A model for real-time bus holding subject to vehicle capacity limits

Konstantinos Gkiotsalitis
Assistant Professor
University of Twente

Center for Transport Studies (CTS)
Department of Civil Engineering
P.O. Box 217
7500 AE Enschede
The Netherlands
Email: k.gkiotsalitis@utwente.nl

Eric C. van Berkum
Full Professor
University of Twente

Center for Transport Studies (CTS)
Department of Civil Engineering
P.O. Box 217
7500 AE Enschede
The Netherlands
Email: e.c.vanberkum@utwente.nl

99th Annual Meeting of the Transportation Research Board
Paper number: 20-00102

July 11, 2019

ABSTRACT
Two distinct directions of research have emerged for the vehicle holding problem: (i) single variable optimization approaches that determine the holding time of a single vehicle when it is about to depart from a bus stop; and, (ii) multivariable, periodic optimization approaches that use rather complex mathematical programs to determine the holding times of all running vehicles. Comprehensive mathematical programs that consider multiple decision variables cannot be easily solved in real time, and are typically reserved for periodic control in longer time horizons. For this reason, this study focuses on single variable optimization approaches which determine the holding time of a vehicle when it arrives at a control point stop. Up to now, single variable optimization methods resort to rather simple, rule-based control logics. One of them is the one-headway-based logic which determines the holding time of a bus based on its headway with its preceding bus without addressing other implications, such as overcrowding. To rectify this, we introduce a new nonlinear model for the single variable bus holding problem that considers the passenger demand and vehicle capacity limits. Then, we reformulate this problem to an easier-to-solve program with the use of slack variables and we prove that it can be solved to global optimality. A simulation-based investigation of the performance of our model against the performance of classic control logics that do not consider vehicle capacity limits is finally performed in bus line 302 in Singapore.

Keywords: timetabling; high-frequency services; robust optimization; transfer coordination; nonlinear programming

INTRODUCTION

Decisions regarding the operations of bus services are made at different planning stages. At the tactical planning stage, one has to determine the frequency Yu et al., Gkiotsalitis and Cats (*1, 2*), the timetable Sun et al., Wu et al. (*3, 4*), and the crew and vehicle schedules Wren and Rousseau, Gintner et al., Kliewer et al. (*5, 6, 7*) of every bus line. Tactical plans are communicated well in advance, and all stakeholders (i.e., public transport authorities/operators, bus drivers, passengers) are aware of them prior to the start of the daily operations Ceder (*8*).

The fixed service interval (time headway) of every bus line is determined from the tactical planning stage and is equal to the inverse of the service frequency Ceder (*8*). That is to say, a bus line with a service frequency of 6 trips per hour operates under a 10-minute time headway. The time headway of two trips, which is the time difference between the time instances they were at the same location, will henceforth be simply called *headway*.

The main challenge in high-frequency services with more than 5 trips per hour is to maintain the planned headways among buses at every bus stop Trompet et al. (*9*). If the demand and the travel times of all bus trips that operate in a service line are equal and stable, bus trips will maintain their even headways at all downstream stops. This will result in a regular service where the actual passenger waiting times at stops meet the passengers' expectations. Nevertheless, travel time and passenger demand variations during the actual operations result in unreliable and inconsistent services Chen et al., Daganzo (*10, 11*). Knoppers and Muller, Berrebi et al., Gkiotsalitis and Maslekar (*12, 13, 14*) and Knoppers and Muller (*12*) have shown that the fixed service intervals cannot be maintained at all stops. Indeed, even if buses are dispatched according to their planned headways, their headways are expected to deviate from their scheduled values as they are moving towards downstream stops Hans et al. (*15*). This leads to irregular services where buses are too close or too far away from each other instead of maintaining their scheduled headway.

To address the adverse effects of the demand and travel time variability, several flexible scheduling approaches have emerged over the past 40 years. Such flexible approaches have a shifted focus towards operational control that reacts to changes in quasi-real-time. Operational control includes a variety of options, such as bus holding Bartholdi III and Eisenstein, Delgado et al. (*16, 17*), stop-skipping Liu et al., Chen et al. (*18, 19*), short-turning Cortés et al. (*20*), interlining Gkiotsalitis et al. (*21*), re-scheduling Gkiotsalitis and Stathopoulos (*22*), and speed control Daganzo and Pilachowski, Muñoz et al. (*23, 24*). All options aim at improving the reliability of services during the actual operations and correcting potential inconsistencies due to operational disruptions.

In this study, we specifically focus on the problem of real-time bus holding that holds buses at specific bus stops to reduce the deviation between the actual and the planned headways. In so doing, bus trips will maintain their even distribution, and the waiting times of passengers will be closer to their expected values. In its simplest form, bus holding holds a trip n at a stop s for a time period $x \geq 0$ if its actual headway with its preceding trip, $n-1$, is lower than the planned headway, H_s. This is the well-known one-headway-based holding logic which strives to maintain the planned headway between a trip n and its preceding one, $n-1$ (Fu and Yang (*25*)). Its simplicity is very useful when one wants to apply bus holding in real time because then the holding time of every trip should be determined immediately upon its arrival at the respective bus stop (Daganzo (*11*)). Other approaches do not consider only the headway between one bus trip, n, and its preceding trip, $n-1$, but also the headway with the following trip, $n+1$. Such approaches are known as two-headway-based methods (see Fu and Yang (*25*)).

An entirely different line of research determines the holding times of multiple bus trips, instead of only trip n, following a periodic optimization approach Gkiotsalitis and Cats (*26*). Periodic optimization approaches consider multiple decision variables and are based on iterative, finite-horizon optimization(s) of a bus holding model. At time t, the current state of the bus operations (i.e., current positions of running trips) is used as input and, together with the expected travel times within a relatively short time horizon $t+T$, the holding times of multiple running trips are determined. This is equivalent to scheduling the bus holding times of all running trips within a short time horizon with the use of travel time expectations Eberlein et al. (*27*). There are two main issues with the periodic optimization methods:

- if the number of bus trips, $i = \{1,2,...,\}$, that are expected to visit control point stops within a time period $t+T$ is too big, determining the holding times of all those trips results in complex, multivariable optimization problems that cannot be solved in real time (see Hickman, Sánchez-Martínez et al. (*28, 29*));
- if the short-term travel time predictions are not close to the realized travel times, the scheduled holding times of trips $i = \{1,2,...,\}$ might have a limited effect - or even be counterproductive. Due to that, their values should be recomputed every time new information becomes available. This will result in receding horizon control, or “rolled" rolling horizon optimization Eberlein et al. (*27*).

Evidently, periodic optimization approaches are not computationally efficient because they require to solve complex mathematical models with many holding decisions, from which only some will be implemented in practice by the time new information becomes available. This inefficiency is well-known in model predictive control (MDP), where multiple decisions are made but only some of them have the chance to be implemented by the time new information becomes available triggering a repeat of the optimization process Nikolaou (*30*).

Mathematical models for periodic bus holding control are very advanced, and some of them, such as Delgado et al., Sánchez-Martínez et al. (*17, 29*), incorporate the bus loads and vehicle capacity limitations in the bus holding optimization process. Notwithstanding this, their complex nature does not allow to compute a globally optimal bus holding solution in real time. This motivates our work: our study proposes an easy-to-solve mathematical program for the bus holding problem under capacity limitations that can determine (immediately) the holding time of a bus trip upon its arrival at a bus stop. To the best of the authors' knowledge, our proposed model is the first of its kind and is based on the modeling of the real-time bus holding problem as a regularity-based optimization problem under bus load variations and capacity limitations.

The remainder of this paper is structured as follows: in section 2, we model the bus holding problem with the objective of maintaining the service regularity while meeting the vehicle capacity limits. This problem is proved to be nonlinear and non-smooth; thus, it cannot be solved to global optimality because its functions are not differentiable at every point in their domain. In section 3, we reformulate the aforementioned bus holding problem by introducing slack variables. Then, we prove that its reformulated version has a globally optimal solution because its objective function is convex, and its feasible region is a convex set. In section 4 we perform numerical experiments in idealized, toy networks and bus line 302 in Singapore to demonstrate the potential improvement of using our model, instead of control logics that do not consider the capacity limits of the running vehicles. Finally, in section 5, we summarize our findings and propose potential future directions.

PROBLEM DEFINITION AND MATHEMATICAL PROGRAM

A typical objective of bus holding strategies is to minimize the variation of the actual (realized) headways from their scheduled values (also known as *target* or *ideal* headway values) Berrebi et al. (*13*). One of the most common approaches to achieving that is the two-headway-based bus holding method Ibarra-Rojas et al. (*31*). The two-headway-based bus holding strategy is not a periodic optimization approach. Instead, it is a rule-based method that determines the holding time of a trip n when it arrives at a control point stop s based on the realized and expected headway(s) with its preceding, $n-1$, and following, $n+1$, bus trips (see Fu and Yang (*25*)).

The holding logic of Fu and Yang (*25*), which does not consider capacity limitations, is summarized in algorithm 1. Alg.1 determines the holding time of trip n at stop s, and the holding decision is made when trip n has completed all its boardings/alightings and is ready to depart.

The notation used in Alg.1 is summarized in Fig.1, where

- t is the time when trip n has completed its boardings/alightings at stop s and is ready to depart,
- $d_{n,s}$ is the determined departure time of trip n from stop s,
- $d_{n,s}-t$ is the determined holding time of trip n at stop s,
- $d_{n-1,s}$ is the realized departure time of the preceding trip $n-1$ from stop s,
- H_s is the scheduled (target) headway of adjacent bus trips at stop s,
- $\tilde{d}_{n+1,s}$ is the expected departure time of the following trip, $n+1$, from stop s.

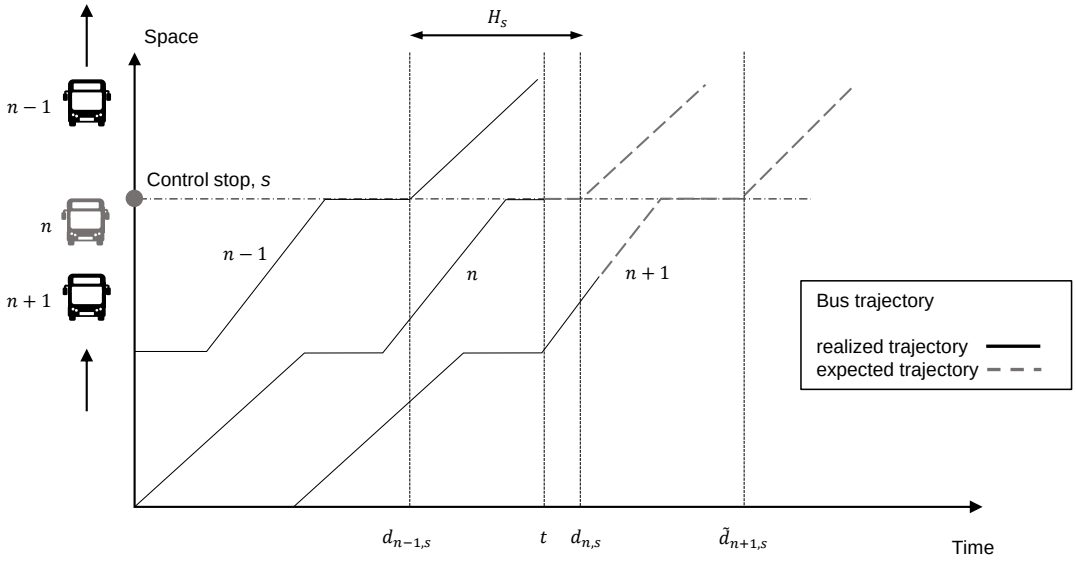

FIGURE 1 : Realized and expected trajectories of the preceding, $n-1$, and following, $n+1$, bus trip of n.

Algorithm 1 Two-headway-based holding control logic of Fu and Yang (*25*)

0: If $t < d_{n-1,s} + H_s$, then:

1: If $\frac{1}{2}(\tilde{d}_{n+1,s} - d_{n-1,s}) < H_s$, then:

2: $d_{n,s} = d_{n-1,s} + H_s$

3: Else:

4: $d_{n,s} = d_{n-1,s} + [\frac{1}{2}(\tilde{d}_{n+1,s} - d_{n-1,s}) + H_s]/2$

5: Else:

6: $d_{n,s} = t$

In the control logic of Alg.1, if bus trip n completes its boardings/alightings at time $t \geq d_{n-1,s} + H_s$, it has to depart immediately because it is behind schedule. On the contrary, if $t < d_{n-1,s} + H_s$, trip n has some buffer time which can be spent at stop s to reduce the deviation from the target headway(s). This simple control logic will be used as a *benchmark* in our numerical experiments to investigate the potential benefits of our single variable bus holding model that considers vehicle capacity limitations.

Proceeding to the introduction of our method, we present the main assumptions of our work, which are also commonly used in past literature related to the bus holding problem of high-frequency services:

(1) In high-frequency services, passengers who cannot board a bus will wait for the next trip of the same bus line because their waiting times are relatively small Delgado et al., Muñoz et al., Delgado et al. (*17, 24, 32*).

(2) Passengers cannot coordinate their arrivals at stops to the arrival times of buses at high-frequency services Berrebi et al. (*33*). Thus, we assume a demand-based passenger arrival rate, λ_s, at any stop s Delgado et al., Fu and Yang (*17, 25*).

(3) The allowed holding time of buses at stops has an upper (maximum) limit, ζ, due to the inconvenience caused to on-board passengers (typically, this is set to 90 seconds Cortés et al. (*34*)).

To formulate our bus holding problem that considers vehicle capacity limits, we introduce the following notation.

Notation

Sets/Indices

$S = \langle 1, 2, ... \rangle$	ordered set of bus stops.
n	index of the trip for which a holding decision needs to be made at the current time instance.
$n-1$	index of the preceding trip of trip n.
$n+1$	index of the following trip of trip n.
s	specific bus stop at which a holding decision for trip n needs to be made. Note that $s \in S \setminus \{1, \|S\|\}$.

Parameters

t	time when bus trip n has completed its boardings/alightings at stop s and is ready to depart if there is no further holding.
$d_{n-1,s}$	realized departure time of trip $n-1$ from stop s.
λ_s	arrival rate of passengers at stop s (i.e., passengers per sec).
c_j	capacity of bus trip j, where $j \in \{n-1, n, n+1\}$.
ϕ_n	observed bus load of trip n at time t including the number of passengers who are refused to board trip n at stop s due to overcrowding. By definition, ϕ_n can be greater than c_n.
$\tilde{l}_{n+1}$	expected bus load of trip $n+1$ at the time of its arrival at stop s.
$\tilde{\beta}_{n+1}$	expected passenger alightings of bus trip $n+1$ at stop s.
$\tilde{a}_{n+1,s}$	expected arrival time of trip $n+1$ at stop s.
H_s	target (ideal) headway of adjacent trips at stop s.
t_b	required time for each passenger boarding.
t_a	required time for each passenger alighting.
ζ	maximum allowed holding time.
M_1, M_2	very large numbers, where $M_1 \gg M_2 \gg 0$.

Decision Variable

x	holding time of trip n at stop s. Note that $\{x \in \mathbb{R} \mid 0 \leq x \leq \zeta\}$ according to assumption (3).

Variables

$d_{n,s}$	departure time of trip n from stop s. Note that $d_{n,s} \triangleq t + x$.
$\tilde{d}_{n+1,s}$	expected departure time of trip $n+1$ from stop s.
l_n	stranded passengers by bus trip n at stop s.

Problem Objective

The objective of the bus holding problem in high-frequency services is to adhere to the target (scheduled) headways. When we determine the holding time of trip n at stop s, we strive to minimize the *squared deviation* between the realized/expected headways with its adjacent trips, $n-1$, $n+1$, and the ideal headway, H_s.

This is expressed in Eq.(1) where $(t+x)$ is the determined departure time of trip n from stop s, $d_{n-1,s}$ the realized departure time of trip $n-1$ from stop s, and $\tilde{d}_{n+1,s}$ the expected departure time of trip $n+1$ from stop s. Note that $\tilde{d}_{n+1,s}$ is an expected value because trip $n+1$ has not arrived at

stop s when the holding decision of trip n is made.

$$f(x) \triangleq \big((t+x) - d_{n-1,s} - H_s\big)^2 + \big(\tilde{d}_{n+1,s} - (t+x) - H_s\big)^2 \tag{1}$$

We should note here that Eq.(1) uses the *squared deviation* between the expected/realized headways and their target values. This is in line with the key performance indicators used to monitor the regularity of bus services Trompet et al., Newell (*9, 35*).

Constraints and Infeasibility

A first constraint when we consider the vehicle capacity limits is that trip n cannot serve more passengers than its capacity, c_n. This can be expressed as:

$$\phi_n + x\lambda_s \leq c_n \tag{2}$$

where $x\lambda_s$ is the number of additional passengers that are willing to board bus trip n if it is held at stop s for time x after it completes its boardings/alightings. Additionally, ϕ_n is the sum of the bus load of trip n and the number of (potentially) stranded passengers when it has completed its boardings/alightings at stop s.

Lemma 2.1. $\exists\ \phi_n \mid \phi_n + x\lambda_s > c_n,\ \forall x \in \mathbb{R}_{\geq 0}$.

Proof. ϕ_n is the observed bus load of trip n at time t plus the number of passengers unable to board trip n at stop s due to overcrowding. Hence, ϕ_n does not have an upper bound in $\mathbb{R}_{\geq 0}$. In contrast, the capacity of trip n, c_n, is a fixed integer number in $\mathbb{R}_{\geq 0}$. Hence, $\exists\ \phi_n \in \mathbb{R}_{\geq 0} \mid \phi_n > c_n$. Additionally, $x\lambda_s \geq 0 \because x, \lambda_s \geq 0$. Therefore, for $\phi_n > c_n \Rightarrow \phi_n + x\lambda_s > c_n$. □

Lemma 2.1 proves that there is no holding time $x \geq 0$ which can guarantee that the capacity of trip n suffices. Thus, the number of stranded passengers, l_n, by bus trip n at stop s can be expressed as:

$$l_n \triangleq \max(0, \phi_n + x\lambda_s - c_n) \tag{3}$$

Since constraint $\phi_n + x\lambda_s \leq c_n$ cannot be always satisfied, it can be perceived as a *soft* constraint which is allowed to be violated if, and only if, our holding time x cannot ensure that there are no stranded passengers by bus trip n at stop s. This soft constraint is added to the objective function as a penalty term $M_1 \max(0, \phi_n + x\lambda_s - c_n)$:

$$f(x) \triangleq \big((t+x) - d_{n-1,s} - H_s\big)^2 + \big(\tilde{d}_{n+1,s} - (t+x) - H_s\big)^2 + M_1 \max(0, \phi_n + x\lambda_s - c_n) \tag{4}$$

Note that the very large positive number M_1 in the penalty term $M_1 \max(0, \phi_n + x\lambda_s - c_n)$ ensures that the satisfaction of constraint $\phi_n + x\lambda_s \leq c_n$ is prioritized over $\big((t+x) - d_{n-1,s} - H_s\big)^2 + \big(\tilde{d}_{n+1,s} - (t+x) - H_s\big)^2$. Indeed, if $\phi_n + x\lambda_s \leq c_n$, then this solution does not add any penalty to the objective function since $M_1 \max(0, \phi_n + x\lambda_s - c_n) = 0$. In reverse, when $\phi_n + x\lambda_s > c_n$, the penalty term penalizes the objective function by a very large number $M_1(\phi_n + x\lambda_s - c_n)$ and directs the program towards another solution x that reduces the value of $M_1 \max(0, \phi_n + x\lambda_s - c_n)$ as much as possible. Consequently, a solution x that minimizes the objective function would be such that the number of stranded passengers by bus trip n at stop s is reduced to the greatest extent.

A second constraint is related to the vehicle capacity limit of the following trip, $n+1$. Note that the vehicle capacity limit of the preceding trip, $n-1$, is not considered because our decision variable, x, cannot affect its value. When trip $n+1$ arrives at stop s it has a bus load $\tilde{l}_{n+1}$ and is expected to alight $\tilde{\beta}_{n+1}$ passengers. Because of the time needed for the alightings, $\tilde{\beta}_{n+1}t_a$, we get $\tilde{\beta}_{n+1}t_a\lambda_s$ more passenger boardings. In addition, the stranded passengers by trip n, l_n, are willing to board trip $n+1$. Furthermore, by the time trip n departs stop s, $(t+x)$, until trip $n+1$ arrives there, we have $(\tilde{a}_{n+1,s}-(t+x))\lambda_s$ more passengers willing to board trip $n+1$. Thus, the expected bus load of trip $n+1$ when it departs from stop s is $\tilde{l}_{n+1}-\tilde{\beta}_{n+1}+\tilde{\beta}_{n+1}t_a\lambda_s+l_n+(\tilde{a}_{n+1,s}-(t+x))\lambda_s$. Note that this is the lowest possible bus load of trip $n+1$ when it departs from stop s because the holding time of trip $n+1$ at stop s is not factored in since it is not a decision variable at this time instance.

Remark 1. *At stop s the number of passengers willing to board trip $n+1$ is $\tilde{\beta}_{n+1}t_a\lambda_s+l_n+(\tilde{a}_{n+1,s}-(t+x))\lambda_s$. While boarding those passengers, $(\tilde{\beta}_{n+1}t_a\lambda_s+l_n+(\tilde{a}_{n+1,s}-(t+x))\lambda_s)t_b\lambda_s$ more passengers will arrive at stop s and will be willing to board trip $n+1$. While boarding the new passengers, more passengers, $(\tilde{\beta}_{n+1}t_a\lambda_s+l_n+(\tilde{a}_{n+1,s}-(t+x))\lambda_s)t_b^2\lambda_s^2$, will arrive and this iterative procedure results in a vicious circle. Thus, it is not possible to establish a closed-form mathematical expression that determines exactly the number of passengers willing to board trip $n+1$ if we assume that the time needed for every extra boarding will always generate new boarding demand. To alleviate this, we consider only the passengers that will arrive while boarding passengers $\tilde{\beta}_{n+1}t_a\lambda_s+l_n+(\tilde{a}_{n+1,s}-(t+x))\lambda_s$ and we assume that the number of passenger arrivals during subsequent boardings is negligibly small. That is to say, while boarding passengers $(\tilde{\beta}_{n+1}t_a\lambda_s+l_n+(\tilde{a}_{n+1,s}-(t+x))\lambda_s)t_b\lambda_s$ the number of new passengers arriving at the stop is insignificant because the time duration of $(\tilde{\beta}_{n+1}t_a\lambda_s+l_n+(\tilde{a}_{n+1,s}-(t+x))\lambda_s)t_b^2\lambda_s$ is infinitesimal and $(\tilde{\beta}_{n+1}t_a\lambda_s+l_n+(\tilde{a}_{n+1,s}-(t+x))\lambda_s)t_b^2\lambda_s^2\approx 0$. Despite this assumption, our formulation offers a more accurate representation of the potential passenger boardings compared to past works that oversimplify the problem by ingoring all passenger arrivals at a stop while the bus is dwelling (see Fu and Yang, Hickman, Marguier (25, 28, 36)).*

The assumption in Remark 1 allows us to determine a closed-form expression of the expected bus load of trip $n+1$ from stop s. This bus load should be lower or equal to the capacity of the bus that operates trip $n+1$. This is expressed in the inequality constraint of Eq.(5).

$$\tilde{l}_{n+1}-\tilde{\beta}_{n+1}+\left(\tilde{\beta}_{n+1}t_a\lambda_s+l_n+(\tilde{a}_{n+1,s}-(t+x))\lambda_s\right)(1+t_b\lambda_s)\leq c_{n+1} \tag{5}$$

Considering the capacity limit of trip $n+1$, the inequality constraint of Eq.(5) cannot be always satisfied for $x\in\mathbb{R}\mid 0\leq x\leq\zeta$.

Similarly to the capacity constraint of trip n, the capacity constraint of trip $n+1$ expressed in Eq.(5) can be perceived as a *soft* constraint which is allowed to be violated if, and only if, our holding time x cannot ensure that there are no stranded passengers by bus trip $n+1$ at stop s. This soft constraint is added to the objective function as a penalty term $M_2\max\Big[0,\tilde{l}_{n+1}-\tilde{\beta}_{n+1}+\left(\tilde{\beta}_{n+1}t_a\lambda_s+l_n+(\tilde{a}_{n+1,s}-t-x)\lambda_s\right)(1+t_b\lambda_s)-c_{n+1}\Big]$:

$$\begin{aligned} f(x)\triangleq\left((t+x)-d_{n-1,s}-H_s\right)^2+\left(\tilde{d}_{n+1,s}-(t+x)-H_s\right)^2+M_1\max(0,\phi_n+x\lambda_s-c_n)\\ +M_2\max\left[0,\tilde{l}_{n+1}-\tilde{\beta}_{n+1}+\left(\tilde{\beta}_{n+1}t_a\lambda_s+l_n+(\tilde{a}_{n+1,s}-t-x)\lambda_s\right)(1+t_b\lambda_s)-c_{n+1}\right] \end{aligned} \tag{6}$$

Remark 2. *Note that we use very large numbers M_1, M_2 to penalize the soft constraints related to the stranded passengers by bus trips n and $n+1$, respectively. Additionally, we set $M_1 \gg M_2$. $M_1 \gg M_2$ indicates that if trip n reaches its capacity limit, it will depart immediately from stop s even if this is expected to lead to the overcrowding of trip $n+1$. That is to say, we cannot hold an overcrowded bus trip, n, even if this has a positive effect to its following trip, $n+1$. This is realistic in practice because if bus trip n is held after reaching its capacity limit, it will cause inconvenience to both the driver and the passengers who are refused to board Trompet et al. (9).*

The expected departure time of trip $n+1$ from stop s, $\tilde{d}_{n+1,s}$, is equal to the expected arrival time at stop s, $\tilde{a}_{n+1,s}$, plus the required time for boardings/alightings (dwell time). The required time for boardings/alightings is $\tilde{\beta}_{n+1}t_a$ for passenger alightings and $\left(\tilde{\beta}_{n+1}t_a\lambda_s + l_n + (\tilde{a}_{n+1,s} - (t+x))\lambda_s\right)(1+t_b\lambda_s)t_b$ for passenger boardings. Note that all $\left(\tilde{\beta}_{n+1}t_a\lambda_s + l_n + (\tilde{a}_{n+1,s} - (t+x))\lambda_s\right)(1+t_b\lambda_s)$ passengers might not be able to board trip $n+1$ at stop s if its capacity limit is reached. Hence, the required time for passenger boardings is $\min\left[\left(\tilde{\beta}_{n+1}t_a\lambda_s + l_n + (\tilde{a}_{n+1,s} - (t+x))\lambda_s\right)(1+t_b\lambda_s)t_b, (c_{n+1} + \tilde{\beta}_{n+1} - \tilde{l}_{n+1})t_b\right]$.

This results to the *expected* departure time of trip $n+1$ from stop s:

$$\tilde{d}_{n+1,s} \triangleq \tilde{a}_{n+1,s} + \tilde{\beta}_{n+1}t_a + \min\Big[\left(\tilde{\beta}_{n+1}t_a\lambda_s + l_n + (\tilde{a}_{n+1,s} - t - x)\lambda_s\right)(1+t_b\lambda_s)t_b, \; (c_{n+1} + \tilde{\beta}_{n+1} - \tilde{l}_{n+1})t_b\Big] \tag{7}$$

Mathematical Program

The above-mentioned constraints form the following bus holding program, (Q), that determines the holding time x of trip n at time instance t.

$$\begin{aligned} (Q) \quad & \min_{x} \quad f(x) \\ & \text{s.t.} \quad (l_n, f, \tilde{d}_{n+1,s}) \;\big|\; (l_n, f, \tilde{d}_{n+1,s}) \text{ satisfy Eq.(3),(6),(7)} \\ & \quad\;\; 0 \le x \le \zeta \end{aligned} \tag{8}$$

Program (Q) is a nonlinear programming problem (NLP) because of the several non-smooth "max", "min" terms in the objective function $f(x)$, and the variables $l_n, \tilde{d}_{n+1,s}$. Due to the "max", "min" terms, program (Q) is not convex. Consequently, a solution method cannot guarantee the return of a globally optimal solution since the associated functions are not smooth and differentiable at every point in their domain.

REFORMULATION TO A QUADRATIC PROGRAM

Reformulation

Let us consider the nonlinear term $\max(0, \phi_n + x\lambda_s - c_n)$ of our objective function that appears also in the equality constraint $l_n = \max(0, \phi_n + x\lambda_s - c_n)$ expressed in Eq.(3). Note that the "max" term introduces non-smoothness to our objective function and our equality constraint. To rectify this, we introduce a slack variable ν_1 that, due to its bounds and the direction of optimization, will take

the value $\max(0, \phi_n + x\lambda_s - c_n)$ at the solution of the program. With the introduction of this slack variable v_1 that replaces $\max(0, \phi_n + x\lambda_s - c_n)$, the objective function becomes

$$\begin{aligned} f(x, v_1) \triangleq & \left(t + x - d_{n-1,s} - H_s\right)^2 + \left(\tilde{d}_{n+1,s} - t - x - H_s\right)^2 + M_1 v_1 \\ & + M_2 \max\left[0, \tilde{l}_{n+1} - \tilde{\beta}_{n+1} + \left(\tilde{\beta}_{n+1} t_a \lambda_s + v_1 + (\tilde{a}_{n+1,s} - t - x)\lambda_s\right)(1 + t_b \lambda_s) - c_{n+1}\right] \end{aligned} \tag{9}$$

and the expected departure time of trip $n+1$ from stop s:

$$\begin{aligned} \tilde{d}_{n+1,s} \triangleq \tilde{a}_{n+1,s} + \tilde{\beta}_{n+1} t_a + \min\Big[& \left(\tilde{\beta}_{n+1} t_a \lambda_s + v_1 + (\tilde{a}_{n+1,s} - t - x)\lambda_s\right)(1 + t_b \lambda_s) t_b, \\ & (c_{n+1} + \tilde{\beta}_{n+1} - \tilde{l}_{n+1}) t_b \Big] \end{aligned} \tag{10}$$

Hence, we reformulate program (Q) to

$$\begin{aligned} (\bar{Q}) \quad \min_{x, v_1} \quad & f(x, v_1) \\ \text{s.t.} \quad & (f, \tilde{d}_{n+1,s}) \mid (f, \tilde{d}_{n+1,s}) \text{ satisfy Eq.(9), (10)} \\ & v_1 \geq 0 \\ & v_1 \geq \phi_n + x\lambda_s - c_n \\ & 0 \leq x \leq \zeta \end{aligned} \tag{11}$$

Note that the term $M_1 v_1$ in the reformulated objective function $f(x, v_1)$ forces v_1 to receive its lowest possible value which is always greater than or equal to zero and has the equivalent effect of term $M_1 \max(0, \phi_n + x\lambda_s - c_n)$.

The objective function of program $(\bar{Q})$ has another non-smooth term: $M_2 \max\left[0, \tilde{l}_{n+1} - \tilde{\beta}_{n+1} + \left(\tilde{\beta}_{n+1} t_a \lambda_s + v_1 + (\tilde{a}_{n+1,s} - t - x)\lambda_s\right)(1 + t_b \lambda_s) - c_{n+1}\right]$. With the introduction of another slack variable v_2 that takes the value of the above term at the solution of the program, the objective function becomes

$$f(x, v_1, v_2) \triangleq \left(t + x - d_{n-1,s} - H_s\right)^2 + \left(\tilde{d}_{n+1,s} - t - x - H_s\right)^2 + M_1 v_1 + M_2 v_2 \tag{12}$$

and program $(\bar{Q})$ is reformulated to

$$\begin{aligned} (\hat{Q}) \quad \min_{x, v_1, v_2} \quad & f(x, v_1, v_2) \\ \text{s.t.} \quad & (f, \tilde{d}_{n+1,s}) \mid (f, \tilde{d}_{n+1,s}) \text{ satisfy Eq.(12), (10)} \\ & v_1 \geq 0 \\ & v_1 \geq \phi_n + x\lambda_s - c_n \\ & v_2 \geq 0 \\ & v_2 \geq \tilde{l}_{n+1} - \tilde{\beta}_{n+1} - c_{n+1} + \\ & \left(\tilde{\beta}_{n+1} t_a \lambda_s + v_1 + (\tilde{a}_{n+1,s} - t - x)\lambda_s\right)(1 + t_b \lambda_s) \\ & 0 \leq x \leq \zeta \end{aligned} \tag{13}$$

The equality constraint of Eq.(10) that defines the value of variable $\tilde{d}_{n+1,s}$ is the last non-smooth term in our reformulated program, $\hat{Q}$, due to the nonlinear term $\min\Big[\big(\tilde{\beta}_{n+1}t_a\lambda_s+\nu_1+(\tilde{a}_{n+1,s}-t-x)\lambda_s\big)(1+t_b\lambda_s)t_b, (c_{n+1}+\tilde{\beta}_{n+1}-\tilde{l}_{n+1})t_b\Big]$. As a remedy, we re-write $\tilde{d}_{n+1,s}$ as

$$\tilde{d}_{n+1,s} = \tilde{a}_{n+1,s}+\tilde{\beta}_{n+1}t_a+\big(\tilde{\beta}_{n+1}t_a\lambda_s+\nu_1+(\tilde{a}_{n+1,s}-t-x)\lambda_s\big)(1+t_b\lambda_s)t_b-\nu_2t_b \tag{14}$$

To simplify the notation, let $k \triangleq 1+t_b\lambda_s$, where $k\in\mathbb{R}_{\geq 0}$ because $t_b,\lambda_s\geq 0$. Then, the objective function can be re-written as

$$f(x,\nu_1,\nu_2)\triangleq\big(t+x-d_{n-1,s}-H_s\big)^2+\Big[\tilde{a}_{n+1,s}+\tilde{\beta}_{n+1}t_a+\big(\tilde{\beta}_{n+1}t_a\lambda_s+\nu_1+(\tilde{a}_{n+1,s}-t-x)\lambda_s\big)kt_b-\nu_2t_b-t-x-H_s\Big]^2+M_1\nu_1+M_2\nu_2 \tag{15}$$

and this leads to the reformulation of program $(\hat{Q})$ to

$$\begin{aligned}(\tilde{Q})\quad \min_{x,\nu_1,\nu_2}\quad & f(x,\nu_1,\nu_2)\\ \text{s.t.}\quad & (f)\ \big|\ (f) \text{ satisfies Eq.(15)}\\ & \nu_1\geq 0\\ & \nu_1\geq \phi_n+x\lambda_s-c_n\\ & \nu_2\geq 0\\ & \nu_2\geq \tilde{l}_{n+1}-\tilde{\beta}_{n+1}-c_{n+1}+\big(\tilde{\beta}_{n+1}t_a\lambda_s+\nu_1+(\tilde{a}_{n+1,s}-t-x)\lambda_s\big)k\\ & 0\leq x\leq\zeta\end{aligned} \tag{16}$$

This reformulation has introduced two slack variables (ν_1,ν_2) to transform the non-smooth, nonlinear program (Q) to a program $(\tilde{Q})$ with a quadratic objective function and linear inequality constraints that attains an equivalent solution to (Q). As it is shown in the following theorem, a locally optimal solution of program $(\tilde{Q})$ is also a globally optimal one.

Theorem 3.1. *A local minimizer of $(\tilde{Q})$ is a globally optimal solution.*

Proof. A local minimizer of $(\tilde{Q})$ is a global minimizer of $(\tilde{Q})$ if the objective function is convex and the feasible region is a convex set. The feasible region is defined by linear inequalities and is a polyhedron (thus, it is also a *convex set*). Further, we prove that the objective function $f(x,\nu_1,\nu_2)$ is convex with respect to x,ν_1,ν_2.

The first-order partial derivatives of $f(x,\nu_1,\nu_2)$ are

$$\frac{\partial f}{\partial x}=2x+2(t-d_{n-1,s}-H_s)+2x(\lambda_skt_b+1)^2-2(\lambda_skt_b+1)\Big[\tilde{a}_{n+1,s}+\tilde{\beta}_{n+1}t_a+\big(\tilde{\beta}_{n+1}t_a\lambda_s+\nu_1+(\tilde{a}_{n+1,s}-t)\lambda_s\big)kt_b-\nu_2t_b-t-H_s\Big]$$

$$\frac{\partial f}{\partial \nu_1}=2k^2t_b^2\nu_1+2kt_b\Big[\tilde{a}_{n+1,s}+\tilde{\beta}_{n+1}t_a+\big(\tilde{\beta}_{n+1}t_a\lambda_s+(\tilde{a}_{n+1,s}-t-x)\lambda_s\big)kt_b-\nu_2t_b-t-x-H_s\Big]+M_1$$

$$\frac{\partial f}{\partial \nu_2} = 2t_b^2\nu_2 - 2t_b\Big[\tilde{a}_{n+1,s} + \tilde{\beta}_{n+1}t_a + \big(\tilde{\beta}_{n+1}t_a\lambda_s + \nu_1 + (\tilde{a}_{n+1,s} - t - x)\lambda_s\big)kt_b - t - x - H_s\Big] + M_2$$

Therefore, the *Hessian* matrix of f reads:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial \nu_1} & \frac{\partial^2 f}{\partial x \partial \nu_2} \\ \frac{\partial^2 f}{\partial \nu_1 \partial x} & \frac{\partial^2 f}{\partial \nu_1^2} & \frac{\partial^2 f}{\partial \nu_1 \partial \nu_2} \\ \frac{\partial^2 f}{\partial \nu_2 \partial x} & \frac{\partial^2 f}{\partial \nu_2 \partial \nu_1} & \frac{\partial^2 f}{\partial \nu_2^2} \end{bmatrix} = \begin{bmatrix} 2+2(\lambda_s kt_b+1)^2 & -2(\lambda_s kt_b+1)kt_b & 2(\lambda_s kt_b+1)t_b \\ -2(\lambda_s kt_b+1)kt_b & 2k^2t_b^2 & -2kt_b^2 \\ 2(\lambda_s kt_b+1)t_b & -2kt_b^2 & 2t_b^2 \end{bmatrix}$$

To prove the convexity of f, we should prove that the Hessian matrix, $\mathbf{H}$, with elements $H_{ij} \in \mathbf{H}$, is positive semi-definite (P.S.D.). That is, all the leading principal minors are non-negative:

$$\mathbf{H} \text{ is P.S.D. } \Leftrightarrow H_{11} \geq 0, \begin{vmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{vmatrix} \geq 0, \ \det(\mathbf{H}) \geq 0.$$

In our case, we have $H_{11} = 2 + 2(\lambda_s kt_b + 1)^2 > 0$.

In addition, $\begin{vmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{vmatrix} = (2+2(\lambda_s kt_b+1)^2)2k^2t_b^2 - 4(\lambda_s kt_b+1)^2k^2t_b^2 = 4k^2t_b^2 > 0$.

Furthermore,

$$\begin{aligned} \det(\mathbf{H}) =& (2+2(\lambda_s kt_b+1)^2)\begin{vmatrix} H_{22} & H_{23} \\ H_{32} & H_{33} \end{vmatrix} \\ &+ 2(\lambda_s kt_b+1)kt_b\begin{vmatrix} H_{21} & H_{23} \\ H_{31} & H_{33} \end{vmatrix} \\ &+ 2(\lambda_s kt_b+1)t_b\begin{vmatrix} H_{21} & H_{22} \\ H_{31} & H_{32} \end{vmatrix} \\ =& (2+2(\lambda_s kt_b+1)^2)\cdot 0 + 2(\lambda_s kt_b+1)kt_b \cdot 0 \\ &+ 2(\lambda_s kt_b+1)t_b \cdot 0 = 0. \end{aligned}$$

Thus, f is convex and this completes our proof. We finally note that for *strict* convexity, $\det(\mathbf{H})$ should have been greater than zero. Since this is not the case, we might have more than one globally optimal solutions. □

Demonstration

In this sub-section, we perform a small demonstration to show that the non-smooth nonlinear program (Q) and the reformulated one, $(\tilde{Q})$, attain the same solution. In our demonstration, we use an idealized scenario and report the solutions of both programs. In our idealized scenario, trip n arrives at control point stop s and completes its boardings/alightings at time $t = 1500$s. The parameters of our scenario are presented in Table 1.

TABLE 1 : Parameter values of the idealized scenario

Parameter	Value	Unit	Parameter	Value	Unit
$d_{i-1,s}$	1000	s	t_a	1.5	s
t	1500	s	t_b	4	s
H_s	600	s	$a_{n+1,s}$	2500	s
ϕ_n	40	passengers	ζ	300	s
c_n, c_{n+1}	60	passengers	M_1	10E+14	-
$\tilde{\beta}_{n+1}$	10	passengers	M_2	10E+12	-
$\tilde{l}_{n+1}$	50	passengers	λ_s	0.02	passengers / s

As previously discussed, problem (Q) cannot be solved to global optimality due to the non-smooth terms that yield an objective function which is not differentiable at every point in its domain. To find an *approximate solution* of (Q), we discretize the decision variable, x, and evaluate the performance of the objective function for every value of x using simple enumeration (brute-force). In this discretization, the holding time x is discretized into seconds and $x \in \mathbb{Z} \mid 0 \leq x \leq \zeta$.

Using brute-force, we evaluate the objective function f in program (Q) expressed in Eq.(6) and we plot its value for every $x \in \mathbb{Z} \mid 0 \leq x \leq \zeta$. The results are plotted in Fig.2. From Fig.2 it is evident that the *approximate* solution of (Q) is $x^* = 296$ s. Note that the function f in program (Q) is not smooth. In more detail, for $x \in [0, 89]$ the capacity limit of the following bus trip, $n+1$, is exceeded and this leads to stranded passengers by trip $n+1$. For a holding time of trip n in the range of $90 < x \leq \zeta$, the capacity limit of $n+1$ at stop s is not reached. Hence, the objective function of program (Q) has a non-smooth transition at $x = 89$ s.

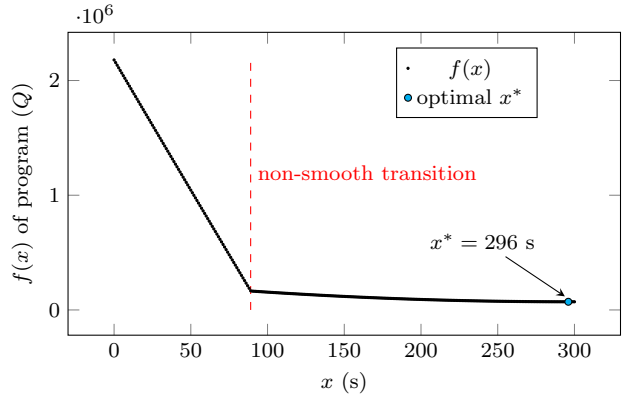

FIGURE 2 : Performance of the objective function f of the discretized program (Q) for every $x \in \mathbb{Z} \mid 0 \leq x \leq \zeta$.

The non-smoothness of (Q) is avoided with our reformulated program $(\tilde{Q})$ that can be solved to *global optimality* with a solution method for quadratic programming. To compute a globally optimal solution of $(\tilde{Q})$, we solve our mathematical program $\tilde{Q}$ in a general-purpose computer with Intel Core i7-455 7700HQ CPU @ 2.80GHz and 16 GB RAM using CPLEX 12.8. The obtained solution is:

$$(v_1, v_2, x) = (0, 0, 296.35 \text{ s})$$

As expected, the globally optimal solution of our reformulated program $(\tilde{Q})$, $x = 296.35$ s, is almost equivalent to the solution of the discretized original program (Q), $x = 296$ s, demonstrating the validity of our reformulated program $(\tilde{Q})$.

NUMERICAL EXPERIMENTS

This section aims to demonstrate the potential improvement when our solution is adopted instead of typical two-headway-based approaches that do not account for the capacity limitations of vehicles. Several experiments are conducted in idealized scenarios to show how our approach mitigates overcrowding while improving the service regularity. Our proposed model is then applied to bus line 302 in Singapore where we investigate the potential effect of holding a bus at one stop.

Performance in idealized scenarios

To demonstrate the effectiveness of our control logic against classic control logic(s) that do not cater for the bus crowding levels, we determine the holding times for each one of the idealized scenarios in Table 2 with (i) our model, and (ii) the classic two-headway-based approach of Fu and Yang (*25*) expressed in Alg.1. The holding times from Alg.1 are presented in the last column of Table 2.

TABLE 2 : Optimal Holding decisions for idealized scenarios with different values of (λ_s, ϕ_n)

scenarios			Solving $\tilde{Q}$ with CPLEX			Solution of Alg.1
	λ_s	ϕ_n	v_1	v_2	x^*	x^*
I	0.02	40	0	0	**296 s**	199 s
II	0.002	40	0	0	**261 s**	181 s
III	0.02	58	0	0	**100 s**	199 s
IV	0.02	55	0	0	**250 s**	199 s
V	0.05	58	0	38.5	**40 s**	229 s
VI	0.02	59	0	0.84	**50 s**	199 s
VII	0.05	40	0	16.9	**300 s**	229 s
VIII	0.02	62	2.00	1.92	**0 s**	199 s

As demonstrated in Table 2, the solution of Fu and Yang (*25*) is not sensitive to the changes of parameter values, ϕ_n, since it does not cater for overcrowding, but merely balances the headways between the preceding and following trip(s) using an estimate of $\tilde{d}_{n+1,s} \approx \tilde{a}_{n+1,s} + \tilde{\beta}_{n+1} t_a + (\tilde{a}_{n+1,s} - t)\lambda_s t_b$.

The results of the comparative analysis between our approach and the classic two-headway-based approach of Fu and Yang (*25*), which is used as a benchmark, are summarized in Fig.3. Fig.3 demonstrates the potential benefit of our control method in comparison to similar approaches that ignore the overcrowding of buses in the optimization process. In the left and right sub-figures of Fig.3 we plot the bus load of trips n and $n+1$ when they depart from stop s for each one of the 8 scenarios in Table 2. From the left sub-figure, our holding solution leads to stranded passengers only in scenario VIII, in which 2 passengers were already waiting for trip n when it arrived at stop s. In contrast, the control logic of Fu and Yang (*25*) results in refused boardings with regards to trip n in 4 cases: III, V, VI, and VIII. Regarding trip $n+1$ (right sub-figure), both control logic(s) result in refused boardings in 3 cases: V, VII, and VIII. The reason is that bus holding cannot reduce the passenger demand that affects those 3 cases; thus, extra measures are needed. This validates

our infeasibility claim in Lemma 2.1 which proved that holding, as a standalone measure, cannot guarantee the accommodation of all passengers.

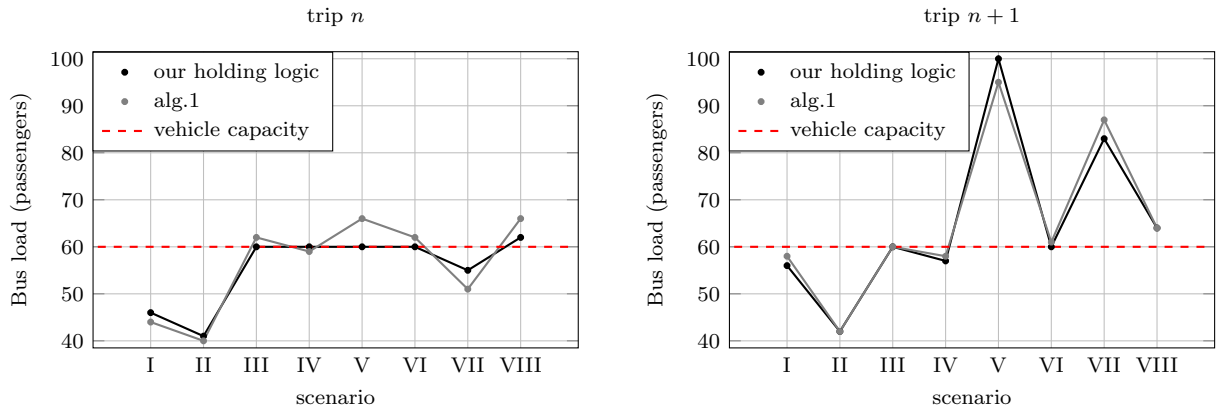

FIGURE 3 : Bus load of trips n and $n+1$ when they depart from stop s in every scenario with the implementation of our model and the one of Fu and Yang (*25*).

Case Study

Our case study is the high-frequency, circular bus line 302 in Singapore. Bus line 302 has 22 stops departing from Choa Chu Kang Loop - Choa Chu Kang Int (44009) and ending at the same stop. It is operated by SMRT and its regularity is monitored by the Land Transport Authority (LTA). Normally starts operating at 05:30 and ends at 00:55. Its route length is 8.1 km and its total travel time typically ranges from 35 to 40 minutes. Bus line 302 is selected because it is one of the seven high-frequency bus lines in Singapore that are monitored in terms of service regularity and are placed under the Bus Service Reliability Framework (BSRF) from the LTA Leong et al. (*37*). Under the BSRF framework, bus lines that do not maintain their scheduled headways are penalized, whereas well-performing lines receive monetary incentives (up to 3000$ for every 0.1 min improvement in regularity at the end of each month, as of May 2014).

Bus line 302 is a feeder service that serves residential blocks, schools, and public amenities, connecting them to Choa Chu Kang Town Centre and Yew Tee Mass Rapid Transit (MRT) station. Its primary area of service is Choa Chu Kang Neighbourhoods 5 and 6. Typically, in this bus line operate 12-meter single-decker buses with a seated capacity of 42 passengers and standing capacity of 33 passengers (75 passengers in total). High capacity, articulated buses have also been deployed due to high demand from residents. The total number of operating trips per day is 245, and the scheduled (target) headways differ among peak/off-peak hours.

Our experiments focus on the time period 06:30-08:30, which exhibits the highest frequency of 15 trips per hour. In that period operate 33 trips with a scheduled headway of 4 minutes. Bus holding cannot be applied at any stop because: (i) some stops are used from several bus lines and do not have enough space for holding; (ii) holding a bus at every stop will greatly increase the inconvenience of onboard passengers and the total trip travel time Cortés et al., Cats et al. (*34, 38*); and (iii) several bus trips do not serve all stops. For this reason, past works have used a selected group of stops, known as intermediate time point stops (ITPs) (or control points), when holding is applied van Oort et al. (*39*). The LTA has selected two monitoring points for the service regularity

of bus line 302 that can serve as ITPs: Yew Tee Stn (45321) and Opp Blk 666 (45421). The topology of bus line 302 and the two control point stops are presented in Fig.4.

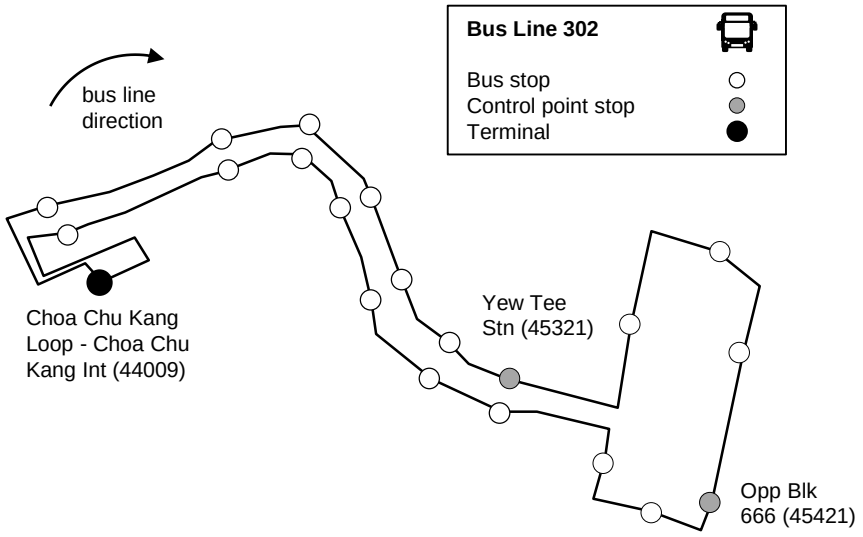

FIGURE 4 : Topology and selected control point stops of bus line 302 in Singapore

In this experimentation, we demonstrate the application of our control logic in one bus trip. This trip is the 2nd trip that operates in the time period 06:30-08:30. To be consistent with our previous notation, that trip is henceforth denoted as n. The holding time of trip n is decided when it is about to depart from stop Yew Tee Stn (45321). Using real data from a weekday, bus trip n departed from stop Yew Tee Stn (45321) at time 06:50. Its preceding trip, $n-1$, departed from the same stop at time 06:48, and its following trip, $n+1$, arrived there at time 06:54. Additionally, the bus load of trip n when it was about to depart from stop Yew Tee Stn (45321) was $\phi_n = 47$ passengers and the bus load of trip $n+1$ at the time of its arrival at stop Yew Tee Stn (45321) was $\tilde{l}_{n+1} = 52$ passengers.

The alighted passengers from trip $n+1$ at stop Yew Tee Stn (45321) were $\tilde{\beta}_{n+1} = 19$ and the boarding passengers 14. Both trips are operated by single decker buses with a total capacity of 75 passengers (including standees). Assuming uniformly distributed passenger arrivals at stop s, the 14 passenger boardings in a time interval of 4 min indicate a passenger arrival rate of $\lambda_s = 3.5$ passengers per minute. This assumption is borrowed from past works which prove that passengers are not able to coordinate their arrival times at stops with the arrival times of buses in high-frequency services Ibarra-Rojas et al. (*31*).

The observed (average) time for an extra passenger boarding and alighting at that stop is 2 and 1 s, respectively. Our observations are in line with the findings of Meng and Qu (*40*) that proposed an extra time of 1.36 s for each boarding/alighting in bus lines in Singapore based on historical data analysis. To summarize, the actual parameter values when holding trip n at stop Yew Tee Stn (45321) are presented in Table 3. Note that, as in Cortés et al. (*34*), we do not allow a holding time of more than 90 s due to the inconvenience caused to on-board passengers.

TABLE 3 : Actual parameter values when determining the holding time of the 2nd trip which is dispatched after 6:30

Parameter	Value	Parameter	Value
$d_{i-1,s}$	6:48 $\rightarrow$ 24480 s	t_a	1 s
t	6:50 $\rightarrow$ 24600 s	t_b	2 s
H_s	4 min $\rightarrow$ 240 s	$\tilde{a}_{n+1,s}$	06:54$\rightarrow$ 24840 s
ϕ_n	47 passengers	ζ	90 s
c_n, c_{n+1}	75 passengers	M_1	10E+14
$\tilde{\beta}_{n+1}$	19 passengers	M_2	10E+12
$\tilde{l}_{n+1}$	52 passengers	λ_s	3.5/60 passengers / s

In the actual operations (do-nothing scenario), bus trip $n+1$ departed from stop Yew Tee Stn (45321) at time $\tilde{d}_{n+1,s} = \tilde{a}_{n+1,s} + \tilde{\beta}_{n+1} t_a + \big(\tilde{\beta}_{n+1} t_a \lambda_s + (\tilde{a}_{n+1,s} - t)\lambda_s\big)\big)(1 + t_b \lambda_s) t_b = 24893$ s. This yielded a squared headway deviation of

$$(t - d_{n-1,s} - H_s)^2 + (\tilde{d}_{n+1,s} - t - H_s)^2 = (-120)^2 + 52.742^2 = 17182\ s^2$$

where the headway between trip n and $n-1$ is 120 s and between trip n and $n+1$ is 293 s.

Now, if one had applied our model, bus trip n would have been held at Yew Tee Stn (45321) for:

$$x = 78.9\ s$$

Our holding time would have yielded $\tilde{d}_{n+1,s} = 24882$ s and a reduced squared headway deviation of 3017 s^2 (82% improvement). The headways among trips $n-1, n, n+1$ and the bus loads of trips n and $n+1$ after departing from Yew Tee Stn (45321) with and without applying our control logic are summarized in Fig.5.

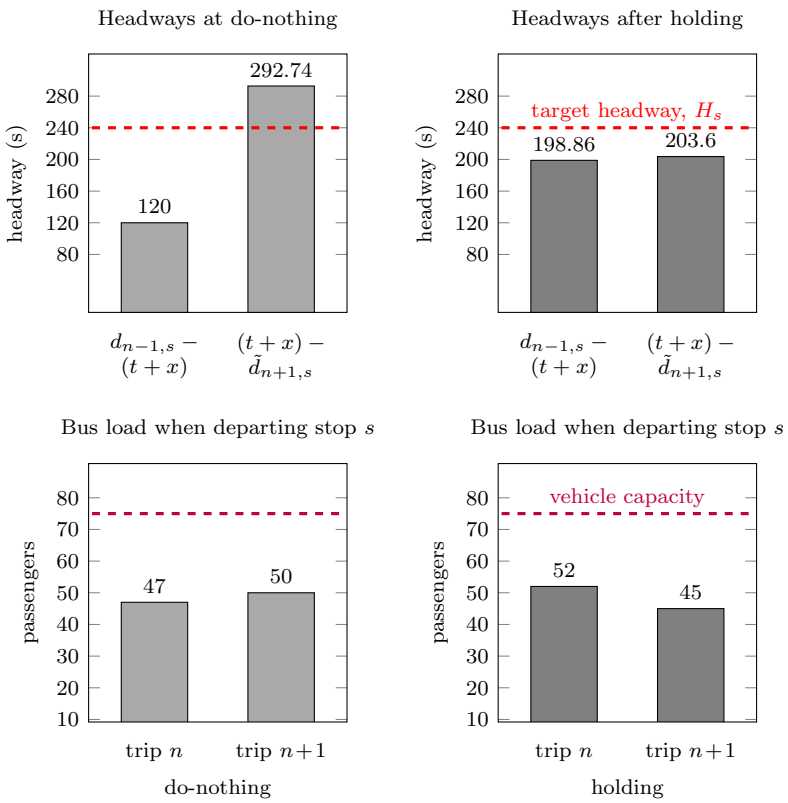

FIGURE 5 : Headways and bus loads at stop s in the do-nothing case (left sub-figures) and in the case where we apply the holding suggested by our model (right sub-figures)

The top two sub-figures in Fig.5 show the headway between trips $n-1$ and n (which is calculated as $d_{n-1,s} - (t+x)$), and trips n and $n+1$ (which is calculated as $(t+x) - \tilde{d}_{n+1,s}$). From those sub-figures, it is evident that the headways among trips $n-1, n$ and $n, n+1$ are more evenly distributed after applying our control logic (values of 198.86 s and 203.6 s, respectively). It is important to note that bus holding cannot guarantee that we can meet the target headway ($H_s = 240$ s), but it can reduce the deviation of headways from that value and provide more evenly distributed headways. This is well-reported in the work of Bartholdi III and Eisenstein (*16*) that started a line of research on equalizing headways, instead of meeting target headway values.

The two sub-figures at the lower part of Fig.5 indicate the bus load of trips n and $n+1$, respectively, when they depart from stop s. Notably, our control logic will hold trip n in stop s resulting in an increased bus load for trip n and a reduced one for trip $n+1$. This action is allowed as long as none of the trips reaches its vehicle capacity limit.

CONCLUSION

This work provided a model, (Q), for real-time bus holding under capacity limitations. The consideration of the bus load and the vehicle capacity limits added another dimension to the traditional bus holding problem, and this resulted in a nonlinear, non-smooth model Q. With the use of slack variables, the nonlinear, non-smooth model Q was transformed into a quadratic program with linear (in)equality constraints. The reformulated program is proved to be convex and have a globally

optimal solution. This easy-to-solve program returned solutions for several idealized scenarios demonstrating the improvement potential in terms of regularity and refused boardings compared to two-headway-based methods that do not consider the capacity and the bus loads in the optimization process.

In the case study of bus line 302 in Singapore, we show that our proposed solution can improve the squared headway deviation by up to 82% compared to the case of no holding. In future research, our approach can be expanded in a wide range of problems involving rail operations. Other advances could be an expansion of our model to incorporate additional constraints related to the timetables and the recommended total trip travel times.

Author Contribution Statement

The authors confirm contribution to the paper as follows: study conception and design: K. Gkiotsalitis, E.C. van Berkum; data collection: K. Gkiotsalitis; analysis and interpretation of results: K. Gkiotsalitis, E.C. van Berkum; draft manuscript preparation: K. Gkiotsalitis. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

[1] Yu, B., Z. Yang, and J. Yao, Genetic algorithm for bus frequency optimization. *Journal of Transportation Engineering*, Vol. 136, No. 6, 2009, pp. 576–583.

[2] Gkiotsalitis, K. and O. Cats, Reliable frequency determination: Incorporating information on service uncertainty when setting dispatching headways. *Transportation Research Part C: Emerging Technologies*, Vol. 88, 2018, pp. 187–207.

[3] Sun, D. J., Y. Xu, and Z.-R. Peng, Timetable optimization for single bus line based on hybrid vehicle size model. *Journal of Traffic and Transportation Engineering (English Edition)*, Vol. 2, No. 3, 2015, pp. 179–186.

[4] Wu, Y., H. Yang, J. Tang, and Y. Yu, Multi-objective re-synchronizing of bus timetable: Model, complexity and solution. *Transportation Research Part C: Emerging Technologies*, Vol. 67, 2016, pp. 149–168.

[5] Wren, A. and J.-M. Rousseau, Bus driver schedulingan overview. In *Computer-aided transit scheduling*, Springer, 1995, pp. 173–187.

[6] Gintner, V., N. Kliewer, and L. Suhl, Solving large multiple-depot multiple-vehicle-type bus scheduling problems in practice. *OR Spectrum*, Vol. 27, No. 4, 2005, pp. 507–523.

[7] Kliewer, N., T. Mellouli, and L. Suhl, A time–space network based exact optimization model for multi-depot bus scheduling. *European journal of operational research*, Vol. 175, No. 3, 2006, pp. 1616–1627.

[8] Ceder, A., *Public transit planning and operation: Modeling, practice and behavior*. CRC press, 2007.

[9] Trompet, M., X. Liu, and D. Graham, Development of key performance indicator to compare regularity of service between urban bus operators. *Transportation Research Record: Journal of the Transportation Research Board*, , No. 2216, 2011, pp. 33–41.

[10] Chen, X., L. Yu, Y. Zhang, and J. Guo, Analyzing urban bus service reliability at the stop, route, and network levels. *Transportation research part A: policy and practice*, Vol. 43, No. 8, 2009, pp. 722–734.

[11] Daganzo, C. F., A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. *Transportation Research Part B: Methodological*, Vol. 43, No. 10, 2009, pp. 913–921.

[12] Knoppers, P. and T. Muller, Optimized transfer opportunities in public transport. *Transportation Science*, Vol. 29, No. 1, 1995, pp. 101–105.

[13] Berrebi, S. J., E. Hans, N. Chiabaut, J. A. Laval, L. Leclercq, and K. E. Watkins, Comparing bus holding methods with and without real-time predictions. *Transportation Research Part C: Emerging Technologies*, Vol. 87, 2018, pp. 197–211.

[14] Gkiotsalitis, K. and N. Maslekar, Multiconstrained timetable optimization and performance evaluation in the presence of travel time noise. *Journal of Transportation Engineering, Part A: Systems*, Vol. 144, No. 9, 2018, p. 04018058.

[15] Hans, E., N. Chiabaut, L. Leclercq, and R. L. Bertini, Real-time bus route state forecasting using particle filter and mesoscopic modeling. *Transportation Research Part C: Emerging Technologies*, Vol. 61, 2015, pp. 121–140.

[16] Bartholdi III, J. J. and D. D. Eisenstein, A self-coördinating bus route to resist bus bunching. *Transportation Research Part B: Methodological*, Vol. 46, No. 4, 2012, pp. 481–491.

[17] Delgado, F., J. C. Munoz, and R. Giesen, How much can holding and/or limiting boarding improve transit performance? *Transportation Research Part B: Methodological*, Vol. 46, No. 9, 2012, pp. 1202–1217.

[18] Liu, Z., Y. Yan, X. Qu, and Y. Zhang, Bus stop-skipping scheme with random travel time. *Transportation Research Part C: Emerging Technologies*, Vol. 35, 2013, pp. 46–56.

[19] Chen, X., B. Hellinga, C. Chang, and L. Fu, Optimization of headways with stop-skipping control: a case study of bus rapid transit system. *Journal of advanced transportation*, Vol. 49, No. 3, 2015, pp. 385–401.

[20] Cortés, C. E., S. Jara-Díaz, and A. Tirachini, Integrating short turning and deadheading in the optimization of transit services. *Transportation Research Part A: Policy and Practice*, Vol. 45, No. 5, 2011, pp. 419–434.

[21] Gkiotsalitis, K., Z. Wu, and O. Cats, A cost-minimization model for bus fleet allocation featuring the tactical generation of short-turning and interlining options. *Transportation Research Part C: Emerging Technologies*, Vol. 98, 2019, pp. 14–36.

[22] Gkiotsalitis, K. and A. Stathopoulos, Demand-responsive public transportation re-scheduling for adjusting to the joint leisure activity demand. *International Journal of Transportation Science and Technology*, Vol. 5, No. 2, 2016, pp. 68–82.

[23] Daganzo, C. F. and J. Pilachowski, Reducing bunching with bus-to-bus cooperation. *Transportation Research Part B: Methodological*, Vol. 45, No. 1, 2011, pp. 267–277.

[24] Muñoz, J. C., C. E. Cortés, R. Giesen, D. Sáez, F. Delgado, F. Valencia, and A. Cipriano, Comparison of dynamic control strategies for transit operations. *Transportation Research Part C: Emerging Technologies*, Vol. 28, 2013, pp. 101–113.

[25] Fu, L. and X. Yang, Design and implementation of bus-holding control strategies with real-time information. *Transportation Research Record: Journal of the Transportation Research Board*, , No. 1791, 2002, pp. 6–12.

[26] Gkiotsalitis, K. and O. Cats, Multi-constrained Bus Holding Control in Time Windows with Branch and Bound and Alternating Minimization. *Transportmetrica B: Transport Dynamics*, Vol. 7, 2019, pp. 1258–1285.

[27] Eberlein, X. J., N. H. Wilson, and D. Bernstein, The holding problem with real–time information available. *Transportation science*, Vol. 35, No. 1, 2001, pp. 1–18.

[28] Hickman, M. D., An analytic stochastic model for the transit vehicle holding problem. *Transportation Science*, Vol. 35, No. 3, 2001, pp. 215–237.

[29] Sánchez-Martínez, G., H. Koutsopoulos, and N. Wilson, Real-time holding control for high-frequency transit with dynamics. *Transportation Research Part B: Methodological*, Vol. 83, 2016, pp. 1–19.

[30] Nikolaou, M., Model predictive controllers: A critical synthesis of theory and industrial needs. *Advances in Chemical Engineering*, Vol. 26, 2001, pp. 131–204.

[31] Ibarra-Rojas, O., F. Delgado, R. Giesen, and J. Muñoz, Planning, operation, and control of bus transport systems: A literature review. *Transportation Research Part B: Methodological*, Vol. 77, 2015, pp. 38–75.

[32] Delgado, F., J. C. Munoz, R. Giesen, and A. Cipriano, Real-time control of buses in a transit corridor based on vehicle holding and boarding limits. *Transportation Research Record*, Vol. 2090, No. 1, 2009, pp. 59–67.

[33] Berrebi, S. J., K. E. Watkins, and J. A. Laval, A real-time bus dispatching policy to minimize passenger wait on a high frequency route. *Transportation Research Part B: Methodological*, Vol. 81, 2015, pp. 377–389.

[34] Cortés, C. E., D. Sáez, F. Milla, A. Núñez, and M. Riquelme, Hybrid predictive control for real-time optimization of public transport systems operations based on evolutionary multi-objective optimization. *Transportation Research Part C: Emerging Technologies*, Vol. 18, No. 5, 2010, pp. 757–769.

[35] Newell, G. F., Control of pairing of vehicles on a public transportation route, two vehicles, one control point. *Transportation Science*, Vol. 8, No. 3, 1974, pp. 248–264.

[36] Marguier, P., *Bus route performance evaluation under stochastic conditions*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1985.

[37] Leong, W., K. Goh, S. Hess, and P. Murphy, Improving bus service reliability: The Singapore experience. *Research in Transportation Economics*, Vol. 59, 2016, pp. 40–49.

[38] Cats, O., A. Larijani, H. Koutsopoulos, and W. Burghout, Impacts of holding control strategies on transit performance: Bus simulation model analysis. *Transportation Research Record: Journal of the Transportation Research Board*, , No. 2216, 2011, pp. 51–58.

[39] van Oort, N., J. Boterman, and R. van Nes, The impact of scheduling on service reliability: trip-time determination and holding points in long-headway services. *Public Transport*, Vol. 4, No. 1, 2012, pp. 39–56.

[40] Meng, Q. and X. Qu, Bus dwell time estimation at bus bays: A probabilistic approach. *Transportation Research Part C: Emerging Technologies*, Vol. 36, 2013, pp. 61–71.