar lengths. Thus, neither queue is too long. In addition, a 4-lane approach (dashed curves in the figure) generally requires a slightly longer sorting area than a 3-lane one (solid curves). Finally, d_{min} is independent on t_m except for the 3-lane case with a small l; see the thick solid curve in Figure 8a.

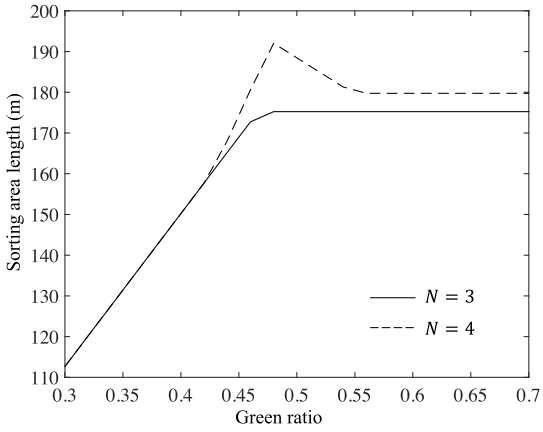
Figure 8b plots d_{min} versus $\frac{G}{T}$ for $N = 3$ and 4. The t_m can take any value since we find it has no effect on d_{min}. The l is now set to 0.3 and all other parameter values are the same as in Figure 8a. The figure shows that d_{min} increases with $\frac{G}{T}$ for $\frac{G}{T} \leq 0.48$. This is intuitive since a larger G renders higher car discharge capacities and thus longer queue lengths in the sorting area. However, d_{min} stops growing when the green ratio exceeds 0.48, since the pre-signal's maximum discharge capacity is attained at $\frac{G}{T} = 0.48$. For the 4-lane case (the dashed curve), d_{min} even diminishes modestly as $\frac{G}{T}$ grows beyond 0.48. This is because when $\frac{G}{T} = 0.48$, a small T-queue is formed on top of a L-queue (see the case illustrated in Figure 4b); and when $\frac{G}{T} > 0.48$, this small T-queue gradually vanishes (see the case shown in Figure 4a).

Figure 8c further illustrates how d_{min} varies with T for $N = 3$ and 4. The $\frac{G}{T}$ is again set to 0.5, and all other parameter values are the same as in Figure 8b. As expected, the figure shows that d_{min} increases as T grows.

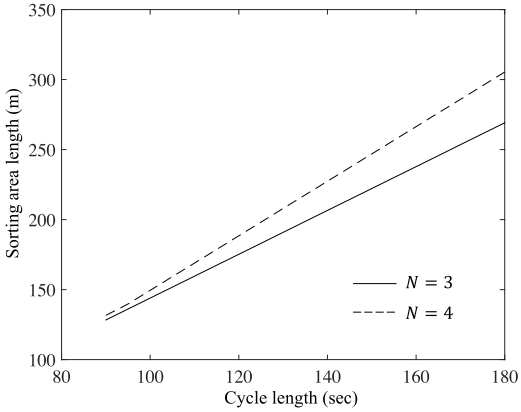
Note in Figures 8a-c that d_{min} can sometimes be large, e.g., greater than 200 m. Such a long sorting area may not fit in short blocks. However, the integrated design can still be implemented even if the calculated d_{min} is longer than what a real intersection approach can offer. Under this situation, the sorting area length will be determined by practical constraints, and the pre-signal phase durations (g_L or g_T, or both) must be trimmed to ensure the reduced queues can be fully contained in the sorting area. The trimmed g_L or g_T can be calculated by reversing (3).



(a) For $l \in [0.1, 0.5]$



(b) For $\frac{G}{T} \in [0.3, 0.7]$



(c) For $T \in [90, 180]$ s

Figure 8. Minimum length of sorting area

Also note that due to heterogeneous driving behavior and vehicle sizes, the actual queue lengths in the sorting area may be greater than d_{min} predicted by (3). However, this effect can be accounted for by using conservative values for queue formation and backward shockwave speeds, i.e., larger w_L and w_T, and smaller w values; see, e.g., Li et al. (2017) and Mei et al. (2019) for the estimation of these shockwave speeds.

5. Incorporating Real-World Operating Features

Results presented in Section 4 were derived under idealized assumptions. Specifically, assumption iii) in Section 2.1 specifies that bus arrival times can be predicted accurately; assumption iv) ignores the impacts of BSP on traffic in the opposite direction; and assumption v) implies that queued cars discharge at a uniform rate during a green period of either signal. We show in this section that by relaxing these assumptions and incorporating some real-world features, the aforesaid benefits of the integrated design will only diminish moderately. Section 5.1 discusses the effects of incorporating the opposite approach into our models. Section 5.2 studies the effects of inaccurate bus arrival time predictions. Section 5.3 examines the effects of random car discharge headways.

5.1 Joint design of two opposing approaches of the intersection

With some modifications, the models presented in Section 3 can be applied to study the joint design of two opposing approaches, i.e., the subject approach and its opposite one. However, considering both approaches jointly would add a number of new parameters. For example, the opposite approach may

have a different number of lanes (with or without a DBL) and different traffic characteristics (e.g., left-turning ratio and bus frequency). This will render a full parametric analysis much more complicated and lengthier. Thus, this section only analyzes a simple case for illustration purpose, where the two approaches have the same lane layout and traffic characteristics. We further assume that the opposite approach also includes a DBL, which is common since bus operations are bi-directional. The same t_m for green extension is also assumed for both approaches. We consider the following three scenarios that may occur depending on the traffic load in the opposite approach:

i) The integrated design is applied to both approaches. This may occur where traffic loads in both approaches are heavy. Under this scenario, models presented in Section 3.4 can be directly applied to estimate the performance metrics of both approaches. The only change is that now the λ_b in these models should be the sum of bus flows in both directions, because a green extension serves buses in both directions.
ii) The integrated design is still applied to the subject approach, but the pre-signal is deactivated in the opposite approach due to the low traffic load. This may occur in a morning peak period when the opposite approach serves the low outbound traffic from city center to suburban areas. This results in a BSP-only design in the opposite approach. Under this scenario, models presented in Section 3.3 can be applied to estimate the performance metrics of the opposite approach, except that λ_b should be the sum of bus flows in both directions.
iii) Not only the pre-signal, but the BSP is also deactivated for the opposite approach (i.e., only buses traveling in the subject approach can request BSP), possibly due to a low transit patron load. This will reduce the number of BSP requests and thus benefit the car traffic. Under this scenario, the performance metrics for the subject approach should be estimated using models in Section 3.4 where λ_b represents the bus flow of the subjective approach only. The expected bus delay in the opposite approach can be estimated as if there is no BSP. Note that under our green extension scheme, a red period duration for T-traffic is always $T - G_T$. Thus, the expected bus delay in the opposite approach is $\frac{1}{2T}(T - G_T)^2$. Finally, the T-car capacity for the opposite approach is the same as in scenario ii).

In the following numerical example, we set $l = 0.4$, $\lambda_b = 90$ bus/h and $N = 4$ for either direction, and the other parameter values are the same as for Figure 6b. The main signal timing plan is optimized under the integrated design using models presented in Appendix C. Figure 9 plots the T-car capacity against expected bus delay for either approach under the three scenarios.

First note in scenario i), due to the symmetry between the two approaches, their pre-signal timing plans and performance metrics are also the same. The metrics are plotted as the thin black curve in Figure 9. The single-approach integrated design metrics (i.e., the thick blue curve in Figure 6b) are copied here for comparison. The comparison shows that jointly considering two opposing approaches renders a lower car capacity and a greater expected bus delay for larger t_m values (see the left parts of the curves). However, significant car capacity gains and bus delay savings are still observed as compared to the conventional design (the red square).

In scenario ii), the metrics of the subject approach are represented by the same black curve, and those of the opposite approach are represented by the thin cyan line. The figure shows that buses in the opposite approach experience the same expected delay as those in the subject approach, and the T-car capacity of the opposite approach is moderately lower than under the conventional design. However, since in this scenario the car inflow in the opposite approach is low, the lower capacity may not be a problem. On the other hand, the present scenario can produce lower car delays than scenario i) in the opposite approach, thanks to the deactivation of the pre-signal.

In scenario iii), the metrics of the opposite approach are represented by the orange circle, showing that the bus delay is relatively large in the absence of BSP. On the other hand, the metrics of the subject approach are represented by the thick blue curve, meaning that cars and buses in the subject approach would enjoy greater benefits. Hence, this scenario triumphs over scenario ii) when the subject approach is very congested, and the opposite approach only serves a light traffic.

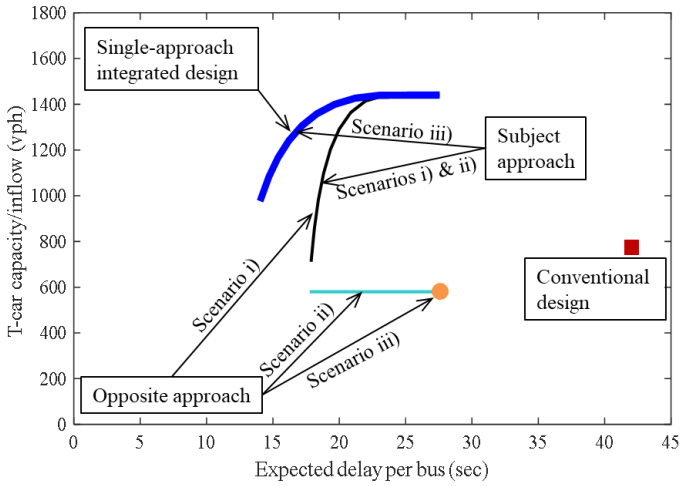


Figure 9. T-car capacity and expected bus delay for two opposing approaches

The above method can be applied to analyze two opposing approaches with distinct lane layouts and traffic characteristics. Only modest modifications are needed. For example, the models presented in Appendix C need be modified to optimize the lane assignments and signal timings for both approaches jointly.

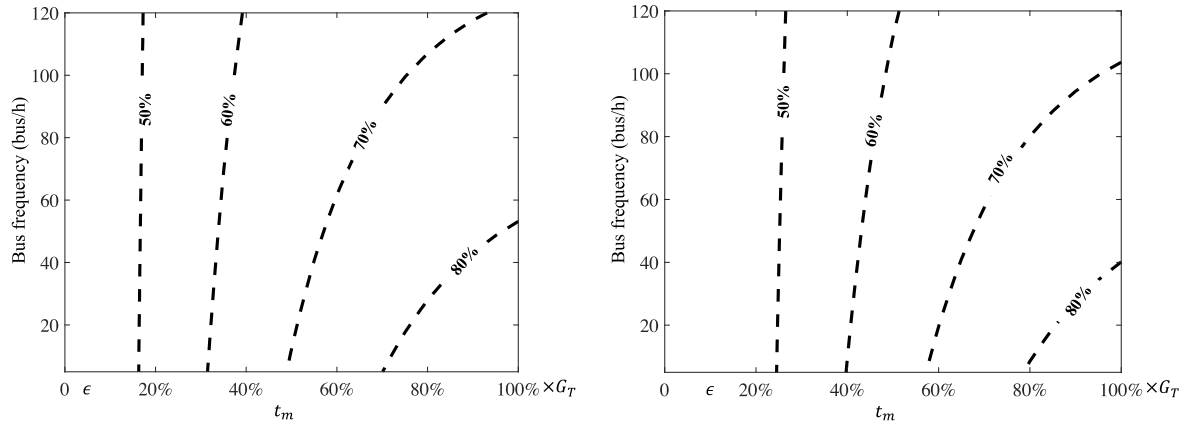
5.2 Effects of inaccurate bus arrival time predictions

Previous studies show that, using suitably placed detectors or real-time on-board GPS data, the error of bus arrival time prediction for BSP can be within a few seconds (e.g., Tan et al., 2008; Shi, 2016), given that no near-side stop is present in the approach. We denote ϵ as the maximum error between the predicted and actual bus arrival times; i.e., if a predicted bus arrival time is t_{ai}, then the actual arrival time would be in the range $[t_{ai} - \epsilon, t_{ai} + \epsilon]$. To prioritize this bus, the extended green period must contain the above range. Hence, the expected bus delay equation (11) under the integrated design should be replaced by the following equation:

$$E[W_{B,4}] = \frac{1}{\lambda_b T}\left(\left(\frac{\lambda_b}{2}(T - G_T + \epsilon)^2 - (T - G_T + \epsilon) + \frac{1}{\lambda_b}\right) + e^{-\lambda_b(t_m - \epsilon)}\left(T - G_T + 2\epsilon - t_m - \frac{1}{\lambda_b}\right)\right) \quad (14)$$

Derivation of (14) is relegated to Appendix E.1. The T-car capacity equations (12) and (13) under the integrated design still hold, for reasons that are also explained in Appendix E.1.

With the modified model (14), we re-calculate the percentage bus delay savings (as compared against the conventional design) for $\lambda_b \in [5,120]$ bus/h, $t_m \in [\epsilon, G_T]$, and $\epsilon \in \{3,5\}$ s. They are illustrated by contour lines in Figures 10a and b. Here we assume $N = 3$ and the other parameter values are the same as in Figure 7a. Comparing both figures against the dashed contour lines in Figure 7a, we find that the bus delay savings are only modestly smaller when prediction error is taken into account. Thus, the benefit of the integrated design is still significant as long as ϵ is kept reasonably small.



(a) $\epsilon = 3$ s **(b) $\epsilon = 5$ s**

Figure 10. Effects of bus arrival time prediction error

5.3 Effects of random car discharge headways

The T-car capacity predicted by (12) is valid only if the L- and T-car queues formed in the sorting area are cleared respectively by the end of each G_L and G_T phases. However, real-world car discharge headways under saturated traffic are random. Those queues in the sorting area may take longer than predicted by our idealized models to discharge into the intersection. Unfortunately, if one T-car is left in the sorting area at the end of a G_T phase, it will block a lane. That lane would otherwise be used by upstream L-cars to discharge in a following G_L phase, and in turn by further upstream T-cars to discharge in a yet following G_T phase. Thus, the car discharge capacity would diminish. To avoid these lane blockages in the sorting area, the number of cars released into the sorting area per cycle must be further reduced.

To account for the above effect, (12) is modified to:

$$Q_{T,4}^{max} = \min\left\{\frac{q_S n_T g_T}{T}, \frac{\alpha^2\gamma^2+2q_S G_T-\alpha\gamma\sqrt{\alpha^2\gamma^2+4q_S G_T}}{2T}(N-1)\Phi(\alpha)\right\}\left(1-\frac{E[t_L]}{g_T}\right), \quad (15)$$

where γ denotes the coefficient of variation of car discharge headways; α a prespecified constant; and $\Phi(\cdot)$ the cumulative distribution function of the standard normal distribution. The $\Phi(\alpha)$ represents the probability that the lane is not blocked. Thus, the value of α should be selected to ensure $\Phi(\alpha)$ is close to 1. (Note here we only consider lane blockages due to residual T-queues, but the modeling of lane blockages due to residual L-queues would be similar.) Derivation of (15) is relegated to Appendix F.

The effects of the random discharge headways on T-car capacity are illustrated in Figure 11 for $N = 3$. The figure shows contour lines of T-car capacity gains as compared against the conventional design for $T \in [90,180]$ s, $t_m \in [0, G_T]$, and $\lambda_b = 90$ bus/h. We set $\gamma = 0.25$ (Jin et al., 2009) and $\alpha = 3$ to ensure a lane is not blocked with a probability of 99.86%. All other parameter values are the same as in Figure 7d. Comparison between Figure 11 and the solid contour lines in Figure 7d shows that the T-car capacity gains are moderately smaller after considering random discharge headways. However, capacity gains are still observed for most instances. This again manifests the benefit of the integrated design in the more realistic settings.

6. Conclusions, Deployment Issues, and Future Work

We examine a novel intersection approach design that integrates a mid-block pre-signal, a DBL, and a simple BSP scheme. The pre-signal is used to sort left-turning and through-moving car traffic streams tandemly so that the green periods of the main signal are more efficiently utilized. Key design features

and parameters are derived, including the coordinated pre-signal and main signal timing plans, adaptive phase alterations for BSP, and the sorting area length. These can be used by practitioners for sketching the design in a real intersection approach. Models for estimating expected bus delay and through-moving car capacity are developed. These models facilitate the assessment of the integrated design via large batches of numerical experiments under various operating environments.

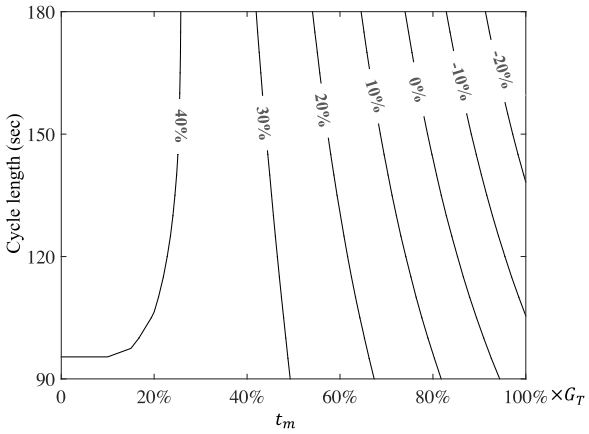


Figure 11. Effects of random discharge headways ($N = 3$)

Our numerical experiments unveil insights that have practical implications. Above all, we show for various operating conditions that the incorporation of pre-signal can compensate the car capacity loss due to bus priority measures, and that car capacity gains and bus delay savings can be achieved simultaneously. We further show that the above benefits of the integrated design are fairly robust when real-world operating features are considered. This means bus priority measures can be implemented at very busy intersections without damaging the car traffic. Moreover, the integrated design seems to be especially competitive for intersections with heavy bus traffic (see again Figure 7a), where frequent activation of BSP may significantly disrupt car traffic (Lin et al., 2015).

We conclude this paper by briefly discussing: i) the choice of a key parameter, t_m; ii) deployment issues; and iii) potential modifications and extensions of the present study for future work.

6.1 Choice of t_m

In general, a larger t_m renders higher bus priority, but also greater damage to the car traffic, especially when the bus frequency is high. For example, Figure 6b unveils the significant reduction in car capacity as t_m approaches its maximum. A long green extension may also result in residual car queues that form in the following cycle where signal phases are delayed or shortened. On the other hand, implementing a limited green extension (under a smaller t_m) can often yield considerable bus delay savings with nearly no negative impact on car capacity; see again the right, flat part of the blue, bolded curves in Figures 6a and b. This implies that a relatively small t_m may suffice to prioritize buses without incurring damages to cars. In addition, other practical matters, e.g., the minimum green time to ensure safe pedestrian street-crossing, should also be taken into consideration when determining the value of t_m.

6.2 Deployment issues

Real-world implementation of the integrated design requires coordinated adaptive signal control that receives real-time inputs from onboard GPS devices or roadside sensors, and that can coordinate the pre-signal and main signal in real time. Intelligent signal control systems like SCATS (Wolshon and

Taylor, 1999) can be used for this purpose. Variable message signs would be a good assistance for instructing car drivers to conform to the dynamic lane assignment rule of the pre-signal (Luo, 2011).

6.3 Future work

The main purpose of this paper is to examine the viability and potential benefits of the integrated intersection design. Necessary simplifications were made to this end. These simplifications can be further relaxed via modifications and extensions of our work. Some of those potential modifications and extensions are summarized below:

i) The present models ignore the bounded acceleration of vehicles. In reality, queued vehicles need first to accelerate from 0 to the free flow speed when discharging. Thus, they may take longer time to traverse the sorting area from the pre-signal. This effect can be simply accounted for by increasing the offset between the ends of pre-signal and main signal green phases (d/v_f) to match the actual travel time through the sorting area.

ii) This paper only considers through-moving buses. In practice, left-turning buses can use the same DBL as through-moving ones, and they can receive their own signal priority via, e.g., extension of the G_L phase. However, L- and T-buses queueing in the same bus lane may impede each other from discharging into the intersection. Thus, special signal timing plan is needed to sort out the conflicts between L- and T-buses where only one DBL is available.

iii) Our work assumes green extension as the only BSP scheme. Other BSP schemes, e.g., red truncation and green insertion, can be designed and modeled in a similar way. Incorporating more BSP schemes can further reduce bus delays.

iv) A second pre-signal can be installed on the bus lane to convert it into an IBL (e.g., Guler and Menendez, 2014). This requires more sophisticated signal control schemes, but has a potential to further increase the car capacity at the intersection.

v) Finally, researchers and practitioners might also be interested in knowing car delays at an intersection under given car and bus demands, so as to, e.g., minimize the overall passenger delays (He et al., 2016). The method built upon cumulative vehicle count diagrams (see Appendix D for instance) can be used to develop car delay models under the integrated design. Modeling complexity may arise due to residual car queues that may occur at the end of pre-signal green phases, since a residual queue may last for several consecutive cycles. These residual queues can be modeled via Markov chains (see, e.g., Viti and Van Zuylen, 2010).

Acknowledgements

This work was supported by a General Research Fund (Project No. 15224317) provided by the Research Grants Council of Hong Kong and a Platform Project Fund (Project No. ITP/020/18LP) provided by the Innovation and Technology Commission of Hong Kong.

Appendix A. List of Notations

Table A1. List of notations

Notation	Definition
T	Signal cycle length.
G	Total green duration for the subject approach.
G_L, G_T	Green periods for L- and T-traffic in the subject approach, respectively.
R_L, R_T	Red periods of the main signal.
g_L, g_T	The pre-signal's green periods for L- and T-traffic, respectively.

N	Number of lanes in the subject approach.
M	Number of lanes in the sorting area.
N_L, N_T	Numbers of L- and T-car lanes in the subject approach, respectively, in the absence of pre-signal.
n_L, n_T	Numbers of L- and T-car lanes at the pre-signal, respectively.
d	Length of the sorting area.
d_{min}	Minimum sorting area length.
l	Left-turning ratio.
t_y	Amber period.
t_m	Maximum extension of a G_T phase.
t_L	The green time loss per cycle for T-traffic due to a green extension under the integrated design.
t_e	The green time extension following a G_T phase.
t_R	The redundant green time in a G_T phase.
$\tilde{G}_T$	Duration of a truncated G_T phase due to green extension.
$\tilde{g}_T$	Duration of a truncated g_T phase due to green extension.
q_S	Capacity of a car lane.
v_f	Vehicles' free-flow speed.
w	Backward wave speed.
q_L, q_T	Saturation flows of the L- and T-traffic at the pre-signal, respectively.
w_L, w_T	The queue formation shock wave speeds in the sorting area for L- and T-vehicles, respectively.
$E[W_{B,i}]$	Expected bus delay under each alternative ($i = 1,2,3,4$).
$E[W_{B,i}]^{max}$	Maximum expected bus delay under the conventional and pre-signal-only alternatives ($i = 1,2$).
$Q_{T,i}$	T-car inflow under the conventional and pre-signal-only alternatives ($i = 1,2$).
$Q_{T,i}^{max}$	Capacity of T-cars of the subject approach under each alternative ($i = 1,2,3,4$).
q_A	Vehicle inflow per T-lane under the conventional and pre-signal-only alternatives.
λ_b	Bus arrival rate.
δ	Number of passenger cars equivalent to a bus.
ϵ	Prediction error of bus arrival times.
γ	Coefficient of variation in cars' saturation headways.

Appendix B. Derivation of the minimum sorting area length in (3)

The minimum sorting area length is equal to the maximum queue length in the sorting area. Its exact solution has a quite complicated mathematical form. Thus, we instead develop a good upper bound which is much simpler.

We first consider a signal cycle without BSP; see Figures 4a-c. For the case of separated queues shown in Figure 4a, the L-queue length never exceeds $g_L\left(w_L^{-1} + v_f^{-1}\right)^{-1}$. This upper bound is attained if all the L-vehicles in a g_L phase join the queue. Similarly, the T-queue length never exceeds $g_T\left(w_T^{-1} + v_f^{-1}\right)^{-1}$. Thus, an upper bound of queue lengths under this case is $\max\left\{g_L\left(w_L^{-1} + v_f^{-1}\right)^{-1}, g_T\left(w_T^{-1} + v_f^{-1}\right)^{-1}\right\}$.

Now we look at the case of tandem queues shown in Figure 4b. First of all, the T- and the (separated) L-queue lengths are again bounded by $g_T\left(w_T^{-1} + v_f^{-1}\right)^{-1}$ and $g_L\left(w_L^{-1} + v_f^{-1}\right)^{-1}$, respectively. For the small L-queue formed on top of the T-queue, its length is equal to $\max\{t_2 - t_y, 0\} \cdot (w_L^{-1} - w^{-1})^{-1}$; see the definition of t_2 in the figure. To find t_2, note first that $t_1 = g_T + g_L - \left(G_T + R_L + t_y + G_L\right)$; see the definition of t_1 in the figure. Then, by geometry we have $t_2 = \max\left\{t_1 - g_T \cdot \left(w_T^{-1} + v_f^{-1}\right)^{-1} \cdot (w_T^{-1} - w^{-1}), 0\right\} = \max\left\{\frac{g_T n_T}{M} + g_L - \left(G_T + R_L + t_y + G_L\right), 0\right\}$.

Thus, the L-queue length is $\max\left\{\frac{g_T n_T}{M} + g_L - \left(G_T + R_L + G_L + 2t_y\right), 0\right\} \cdot (w_L^{-1} - w^{-1})^{-1}$. The tandem queue length's upper bound is thus the sum of bounds for the T-queue and the small L-queue, i.e., $g_T\left(w_T^{-1} + v_f^{-1}\right)^{-1} + \max\left\{\frac{g_T n_T}{M} + g_L - \left(G_T + R_L + G_L + 2t_y\right), 0\right\} \cdot (w_L^{-1} - w^{-1})^{-1}$. The queue lengths in a cycle is therefore bounded by $\max\left\{g_L\left(w_L^{-1} + v_f^{-1}\right)^{-1}, g_T\left(w_T^{-1} + v_f^{-1}\right)^{-1} + \max\left\{\frac{g_T n_T}{M} + g_L - \left(G_T + R_L + G_L + 2t_y\right), 0\right\} \cdot (w_L^{-1} - w^{-1})^{-1}\right\}$.

By symmetry, the upper bound of queue lengths for the last case shown in Figure 4c is $\max\left\{g_T\left(w_T^{-1} + v_f^{-1}\right)^{-1}, g_L\left(w_L^{-1} + v_f^{-1}\right)^{-1} + \max\left\{\frac{g_L n_L}{M} + g_T - \left(G_L + R_T + G_T + 2t_y\right), 0\right\} \cdot (w_T^{-1} - w^{-1})^{-1}\right\}$.

Now we examine the cases where a green extension is granted.

For the case of separated queues (see Figure 5a), the queue lengths are still bounded by the same upper bound derived above. Although the L-queue following a green extension could be longer than that in a regular cycle, this queue length is still no greater than $g_L\left(w_L^{-1} + v_f^{-1}\right)^{-1}$.

For the case of tandem queues where $g_L > R_L + G_L + t_y$ (Figure 5b), we have the following facts: i) the length of the first pair of tandem queues in Figure 5b is not affected by the green extension; ii) the length of the following L-queue is no greater than $g_L\left(w_L^{-1} + v_f^{-1}\right)^{-1}$; and iii) the second tandem queue in the figure is smaller than a tandem queue formed in a regular cycle. Thus, the queue lengths are still bounded by the same upper bound derived above.

For the last case (Figure 5c), however, the tandem queues following the green extension may extend further upstream than in a regular cycle. This is because the small T-queue formed on top of the L-queue may become longer. Its length is now bounded by $\max\left\{t_m + \frac{g_L n_L}{M} + g_T - \left(G_L + R_T + G_T + 2t_y\right), 0\right\} \cdot (w_T^{-1} - w^{-1})^{-1}$. Note that t_m is the maximum time by which the phases following an extended G_T are postponed. Hence, the queue length upper bound under this case is updated to: $\max\left\{g_T\left(w_T^{-1} + v_f^{-1}\right)^{-1}, g_L\left(w_L^{-1} + v_f^{-1}\right)^{-1} + \max\left\{t_m + \frac{g_L n_L}{M} + g_T - \left(G_L + R_T + G_T + 2t_y\right), 0\right\} \cdot (w_T^{-1} - w^{-1})^{-1}\right\}$.

The above results are summarized as equation (3) in Section 2.5.

Appendix C. Optimization models for signal phases and lane assignment

For simplicity, we choose to maximize the overall car discharge capacity from the intersection approach without considering the vehicular traffic in other approaches, pedestrian traffic, and impacts of neighboring intersections.

We first present the following integer program for optimizing lane assignment (i.e., N_L and N_T) under the conventional design:

$$Q_M(N) = \max\left\{Q = \frac{q_S G}{T\left(l \cdot N_L^{-1} + (1-l) N_T^{-1}\right)}\right\} \quad \text{(C1a)}$$

$$\text{subject to: } N_L + N_T = N \quad \text{(C1b)}$$

$$N_L, N_T \geq 1 \text{ and are integers;} \quad \text{(C1c)}$$

where the objective $Q_M(N)$ denotes the maximum overall vehicular capacity for an N-lane approach. In addition, the optimal G_L and G_T are given by:

$$G_L = \frac{Tl}{q_S N_L} Q_M(N), \quad \text{(C2a)}$$

$$G_T = \frac{T(1-l)}{q_S N_T} Q_M(N). \quad \text{(C2b)}$$

Model (C1a-C2b) can also be applied to find G_L, G_T, N_L, and N_T under the BSP-only alternative, except that N will be replaced by $N-1$.

The following integer program optimizes the lane assignment under the pre-signal-only alternative:

$$Q_P(N) = \max\left\{Q = q_S \cdot \min\left\{\frac{NG}{T}, \frac{1-2t_y T^{-1}}{l \cdot n_L^{-1} + (1-l) n_T^{-1}}\right\}\right\} \quad \text{(C3a)}$$

$$\text{subject to: } n_L + n_T = N \quad \text{(C3b)}$$

$$n_L, n_T \geq 1 \text{ and are integers;} \quad \text{(C3c)}$$

where $Q_P(N)$ denotes the maximum overall vehicular capacity for an N-lane approach with a pre-signal. The optimal signal phases are:

$$g_T = \frac{T(1-l)}{q_S n_T} Q_P(N) \quad \text{(C4a)}$$

$$g_L = \frac{Tl}{q_S n_L} Q_P(N), \quad \text{(C4b)}$$

$$G_L = \frac{g_L n_L}{N}, \quad \text{(C4c)}$$

$$G_T = G - G_L. \quad \text{(C4d)}$$

Equation (C4d) means any redundant green time in G will be allocated to the G_T phase, while G_L is only enough to discharge L-vehicles that enter the sorting area.

Similarly, model (C3a-C4d) can also be applied to find G_L, G_T, g_L, g_T, n_L, and n_T for the integrated design, except that N is replaced by $N-1$.

Since N is usually no more than 5, the above programs can be solved by exhaustive search. The details of the search algorithm are omitted for simplicity.

Appendix D. Derivation of (7)

The expected T-vehicle delay under the pre-signal-only alternative is derived via diagrams of cumulative vehicle counts measured at the intersection's stop line. There are totally three cases, each described by a different cumulative count diagram as shown in Figures D1a-c. In each figure, three cumulative count curves are plotted. The dashed line with slope $q_A n_T$ represents the count that would occur in the absence of any signal, i.e., the "virtual arrival" curve (Daganzo, 1997). Here q_A still denotes the T-vehicle inflow per lane, and is given by $q_A = \frac{Q_{T,2} + \lambda_b \delta}{n_T}$. The dotted curve represents the count that would occur if only the pre-signal is present. The solid curve represents the actual count at the stop line, i.e., the departure curve. The dotted curve can never exceed the virtual arrival curve, and the departure curve can never exceed the dotted one. We set time 0 as the end of the last g_T phase. The shaded area

between the virtual arrival and departure curves represents the total vehicular delay per cycle.[8] The three cases differ in the relative positions of the g_T and G_T phases, which entail different departure curves shown in the diagrams.

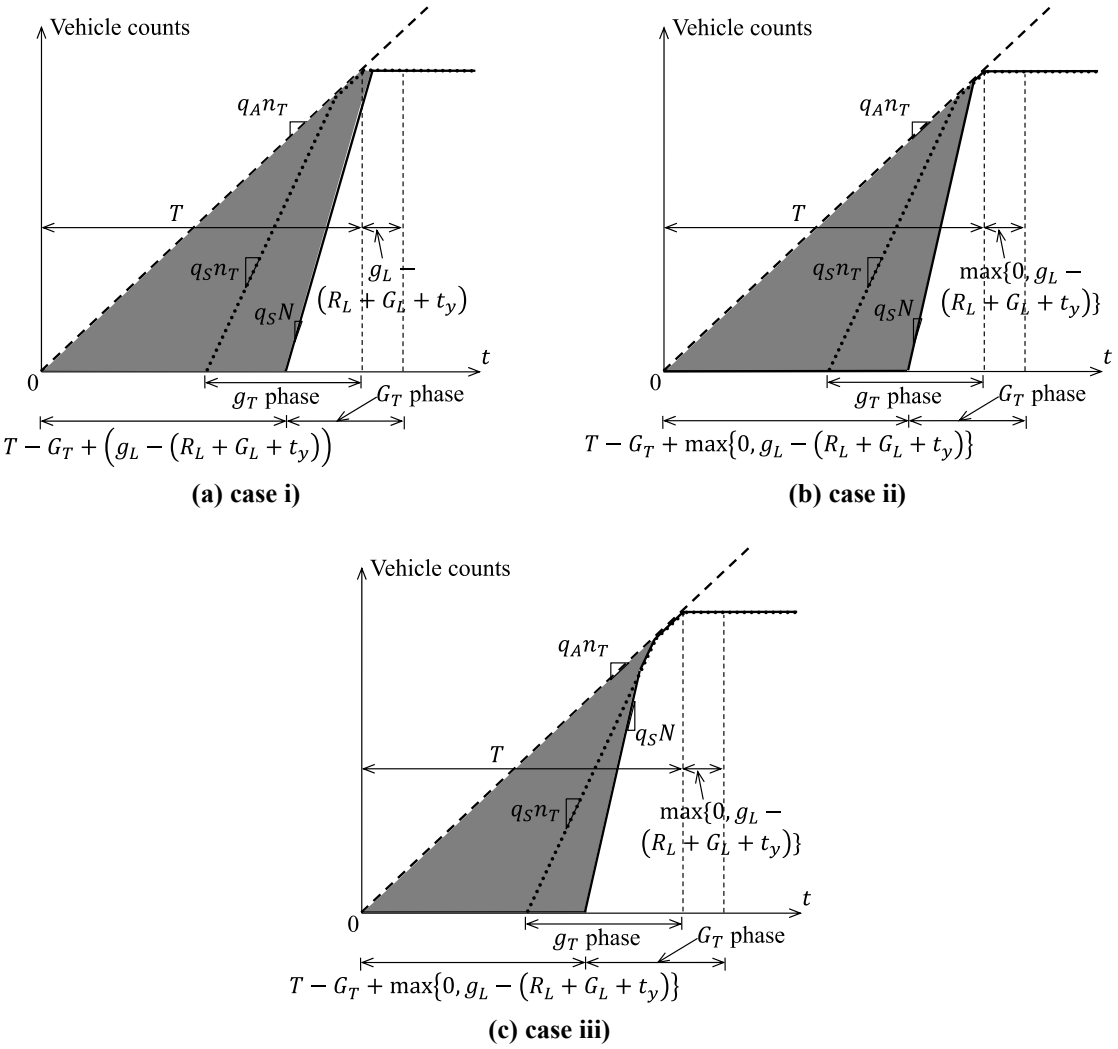


Figure D1. Diagrams of cumulative vehicle counts under the pre-signal-only alternative

Figure D1a shows the case where all T-vehicles have to join the queue in the sorting area before discharging into the intersection. By geometry, the case is found to arise where $q_S N\left(G_T - \left(g_L - \left(R_L + G_L + t_y\right)\right)\right) < q_A n_T T$. The condition also implies that $g_L - \left(R_L + G_L + t_y\right) > 0$, since otherwise we would have $q_S N G_T < q_A n_T T$, meaning that the main signal would be oversaturated. This indicates that the present case belongs to the situation described by Figure 4b; i.e., the first a few L-vehicles will form a small queue on top of the T-queue. Thus, there is an offset of $g_L - \left(R_L + G_L + t_y\right)$ between the ends of g_T and G_T phases in a moving-time coordinate (Newell, 1993), meaning that it would be measured by an observer that travels at the free-flow speed of vehicles. This offset is marked in Figure D1a. Given the above information and calculated by geometry, the average delay per vehicle in this case is $\frac{q_A n_T T}{2 q_S N} + \frac{T}{2} - G_T + \left(g_L - \left(R_L + G_L + t_y\right)\right)$.

[8] The area between the dashed and dotted curves represents the vehicular delays incurred upstream of the pre-signal, and the area between the dotted and solid curves represents the delays incurred in the sorting area.

Figure D1b shows the case where: i) the last portion of T-vehicles in the cycle pass both the pre-signal and the intersection without delay; and ii) some vehicles that pass the pre-signal without delay will still join the queue in the sorting area before discharging into the intersection. By geometry, we find that the above condition i) occurs where $q_S N\left(G_T - g_L + R_L + G_L + t_y\right) \geq q_A n_T T$, and condition ii) is equivalent to $\left(\frac{q_A(T-g_T)}{q_S-q_A} - \left(g_T - G_T + \max\{0, g_L - \left(R_L + G_L + t_y\right)\}\right)\right) N < n_T \frac{q_A(T-g_T)}{q_S-q_A}$. The offset between the ends of g_T and G_T in a moving-time coordinate is $\max\{0, g_L - \left(R_L + G_L + t_y\right)\}$, as shown in Figure D1b. Geometric calculation reveals that the average delay per vehicle in this case is $\frac{q_S N\left(T-G_T+\max\{0,g_L-\left(R_L+G_L+t_y\right)\}\right)^2}{2T(q_S N-q_A n_T)}$.

The last case is illustrated in Figure D1c. In this case, the last portion of T-vehicles traverse the sorting area without delay (but some of those vehicles may experience delays before discharging through the pre-signal). Geometric calculation shows that this case occurs where $\left(\frac{q_A(T-g_T)}{q_S-q_A} - \left(g_T - G_T + \max\{0, g_L - \left(R_L + G_L + t_y\right)\}\right)\right) N \geq n_T \frac{q_A(T-g_T)}{q_S-q_A}$. The average delay per vehicle in this case is found to be $\frac{q_S(T-g_T)^2}{2T(q_S-q_A)} + \frac{q_S N\left(g_T-G_T+\max\{0,g_L-\left(R_L+G_L+t_y\right)\}\right)^2}{2q_A T(N-n_T)}$.

Combining the above results, we have equation (7).

Appendix E. Derivation of (9), (11), and (13)

To derive $E\left[W_{B,3}\right]$, we consider an arbitrary cycle. Without loss of generality, we set $t = 0$ at the end of a regular G_T phase.

We first find the distribution of $t_e \in [0, t_m]$. Since bus arrivals follow a Poisson process with rate λ_b, the probability of $t_e = 0$ (i.e., no bus arrives between $t = 0$ and t_m) is $e^{-\lambda_b t_m}$. The probability that $k > 0$ buses arrive between 0 and t_m is $e^{-\lambda_b t_m} \frac{(\lambda_b t_m)^k}{k!}$.

For a fixed k, $t_e = \max_{1 \leq i \leq k}\{t_{ai}\} = t_{ak}$, where t_{ai} denotes the arrival time of the i-th bus ($1 \leq i \leq k$). For Poisson arrivals, t_{a1}, t_{a2}, …, t_{ak} are statistically equivalent to the order statistics of k random variables that are uniformly distributed in $[0, t_m]$ (Ross, 2014). Hence, the cumulative distribution function of t_e given k is:

$$F_{t_e|k}(s) = \Pr(t_e \leq s|k) = \Pr(t_{a1}, \ldots, t_{ak} \leq s|k) = \left(\frac{s}{t_m}\right)^k, s \in [0, t_m].$$

The probability density function of t_e given $k \geq 1$ is therefore:

$$f_{t_e|k}(s) = k\frac{s^{k-1}}{t_m^k}, s \in [0, t_m].$$

Now we consider the signal delay of T-cars in the cycle following the green extension. Since the G_T phase in this cycle is trimmed by t_e, any bus arrivals in $[t_m, T - G_T + t_e]$ will have to wait till time $T - G_T + t_e$ to discharge into the intersection. Those buses' average delay is $\frac{1}{2}(T - G_T + t_e - t_m)$. The expected total delay of those buses is $\frac{\lambda_b}{2}(T - G_T + t_e - t_m)^2$ for a given t_e. Since an extension of the previous G_T phase only affects the timing of the present cycle (but not any later cycles),

the expected total bus delay in a cycle can be obtained by taking the expectation of $\frac{\lambda_b}{2}(T-G_T+t_e-t_m)^2$:

$$E\left[\frac{\lambda_b}{2}(T-G_T+t_e-t_m)^2\right]$$

$$=E\left[E\left[\frac{\lambda_b}{2}(T-G_T+t_e-t_m)^2|k\right]\right]$$

$$=e^{-\lambda_b t_m}\cdot\frac{\lambda_b}{2}(T-G_T-t_m)^2+\sum_{k=1}^{\infty}e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^k}{k!}\cdot E\left[\frac{\lambda_b}{2}(T-G_T+t_e-t_m)^2|k\right]$$

$$=e^{-\lambda_b t_m}\cdot\frac{\lambda_b}{2}(T-G_T-t_m)^2+\sum_{k=1}^{\infty}e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^k}{k!}\cdot\int_{s=0}^{t_m}\frac{\lambda_b}{2}(T-G_T+s-t_m)^2\cdot k\frac{s^{k-1}}{t_m^k}ds$$

$$=e^{-\lambda_b t_m}\cdot\frac{\lambda_b}{2}(T-G_T-t_m)^2+\sum_{k=1}^{\infty}e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^k}{k!}\cdot\frac{\lambda_b}{2}\frac{k}{t_m^k}\int_{s=0}^{t_m}\left(s^{k+1}+2(T-G_T-t_m)s^k+(T-G_T-t_m)^2s^{k-1}\right)ds$$

$$=e^{-\lambda_b t_m}\cdot\frac{\lambda_b}{2}(T-G_T-t_m)^2+\sum_{k=1}^{\infty}e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^k}{k!}\cdot\frac{\lambda_b}{2}\frac{k}{t_m^k}\cdot\left(\frac{t_m^{k+2}}{k+2}+\frac{2(T-G_T-t_m)}{k+1}t_m^{k+1}+\frac{(T-G_T-t_m)^2}{k}t_m^k\right)$$

$$=e^{-\lambda_b t_m}\cdot\frac{\lambda_b}{2}(T-G_T-t_m)^2+\frac{\lambda_b}{2}\sum_{k=1}^{\infty}e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^k}{k!}\cdot\left(\frac{kt_m^2}{k+2}+\frac{2k(T-G_T-t_m)}{k+1}t_m+(T-G_T-t_m)^2\right)$$

$$=e^{-\lambda_b t_m}\cdot\frac{\lambda_b}{2}(T-G_T-t_m)^2+\frac{\lambda_b}{2}\sum_{k=1}^{\infty}e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^k}{k!}\cdot\left((T-G_T)^2+\left(\frac{-2t_m(T-G_T)}{k+1}\right)+\frac{2t_m^2}{(k+1)(k+2)}\right)$$

$$=e^{-\lambda_b t_m}\cdot\frac{\lambda_b}{2}(T-G_T-t_m)^2+\frac{\lambda_b}{2}\left((T-G_T)^2\cdot\sum_{k=1}^{\infty}e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^k}{k!}+\left(\frac{-2(T-G_T)}{\lambda_b}\right)\cdot\sum_{k=1}^{\infty}e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^{k+1}}{(k+1)!}+\frac{2}{\lambda_b{}^2}\cdot\sum_{k=1}^{\infty}e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^{k+2}}{(k+2)!}\right).$$

From the fact $\sum_{k=0}^{\infty}e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^k}{k!}=1$, we have: $\sum_{k=1}^{\infty}e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^k}{k!}=1-e^{-\lambda_b t_m}$, $\sum_{k=1}^{\infty}e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^{k+1}}{(k+1)!}=\sum_{k=2}^{\infty}e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^k}{k!}=1-e^{-\lambda_b t_m}(1+\lambda_b t_m)$, and $\sum_{k=1}^{\infty}e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^{k+2}}{(k+2)!}=1-e^{-\lambda_b t_m}\left(1+\lambda_b t_m+\frac{(\lambda_b t_m)^2}{2}\right)$. Hence, the above expected total delay is equal to:

$$e^{-\lambda_b t_m}\cdot\frac{\lambda_b}{2}(T-G_T-t_m)^2+\frac{\lambda_b}{2}\left((T-G_T)^2\cdot\left(1-e^{-\lambda_b t_m}\right)+\left(\frac{-2(T-G_T)}{\lambda_b}\right)\cdot\left(1-e^{-\lambda_b t_m}(1+\lambda_b t_m)\right)+\frac{2}{\lambda_b{}^2}\cdot\left(1-e^{-\lambda_b t_m}\left(1+\lambda_b t_m+\frac{(\lambda_b t_m)^2}{2}\right)\right)\right)$$

$$=\left(\frac{\lambda_b}{2}(T-G_T)^2-(T-G_T)+\frac{1}{\lambda_b}\right)+e^{-\lambda_b t_m}\left(T-G_T-t_m-\frac{1}{\lambda_b}\right).$$

The expected bus delay equals the above divided by the expected number of buses in a cycle, $\lambda_b T$. Thus, we have equation (9) and (11):

$$E[W_{B,3}]=\frac{1}{\lambda_b T}\left(\left(\frac{\lambda_b}{2}(T-G_T)^2-(T-G_T)+\frac{1}{\lambda_b}\right)+e^{-\lambda_b t_m}\left(T-G_T-t_m-\frac{1}{\lambda_b}\right)\right).$$

Now we derive $E[t_L]$ for the integrated design. Since a regular G_T phase may contain a redundant time of $t_R=G_T-\frac{g_T n_T}{N-1}$ (there are $N-1$ car lanes in the sorting area), there would be no capacity loss if $t_e\le t_R$. In particular, if $t_R\ge t_m$, we will always have $t_L=0$.

On the other hand, if $t_e>t_R$, the corresponding g_T phase will be trimmed by $(t_e-t_R)\frac{N-1}{n_T}$. Hence,

$$t_L = \frac{N-1}{n_T}\max\{0, t_e - t_R\}.$$

Thus, when $t_R < t_m$, we have:

$$E[t_L|t_R < t_m] = E\big[E[t_L|k, t_R < t_m]\big]$$

$$= e^{-\lambda_b t_m}\cdot 0 + \sum_{k=1}^{\infty} e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^k}{k!}\cdot E\left[\frac{N-1}{n_T}\max\{0, t_e - t_R\}\,\middle|\, k, t_R < t_m\right]$$

$$= \frac{N-1}{n_T}\sum_{k=1}^{\infty} e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^k}{k!}\cdot \int_{s=t_R}^{t_m} k\frac{s^{k-1}}{t_m^k}\cdot(s - t_R)ds$$

$$= \frac{N-1}{n_T}\sum_{k=1}^{\infty} e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^k}{k!}\cdot\left(t_m - t_R - \frac{1}{k+1}\left(t_m - \frac{t_R^{k+1}}{t_m^k}\right)\right)$$

$$= \frac{N-1}{n_T}\left((t_m - t_R)\big(1 - e^{-\lambda_b t_m}\big) - \frac{1}{\lambda_b t_m}\sum_{k=1}^{\infty} e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^{k+1}}{(k+1)!}\cdot\left(t_m - \frac{t_R^{k+1}}{t_m^k}\right)\right)$$

$$= \frac{N-1}{n_T}\left((t_m - t_R)\big(1 - e^{-\lambda_b t_m}\big) - \frac{1}{\lambda_b}\sum_{k=1}^{\infty} e^{-\lambda_b t_m}\frac{(\lambda_b t_m)^{k+1}}{(k+1)!} + \frac{1}{\lambda_b}\sum_{k=1}^{\infty} e^{-\lambda_b t_m}\frac{(\lambda_b t_R)^{k+1}}{(k+1)!}\right)$$

$$= \frac{N-1}{n_T}\left((t_m - t_R)\big(1 - e^{-\lambda_b t_m}\big) - \frac{1}{\lambda_b}\Big(1 - e^{-\lambda_b t_m}(1 + \lambda_b t_m)\Big) + \frac{1}{\lambda_b}e^{-\lambda_b(t_m - t_R)}\Big(1 - e^{-\lambda_b t_R}(1 + \lambda_b t_R)\Big)\right)$$

$$= \frac{N-1}{n_T}\left(t_m - t_R - \frac{1}{\lambda_b}\big(1 - e^{-\lambda_b(t_m - t_R)}\big)\right).$$

Combining the cases of $t_R \geq t_m$ and $t_R < t_m$, we have equation (13):

$$E[t_L] = \max\left\{0, (t_m - G_T)\frac{N-1}{n_T} + g_T\right\} - \frac{N-1}{\lambda_b n_T}\left(1 - e^{-\lambda_b \max\left\{0, t_m - G_T + \frac{g_T n_T}{N-1}\right\}}\right).$$

E.1 Considering the prediction error of bus arrival time

Now consider that a green extension will only serve those buses with predicted arrival time in $[0, t_m - \epsilon]$. Thus, the probability of $t_e = 0$ is $e^{-\lambda_b(t_m - \epsilon)}$, and the probability that $k > 0$ buses are served during the green extension is $e^{-\lambda_b(t_m-\epsilon)}\frac{(\lambda_b(t_m-\epsilon))^k}{k!}$.

When $k > 0$, let $t_e' = t_e - \epsilon$, we have $t_e' = \max\limits_{1\leq i\leq k}\{t_{ai}\} = t_{ak}$, where t_{ai} is the *predicted* arrival time of the i-th bus ($1 \leq i \leq k$). Then, $f_{t_e'|k}(s) = k\frac{s^{k-1}}{(t_m-\epsilon)^k}$, $s \in [0, t_m - \epsilon]$.

Note now that a bus that is predicted to arrive in $[t_m - \epsilon, T - G_T + t_e' + \epsilon]$ will have to wait until $T - G_T + t_e' + \epsilon$ to discharge into the intersection. Thus, the average delay per bus for these delayed buses is $\frac{1}{2}(T - G_T + t_e' + 2\epsilon - t_m)$, and the expected total delay of those buses is $\frac{\lambda_b}{2}(T - G_T + t_e' + 2\epsilon - t_m)^2$. (However, some buses that are predicted to arrival after $t_m - \epsilon$ may arrive earlier due to the uncertainty and catch up with the extended green period. Ignoring these buses would overestimate the bus delays. This makes our estimation of expected bus delay conservative.)

Comparing $E\left[\frac{\lambda_b}{2}(T - G_T + t_e' + 2\epsilon - t_m)^2\right]$ in this section against $E\left[\frac{\lambda_b}{2}(T - G_T + t_e - t_m)^2\right]$ in Appendix E, we find that the former comes from the latter when $T - G_T$ and t_m are replaced by $T - G_T - \epsilon$ and $t_m - \epsilon$, respectively. Thus, using the result derived previously, we have (14) immediately:

791 $E[W_{B,3}],\ E[W_{B,4}] = \frac{1}{\lambda_b T}\left(\left(\frac{\lambda_b}{2}(T - G_T + \epsilon)^2 - (T - G_T + \epsilon) + \frac{1}{\lambda_b}\right) + e^{-\lambda_b(t_m-\epsilon)}\left(T - G_T + 2\epsilon -\right.\right.$

792 $\left.\left. t_m - \frac{1}{\lambda_b}\right)\right)$.

793 Regarding the T-car capacity for the integrated design, equation (12) still holds. The only
794 difference may occur in the calculation of $E[t_L]$. However, note that:

795 $E[t_L|t_R < t_m] = E\left[E[t_L|k, t_R < t_m]\right]$

796 $= \frac{N-1}{n_T}\sum_{k=1}^{\infty} e^{-\lambda_b(t_m-\epsilon)}\frac{\left(\lambda_b(t_m-\epsilon)\right)^k}{k!} \cdot \int_{s=t_R-\epsilon}^{t_m-\epsilon} k\frac{s^{k-1}}{(t_m-\epsilon)^k} \cdot \big(s - (t_R - \epsilon)\big)ds.$

797 Comparing the above formula against the one in Appendix E, we find that the former can be
798 obtained by replacing t_m and t_R by $t_m - \epsilon$ and $t_R - \epsilon$ in the latter. Hence, the result is:

799 $E[t_L|t_R < t_m] = \frac{N-1}{n_T}\left((t_m - \epsilon) - (t_R - \epsilon) - \frac{1}{\lambda_b}\left(1 - e^{-\lambda_b\left((t_m-\epsilon)-(t_R-\epsilon)\right)}\right)\right) = \frac{N-1}{n_T}\left(t_m - t_R -\right.$

800 $\left.\frac{1}{\lambda_b}\left(1 - e^{-\lambda_b(t_m-t_R)}\right)\right)$.

801 This means equation (13) also holds true when prediction errors in bus arrival time are
802 accounted for.

803 Appendix F. Derivation of (15)

804 Denote γ as the coefficient of variation of the car discharge headways from a queue. Xuan et al. (2011)
805 showed that if a lane in the sorting area is blocked by a left-behind car, then: i) the following cycle will
806 be fully wasted for that lane; and ii) the other lanes in the sorting area are not affected, given that cars
807 do not change lanes once sorted. Our following analysis is inspired by Xuan et al. (2011).

808 We first consider the case where the green extension scheme is not implemented. For simplicity,
809 here we deviate slightly from the modeling framework of this paper, by assuming that the pre-signal is
810 adaptive, and it can control the exact number of vehicles released into the sorting area in each phase
811 (Xuan et al., 2011). When no lane is blocked, the pre-signal sends exactly $k(N-1)$ T-cars into the
812 sorting area in each g_T phase, so that k T-cars enter each lane of the sorting area. If a lane is blocked,
813 the pre-signal sends enough cars in the following cycle to fill in the remaining lanes only, so that no
814 more vehicle will be stranded in the blocked lane. According to the central limit theorem, the time
815 needed to discharge all the k cars roughly follows a normal distribution with mean $\frac{k}{q_S}$ and standard
816 deviation $\sqrt{k}\frac{\gamma}{q_S}$. Thus, the probability that a lane is blocked at the end of a G_T phase is approximately
817 $1 - \Phi\left(\frac{G_T - \frac{k}{q_S}}{\sqrt{k}\frac{\gamma}{q_S}}\right)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.
818 Let $\alpha = \frac{G_T - \frac{k}{q_S}}{\sqrt{k}\frac{\gamma}{q_S}}$. The α should be set to a fairly large value to ensure that lane blockages rarely occur.
819 For example, if $\alpha = 2$, the probability of lane blockage is 2.5%; while if $\alpha = 3$, that probability is only
820 0.14%.

821 In other words, to ensure the lane blockage probability is no more than $1 - \Phi(\alpha)$, k must
822 satisfy $\alpha \le \frac{G_T - \frac{k}{q_S}}{\sqrt{k}\frac{\gamma}{q_S}}$. This means $k + \alpha\gamma\sqrt{k} - q_S G_T \le 0$. Solving this inequality, we have $k \le k_m =$

$\left(\frac{-\alpha\gamma+\sqrt{\alpha^2\gamma^2+4q_S G_T}}{2}\right)^2 = \frac{\alpha^2\gamma^2+2q_S G_T-\alpha\gamma\sqrt{\alpha^2\gamma^2+4q_S G_T}}{2}$. Thus, without considering the impact of BSP, the T-car capacity is approximately bounded by $\frac{k_m(N-1)}{T} \cdot \Phi(\alpha)$. Recall that for every lane blockage, a full lane-cycle will be wasted.

Now by considering the effect of green extension, the above bound can be modified to $\frac{k_m(N-1)}{T} \cdot \Phi(\alpha) \cdot \left(1-\frac{E[t_L]}{g_T}\right)$, where $E[t_L]$ is given by (13). This modification is conservative. To see why, note that g_T would be redundant when the per-cycle number of T-cars entering the sorting area is further limited by $k_m(N-1)$. Thus, reducing a g_T by t_L due to a green extension may not reduce the number of T-cars admitted into the sorting area proportionally.

Combining this capacity bound with (12), we have (15).

References

Abdy, Z.R., Hellinga, B.R., 2011. Analytical method for estimating the impact of transit signal priority on vehicle delay. *Journal of Transportation Engineering* 137(8), 569-600.

Ahuja, S., Priest, N., Vuren, T., 2003. Public transport priority schemes: Comparing microsimulation with traditional TRANSYT and LINSIG models. In: *Proceedings of the European Transport Conference (ETC)*, Strasbourg, France.

Bie, Y., Wang, D., Qi, H., 2012. Prediction model of bus arrival time at signalized intersection using GPS data. *Journal of Transportation Engineering* 138(1), 12-20.

Bie, Y., Liu, Z., Wang, H., 2020. Integrating bus priority and presignal method at signalized intersection: Algorithm development and evaluation. *Journal of Transportation Engineering, Part A: Systems* 146(6), 04020044.

Chen, H., 2016. Improving public transit systems at intersection and city-wide scales. *Ph.D. Dissertation*. University of California, Berkeley.

Daganzo, C.F., 1997. Fundamentals of transportation and traffic operations. Oxford: Pergamon.

Dion, F., Rakha, H., Zhang, Y., 2004. Evaluation of potential transit signal priority benefits along a fixed-time signalized arterial. *Journal of Transportation Engineering* 130(3), 294-303.

Farid, Y. Z., Christofa, E., Collura, J., 2018. An analytical model to conduct a person-based evaluation of transit preferential treatments on signalized arterials. *Transportation Research Part C* 90, 411-432.

Garrow, M., Machemehl, R., 1997. Development and Evaluation of Transit Signal Priority Strategies. *Technical Report* SWUTC/97/472840-00068-1, University of Texas at Austin.

Gayah, V.V., Guler, S.I., Gu, W., 2016. On the impact of obstructions on the capacity of nearby signalised intersections. *Transportmetrica B: Transport Dynamics* 4(1), 48-67.

Gu, W., Chen, H., Xuan, Y., 2015. Providing bus signal priority without damaging car discharge capacities. *Conference on Advanced Systems in Public Transport*, Rotterdam, the Netherlands, July 2015.

Gu, W., Gayah, V.V., Cassidy, M.J., Saade N., 2014. On the impacts of bus stops near signalized intersections: Models of car and bus delays. *Transportation Research Part B* 68, 123-140.

Gu, W., Cassidy, M.J., Gayah, V.V., Ouyang, Y., 2013. Mitigating impacts of near-side bus stops on cars. *Transportation Research Part B* 47(1), 42-56.

Guler, S.I., Menendez, M., 2014. Analytical formulation and empirical evaluation of pre-signals for bus priority. *Transportation Research Part B* 64, 41-53.

He, H., Guler, S.I., Menendez, M., 2016. Adaptive control algorithm to provide bus priority with a pre-signal. *Transportation Research Part C* 64, 28-44.

Jin, X., Zhang, Y., Wang, F., Li, L., Yao, D., Su, Y., Wei, Z., 2009. Departure headways at signalized intersections: a log-normal distribution model approach. *Transportation Research Part C* 17(3), 318-327.

Li, F., Tang, K., Yao, J., Li, K., 2017. Real-time queue length estimation for signalized intersections using vehicle trajectory data. *Transportation Research Record* 2623(1), 49-59.

Lighthill, M.J., Whitham, G.B., 1955. On kinematic waves. I. Flood movements in long rivers. II. A theory of traffic flow on long crowded roads. In: *Proceedings of the Royal Society* (London) A 229, 281-345.

Lin, Y., Yang, X., Zou, N., 2019. Passive transit signal priority for high transit demand: model formulation and strategy selection. *Transportation Letters* 11(3), 119-129.

Lin, Y., Yang, X., Zou, N., Franz, M., 2015. Transit signal priority control at signalized intersections: a comprehensive review. *Transportation Letters* 7(3), 168-180.

Luo, J., 2011. Promising implementation of the "Mixed Waiting Zone" (translated from Chinese). *Eastday.com*, April 21, 2011. http://xwcb.eastday.com/c/20110421/u1a875464.html (accessed on October 1, 2019.)

Ma, W., Head, K.L., Feng, Y., 2014. Integrated optimization of transit priority operation at isolated intersections: A person-capacity-based approach. *Transportation Research Part C* 40, 49-62.

Ma, W., Liu, Y., Zhao, J., Wu, N., 2017. Increasing the capacity of signalized intersections with left-turn waiting areas. *Transportation Research Part A* 105, 181-196.

Mei, Y., Gu, W., Chung, E.C., Li, F., Tang, K., 2019. A Bayesian approach for estimating vehicle queue lengths at signalized intersections using probe vehicle data. *Transportation Research Part C* 109, 233-249.

Newell, G.F., 1993. A simplified theory of kinematic waves in highway traffic, part I~III. *Transportation Research Part B* 27(4), 281-313.

Newell, G.F., 1965. Approximation methods for queues with application to the fixed-cycle traffic light. *SIAM Review* 7(2), 223-240.

Qiu, F., Li, W., Zhang, J., Zhang, X., Xie, Q., 2015. Exploring suitable traffic conditions for intermittent bus lanes. *Journal of Advanced Transportation* 49(3), 309-325.

Richards, P.I., 1956. Shockwaves on the highway. *Operations Research* 4(1), 42-51.

Ross, S., 2014. *Introduction to Probability Models*, 11th Ed. Academic Press, New York.

Shi, L., 2016. Bus arrival time reliability analyses and dynamic prediction model based on multi-source data. *Master Thesis*, University of Alberta.

Skabardonis, A., Geroliminis, N., 2005. Real-time estimation of travel times on signalized arterials. In *Transportation and Traffic Theory: Flow, Dynamics and Human Interaction*, 387-406.

Tan, C. W., Park, S., Liu, H., Xu, Q., Lau, P., 2008. Prediction of transit vehicle arrival time for signal priority control: Algorithm and performance. *IEEE Transactions on Intelligent Transportation Systems* 9(4), 688-696.

The City of New York Government, 2014. http://www.nyc.gov/html/brt/html/about/bus-lanes.shtml.

Viegas, J., Lu, B., 2004. The intermittent bus lane signals setting within an area. *Transportation Research Part C* 12, 453-469.

Viti, F., Van Zuylen, H.J., 2010. Probabilistic models for queues at fixed control signals. *Transportation Research Part B* 44(1), 120-135.

907 Wolshon, B., Taylor, W.C., 1999. Analysis of intersection delay under real-time adaptive signal control.
908 *Transportation Research Part C* 7(1), 53-72.

909 Wu, J., Hounsell, N., 1998. Bus priority using pre-signals. *Transportation Research Part A* 32(8), 563-
910 583.

911 Wu, J., Liu, P., Qin, X., Zhou, H., Yang, Z., 2019. Developing an actuated signal control strategy to
912 improve the operations of contraflow left-turn lane design at signalized intersections. *Transportation*
913 *Research Part C* 104, 53-65.

914 Xuan, Y., 2011. Increasing the flow capacity of signalized intersections with pre-signals: Theory and
915 case study. *Ph.D. Dissertation*, University of California, Berkeley.

916 Xuan, Y., Daganzo, C.F., Cassidy, M.J., 2011. Increasing the capacity of signalized intersections with
917 separate left turn phases. *Transportation Research Part B* 45(5), 769-781.

918 Yang, X., Cheng, Y., 2017. Development of signal optimization models for asymmetric two-leg
919 continuous flow intersections. *Transportation Research Part C* 74, 306-326.

920 Zhao, J., Ma, W., Head, K.L., Yang, X., 2015. Optimal operation of displaced left-turn intersections: A
921 lane-based approach. *Transportation Research Part C* 61, 29-48.

922 Zhou, G., Gan A., 2005. Performance of transit signal priority with queue jumper lanes. *Transportation*
923 *Research Record* 1925, 265-271.