Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2017) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

23rd International Symposium on Transportation and Traffic Theory, ISTTT 23, 24-26 July 2019, Lausanne, Switzerland

Operational design for shuttle systems with modular vehicles under oversaturated traffic: Continuous modeling method

Zhiwei Chen[a], Xiaopeng Li[a,*], Xuesong Zhou[b]

[a] Department of Civil and Environmental Engineering, University of South Florida, 4202 E Fowler Avenue, ENC 3300, Tampa, FL 33620
[b] School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ 85287, United States

Abstract

Time-varying capacity design holds an opportunity to reduce the energy consumption of urban mass transit systems, e.g., urban rail transits, bus rapid transits, modular autonomous vehicles. In this paper, we investigate the joint design of dispatch headway and vehicle capacity for one to one shuttle systems with oversaturated traffic to achieve the optimal tradeoff between general vehicle dispatching cost (mainly comprised of vehicle energy consumption) and customer waiting cost. We propose a continuum approximation model from a macroscopic point of view to reveal fundamental analytical insights into the optimal design. By introducing the concept of a virtual arrival demand curve at the origin station, we prove that the investigated problem with possibly oversaturated traffic can be equivalently solved with a simpler revised problem where only unsaturated traffic is present. With this property, we decompose the original problem into a set of independent unit-time revised unsaturated problems that can be analytically solved in each neighborhood across the operational horizon. With two sets of numerical experiments, we show that the CA model offers near-optimum solutions with negligible errors very efficiently and we also verify the theoretical properties. Also, the effectiveness of time-varying vehicle capacity design is demonstrated in shuttle systems under both saturated and unsaturated traffic. Overall, the proposed CA model contributes to the CA methodology literature by extending the CA method for traditional transit dispatching problems with unsaturated traffic to the joint design of dispatch headway and vehicle capacity considering oversaturated traffic, adjustable vehicle capacities and other factors (e.g. minimum dispatch headway).

Peer-review under responsibility of the scientific committee of the 23rd International Symposium on Transportation and Traffic Theory.

Keywords: Vehicle scheduling; time-varying capacity design; energy consumption; oversaturated traffic; continuum approximation

[*] Corresponding author. Tel: +1(813) 973-0778; Fax: +1(813)974-2957

Peer-review under responsibility of the scientific committee of the 23rd International Symposium on Transportation and Traffic Theory.

Nomenclature

Parameters	
$A(t)$	Cumulative arrival demand up to t , $\forall t \in [0, T]$
$a(t)$	Arrival demand rate at t, $a(t) := A'(t), \forall t \in [0, T]$
$B(t)$	Virtual cumulative arrival demand up to t , $\forall t \in [0, T]$
C^F	Fixed vehicle energy cost
C^V	Coefficient for capacity-dependent vehicle energy cost
c	Capacity per vehicle unit
f_i	Energy cost of a vehicle in formation i, $\forall i \in \mathcal{I}$
$\underline{h}$	Minimum dispatch headway
$\mathcal{I} := [1,2,\cdots,I]$	Set of vehicle formations
i	Index of vehicle formations, $\forall i \in \mathcal{I}$
T	Time horizon $[0, T]$
t	Time index, $\forall t \in T$
w	Unit time waiting cost per passenger
α	Unitless parameter,$\alpha \leq 1$
Decision variables	
$D(t)$	Cumulative departure up to t, $\forall t \in [0, T]$
d_k	Number of passengers boarding at dispatch $k, \forall k \in \mathcal{K}$
i_k	Formation of vehicle for dispatch $k, \forall k \in \mathcal{K}$
K	Number of dispatches
$\mathcal{K} := [1,2,\cdots,K]$	Set of dispatch indexes
t_k	Time for dispatch $k, \forall k \in \mathcal{K}$

1. Introduction

Temporal demand fluctuations have been observed in urban mass transit systems (UMTS) in many big cities (e.g. Beijing, Los Angeles, and Paris), where passenger arrival rates during peak hours are much more intensive than those during off-peak hours. Such time-variant demand patterns considerably lower UMTS's service quality because of excessive waiting time and overcrowded vehicles during oversaturated periods. For instance, the passenger load in Shenzhen Metro System (http://www.szmc.net/) is usually over 100% during peak hours, so passengers cannot board on a full train and have to wait for another. On the other hand, the demand fluctuations lead to a noticeable energy waste due to a mismatch between the vehicle capacity and passenger load. In many existing systems, the length of a vehicle is designed to just meet the average peak demand over an hour or more, and its operation may be quite rigid to adapt to off-peak periods. With this design, many seats will be empty during off-peak hours, and much energy is just wasted in hauling vehicle units with low occupancy. For example, the Shenzhen Subway System consumed 7.43 billion kWh electricity in 2016, 88% of which was attributed to hauling train units with passenger loads less than 20% (Jian, 2017).

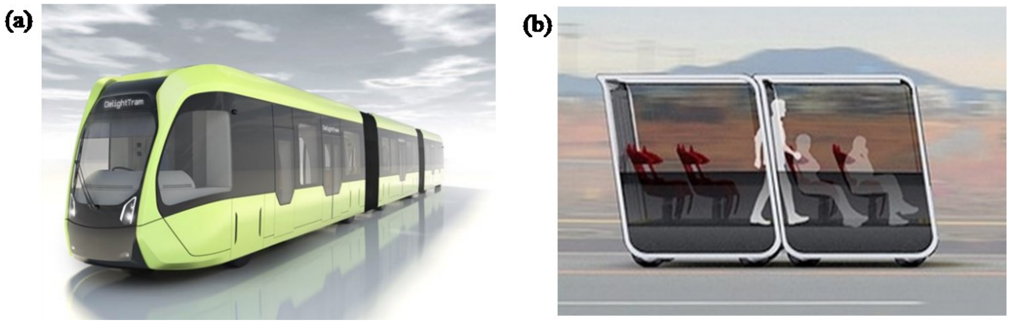

Fig. 1: (a) Autonomous rail rapid transits (source: https://designmuseum.org/) and (b) Modular autonomous vehicles (source: http://www.next-future-mobility.com/).

This observation actually reveals an opportunity for significantly reducing UMTS energy consumption as well as improving the service quality by integrating vehicle scheduling with time-varying capacity design. Indeed, various advanced transportation technologies inspired by this concept have been proposed and tested recently. As is illustrated in Fig. 1(a), pilot experiments on autonomous rail rapid transit with adjustable train lengths are going on in Zhuzhou, China (Lambert, 2017) and flexible grouped electric multiple units (EMU) have been tested in China in 2017 (Rail, 2017). Moreover, modular autonomous vehicles featured with time-varying vehicle capacity (Fig. 1(b)) designed by the Next Future Transportation Inc. are being tested in Dubai (Tarek, 2018) and Singapore (Ackerman, 2016). In these emerging technologies, a modular vehicle is composed of multiple identical modules (or units) that can be dissembled and assembled dynamically either on standard city roads or terminals. Thus, ideal operations shall be able to not only adjust vehicle dispatch headways but also design different vehicle lengths (or capacities) across different dispatches to accommodate time-varying travel demand with minimum energy.

Indeed, increasing research efforts have been seen in recent years to improve the service quality and (or) reduce the energy consumption of UMTS's through demand-driven transit scheduling or timetable design (Yin et al., 2017) but only a handful of them has considered the option of time-varying vehicle capacity (e.g. Albrecht, 2009; Hassold and Ceder, 2012; Hassold and Ceder, 2014; Guo et al., 2014). One approach to solving the transit scheduling problem is discrete modeling, which usually builds models with a discretised time representation and then numerically solves the models (Lin and Kwan, 2016; Niu and Zhou, 2013; Niu et al., 2015; Wang et al., 2015; Xu et al, 2017). Although the discrete models can yield exact solutions, it still takes enormous computational efforts to solve them to optimality even with advanced solution algorithms (Niu and Zhou, 2013; Sun et al., 2014; Zhou and Teng, 2016), which hinders their applications in large-scale instances in the real world. Further, the discrete models cannot provide analytical results on how different demand levels affect the dispatch policy in an UMTS system, which may hinder our understanding of the problem properties and managerial insights (Anasari et al., 2017). See the Chen et al. (2018) for a comprehensive review on this topic.

The other modeling approach, as the focus of this paper, is the classical continuum approximation (CA) model proposed by Newell (1971) for the first time to design the optimal dispatch headways for a transportation route with unsaturated traffic. Later, the CA approach has been extended to consider more realistic constraints, e.g. round-trip constraints and fleet size (Salzborn, 1972; Hurdle, 1973a; Hurdle, 1973b), capacitated vehicles (Sheffi and Sugiyama, M., 1982), bus bunching (Barnett and Kleitman, 1973; Newell, 1974; Daganzo, 2009), multiple periods (Chang and Schonfeld, 1991), transit line spacing (Hurdle, 1973c), interchange (Salzborn, 1980), many-to-many time-varying demand (Wirasinghe, 1990), skip-stop operations (Freyss et al., 2013), etc. Recent studies have applied the CA models to more complicated transit network design problems where the dispatch headway serves as a decision variable (Daganzo, 2010; Estrada et al., 2011; Ouyang et al., 2014; Chen et al., 2015; Fan et al., 2018). See Anasari et al. (2017) for a recent review on the CA methods. Despite these abundant advancements in CA methods, few of them can be directly applied to the investigated problem in this study.

The first challenge is the consideration of oversaturated traffic. One fundamental property for CA to be applicable is that decisions in a local time or space neighborhood have relatively minor impacts on other neighborhoods and the impact deteriorates over the distance. This way, the problem can be approximately decomposed across all neighborhoods and CA can easily solve each neighborhood's decisions with homogenous approximation. This "local-impact" property, however, is not satisfied in the investigated UMTS problem due to oversaturated traffic. Under oversaturated traffic, passengers may not be able to all board the first vehicle reaching their station at its departure time, and thus the passenger queue is built up when this happens. The queuing mechanism actually couples decisions in relatively distant (time) neighborhoods since the current dispatch decision may significantly affect the queuing state and thus the corresponding dispatch decision in a relatively distant future. For example, at time 0, the queue length is 100 after a dispatch of a vehicle of capacity 50, and the arrival rate is the same as the departure capacity in the following few hours. Then the decision of dispatching this vehicle at time 0 will result in a persistent queue of size 100 throughout these few hours, which will cut into the waiting costs not only at time 0 but also for the rest of the few hours. However, if we change the decision of the dispatch at time 0 to a larger vehicle of capacity 100, then the queue length for the remaining few hours will be reduced to 50, which consequentially reduces the waiting cost by a half even in neighborhoods hours away. With this example, we know that the investigated problem does not satisfy the local impact premise and thus, the classic CA methods are not applicable to it.

Though has been rarely investigated under the CA framework, time-dependent queue behavior under oversaturation circumstances of different queuing systems (e.g., intersections, transits, airlines) has been extensively discussed in the seminal work of queueing theory by Newell (1982). Related discussion can also be found in Daganzo (1997). Approximate formulas for performance measures in queueing systems are derived in this book, which can be applied to analyzing any queuing systems (possibly with certain mathematical tricks). Particularly, the deterministic approximation to the total oversaturation delay in Newell's book (see Equation 2.8 in Page 16) is highly related to the investigated problem. Yet to apply this equation to the investigated problem, two problems need to be addressed. First, the approximation is achieved based on the assumption that the cumulative departure curve is rather close to the cumulative arrival curve during oversaturation periods, which might not be true in a heavily congested transit system. Taking a step back, even if this assumption holds, optimization is still needed to determine the optimal service rate in this equation. Since constant continuous departure rates is considered to derive this equation, delays only appear when the system is oversaturated. Therefore, one can easily tell from this equation that the largest service rate should be selected to minimize the (oversaturation) delay in the system. Note that this might not be true if operation cost is taken into cost. Nevertheless, in the investigated problem, passengers are served in batches, which brings passenger delays even when the system is in unsaturated periods. Further, instead of loss time from switching approaches that is almost independent of intersection capacity, the investigated problem also considers vehicle dispatching cost that is mainly comprised of energy consumption depending on the length (or capacity) of the dispatched vehicle. Thus, we have to incorporate the unsaturation delay and operation cost to Newell's approximate equation to obtain the objective function for optimization, to solve which is akin to solve the original problem. Nevertheless, the analysis methods presented in these books, i.e. cumulative plots and fluid approximations, lay a methodological foundation to analyzing the oversaturated queuing behavior in this paper.

In addition to oversaturated traffic, the second challenge in the investigated study is the consideration of time-varying (or variable) vehicle capacity since a vehicle can be assembled with different numbers of modular vehicle units. This will make the vehicle dispatch cost a variable related to the capacities of vehicle dispatched. As far as the authors' knowledge, no CA study has considered such time-varying vehicle capacity and the associated energy cost implications in the vehicle scheduling context. Thus, there still lacks a methodology that can efficiently solve the joint design problem in real-world settings or offer simple analytical insights into the fundamental problem structure.

To bridge these methodological gaps, this paper proposes a CA model for the joint design of dispatch headway and capacity for a one-to-one shuttle system under oversaturated traffic, which has been frequently observed in real UMTS traffic (Niu and Zhou, 2013). The major challenge of this problem is to deal with a series of passenger queue variables across all time points, and each queue variable has a large feasible region that is unbounded as the time elapses due to the consideration of oversaturated demand. The oversaturated traffic even invalidates the fundamental CA premise that local decisions mainly have impacts on the corresponding local neighborhoods. Plus, the multiple capacity options of a dispatch further complicate the dynamics of the queue variables. As a result, this problem is computationally intractable in its original form. To overcome these challenges, we introduce a *virtual arrival demand curve* at the origin station to decompose the oversaturated traffic into a constant term and a revised problem where the arrival demand is always unsaturated throughout the operational horizon. Due to the removal of the excessive queue, the revised unsaturated problem regains the local-impact property enabling the CA concept and thus can be approximately analytically solved across all neighborhoods independently with the CA framework. This also indicates that the computational complexity of the CA model is only linear to the number of the time discretization intervals. In addition to solution efficiency, we have verified that the obtained analytical solutions can produce highly accurate near-optimum designs to the original problem. Overall, this paper makes contributions to the literature from the following three aspects:

(i) We for the first time formulate the joint design problem into a CA model that presents a macroscopic view of the system and yields simple analytical rules into the optimal design. These analytical results will enable efficient solution methods for relevant large-scale transportation problems and offer managerial insights to system operators.

(ii) We prove some elegant theoretical properties of the investigated problem, which show that the original problem can actually be solved by just solving its corresponding revised unsaturated problem. With this, the proposed CA model extends the CA methodology literature from traditional transit dispatching problem with unsaturated traffic to the joint design of dispatch headway and vehicle capacity considering oversaturated traffic, adjustable vehicle capacities and other factors (e.g., minimum dispatch headway).

(iii) Two sets of numerical experiments using real-world data are conducted to verify the validity and efficiency of the proposed solution method and theoretical properties. Results show that the proposed CA model can produce highly accurate near-optimum solutions in almost no time (compared with the exact solutions obtained in Chen et al. (2018)). These results also verify the correctness of the proposed theoretical properties.

Overall, this study advances our knowledge in fundamental mechanisms and critical matters in UMTS operations, creates new opportunities in improving existing engineering practices by incorporating time-varying capacity, and provides both numerical and analytical tools for solving realistic problem instances. We want to note that this paper only focuses on the continuous modeling method. To obtain exact solutions (as well as evaluating the performance of the CA method in both optimality and efficiency), we investigate alternative discrete models with microscopic discrete inputs and propose a customized dynamic programming algorithm that can efficiently solve the problem instances in moderate sizes in Chen et al. (2018).

The remainder of this article is organized as follows. Section 2 introduces the studied joint design problem. Section 3 investigates theoretical properties of the joint design problem. Based on the theoretical properties, Section 4 presents the CA model and the analytical solution approach. Section 5 demonstrates the validity of the proposed CA model and tests its efficiency and the related theoretical properties with two sets of numerical experiments. Finally, Section 6 concludes the paper and briefly discusses the future research topics.

2. Problem statement

This study considers a UMTS shuttle system (e.g. subways, bus rapid transits, autonomous rail transits, modular autonomous vehicles) with one origin and one destination over an operational time horizon $[0, T]$. The time-varying daily passenger demand at the origin is described as a cumulative arrival demand curve $A(t), \forall t \in [0, T]$ as shown in Fig. 2, and thus the arrival demand rate at $t \in [0, T]$ is $a(t) := A'(t)$. A group of vehicles in different formations $\mathcal{I} := [1,2, \cdots, I]$, indexed by $i \in \mathcal{I}$, can be dispatched at the terminal. Each vehicle formation $i \in \mathcal{I}$ has i identical vehicle units and therefore a capacity of ic, where c is the capacity of a vehicle unit. To serve the passenger demand, a number of K vehicles are dispatched at the origin and head to the destination over the operational horizon, and we index these dispatches as $\mathcal{K} := [1,2, \cdots, K]$ where the index increases with the dispatch time. Following Newell (1971), we assume that the availability of vehicles at the origin is independent of when previous vehicles have been dispatched. Therefore, the fleet size at the depot is always sufficient so that there are always some vehicles available in each formation for dispatching at the origin. Please note that the fleet size (i.e., the number of vehicle units) can be determined after solving the investigated operational problem. Therefore, in the planning stage, we can just procure vehicles according to this fleet size (which could be further multiplied by a certain factor for reliability). Since the fleet planning problem is a separate problem in this investigated context, we will focus on vehicle operations assuming that a sufficient number of vehicles are provided. Due to limited resources and operational safety, we assume that the minimum headway between every two consecutive dispatches is $\underline{h}$. These dispatches result in a cumulative departure curve from the origin, denoted by $D(t), \ \forall t \in [0, T]$. The vehicle formation and dispatch time of each dispatch $k \in \mathcal{K}$ are denoted by i_k and t_k, respectively. For the convenience of notation, define $t_0 := 0$ and $t_{K+1} = T$.

Fig. 2: Cumulative passenger counts

Note that since vehicles are only dispatched at discrete time points separated by a minimum interval of $\underline{h}$ while passengers arrive continuously, passengers need to wait before boarding a vehicle. Following previous studies (Niu and Zhou, 2013; Yin et al., 2017), this study adopts the total passenger waiting cost to measure the service quality of a UMTS. Note that in Fig. 2, the time separation between $A(t)$ and $D(t)$ for the same passenger cumulative count denotes the waiting time for this passenger, and thus the total passenger waiting time is just the shaded area between $A(t)$ and $D(t)$. We assume that each passenger has an identical unit-time waiting cost w, and thus the passenger waiting cost is the product of the shaded area in Fig. 2 and w.

Besides the customer waiting cost, the other cost component this study considers is the energy cost (including the traction energy and auxiliary energy) (Huang et al., 2017) consumed by each vehicle from the origin to the destination. The investigated problem only considers the portion of energy associated with a vehicle's mass but not that associated with the number of passengers on board. This is because, first, the energy consumption is actually mainly determined by vehicle mass but not much affected by passengers on board (Zhao et al., 2017). Second, the total transported passengers (i.e., $A(T)$) in this problem is independent of the vehicle dispatching schedule, and thus the associated energy consumption shall not be affected by the dispatch decisions and can be removed from the optimization. Therefore, since the vehicle mass is determined by the vehicle capacity (or the number of vehicle units), we denote the energy cost of a vehicle in formation i as $f_i > 0$ and assume it is concave over i to account for the economies of scale Cohen and Moon (1991) and Holmberg and Tuy (1999), i.e.,

$$\lambda f_i + (1-\lambda) f_j \leq f_k, \forall i, j, k \coloneqq \lambda i + (1-\lambda) j \in \mathcal{I} \cup \{0\}, \lambda \in [0,1], \tag{1}$$

where $f_0 \coloneqq 0$. One common example is that $f_i = C^F + C^V(i)^\alpha, \forall i \in \mathcal{I}$. In this function, the first term C^F is the fixed energy cost regardless of the vehicle capacity (or the number of vehicle units), which accounts for, e.g., the locomotive cost and some auxiliary energy consumption. The second term refers to the variable energy cost attributable to the number of vehicle units. We just use a simple function $C^V(i)^\alpha$ for the variable energy cost, with C^V being a positive coefficient (to account for a single unit cost) and power index $\alpha \leq 1$. Note that this study primarily focuses on an operational problem in UMTS systems while fleet size design is more of a planning level problem. Further, as long as a sufficient number of vehicles are available for operations, the fleet cost, albeit considerable, does not much affect operational decisions. As pointed out in Hurdle (1973a), in a unidirectional shuttle system where vehicles make only one-way trips, it is legitimate to use a model with only two costs, for passenger waiting time and for vehicle operation. Thus, we do not consider the cost of owning the fleet in this paper. Also, this assumption is not restrictive since other cost components related to vehicle dispatch, e.g., capital costs for vehicles, driver costs, crew salaries and vehicle reconfiguration costs, can be easily incorporated into the general cost structure (1) without changing the problem structure. Therefore, for the simplicity of modeling, the investigated problem only considers the energy consumption.

The objective of the joint design problem, then, is to find an optimal arrangement on both the formations of vehicles dispatched, i.e. i_k, and the corresponding dispatch times, i.e. t_k, $\forall k \in \mathcal{K}$ during $[0, T]$. Note that along with i_k and $t_k, \forall k \in \mathcal{K}$, the number of dispatches K and cumulative departure curve $D(t), \forall t \in [0, T]$, which are also decision variables, will be determined. This optimal arrangement aims to achieve the best trade-off between the vehicle energy cost and the passenger waiting cost. The significance of such a trade-off has been well explained in Yin et al. (2017). Moreover, to capture the time-variant passenger process, we denote the vehicle load, i.e., the number of passengers boarding at dispatch k as $d_k, \forall k \in \mathcal{K}$. With these decision variables, we can now formulate the objective function as

$$\min_{K, [t_k, i_k, d_k, \forall k \in \mathcal{K}], \{D(t), \forall t \in [0,T]\}} \sum_{k \in \mathcal{K}} f_{i_k} + w \int_0^T (A(t) - D(t)) dt. \tag{2}$$

This objective function aims to search for the optimal formation of vehicle and time for each dispatch such that the sum of the energy cost and the waiting cost across the operational horizon can be minimized. In addition, to describe the dynamic operation of a UMTS, we consider four groups of constraints as follows.

(i) Minimum headway requirement. These constraints are imposed to ensure the least time separation between two consecutive vehicles due to the safety consideration.

$$t_k - t_{k-1} \geq \underline{h}, \forall k \in \mathcal{K} \backslash \{1\}. \tag{3}$$

(ii) Determination of the vehicle load $d_k, \forall k \in \mathcal{K}$, which is the minimum among the difference between the cumulative arrival demand at k and cumulative departure at dispatch $k-1$ and the capacity of the vehicle dispatched at k.

$$d_k = \min\{A(t_k) - D(t_{k-1}), i_k c\}, \forall k \in \mathcal{K}. \tag{4}$$

(iii) Departure curve conservation. These constraints are imposed to describe the passenger departure dynamics. Constraint (5) is the initialization condition while Constraints (6) describe the departure conservation process. Constraint (7) ensures that all passengers are transported at the end of the given operational horizon. Note that based on Constraint (7), to ensure the feasibility of the investigated problem, the arrival demand pattern in the shuttle system must satisfy $\frac{A(T)}{T} < \frac{cI}{\underline{h}}$. Otherwise passengers cannot be cleared at T even if the maximum transportation capacity is used.

$$D(0) = 0, \tag{5}$$

$$D(t) = \begin{cases} D(t_{k-1}) + d_k, \forall\, t = t_k, k \in \mathcal{K} & ; \\ D(t_k), \forall t \in (t_k, t_{k+1}), k \in \mathcal{K} \cup \{0\}, \end{cases} \tag{6}$$

$$D(t_K) = A(T). \tag{7}$$

(iv) Feasible regions. These constraints define the domains of K, t_k and i_k, respectively.

$$K \in \mathbb{Z}^+; t_k \in [0, T], \forall k \in \mathcal{K}; i_k \in \mathcal{I}, \forall k \in \mathcal{K} \tag{8}$$

Before further analyzing the investigated problem, we use an illustrative example to highlight the significance of incorporating time-varying vehicle capacity design into transit scheduling. We consider an operational horizon with a cumulative arrival curve $A(t)$ as Fig. 3 shows. Two formations of vehicles denoted as $\mathcal{I} := [1,2]$ can be dispatched at the terminal with $c = 5$ passengers/vehicle unit. The cost-related parameters are set as follows: $w = 1, f_1 = 4, f_2 = 7$. If time-varying capacity design is not allowed; i.e., only vehicles in formation 2 (i.e., consisting of 2 units) can be dispatched, then the optimal solution is to dispatch a vehicle in formation 2 at times 2, 3 and 4 respectively, as the blue dotted line in Fig. 3 shows. This results in a total waiting cost of 20.5 and a total energy cost of 21. However, with time-varying capacity design, the optimal solution is to dispatch a vehicle in formation 1 (i.e., consisting of 1 unit) at times 1, 2, 4 and a vehicle in formation 2 at time 3. This solution is represented as the red dashed line in Fig. 3 shows. This strategy results in a total waiting cost of 15.5 and a total energy cost of 19. Thus, time-varying capacity design reduces both the total energy cost and the total waiting cost in the shuttle system.

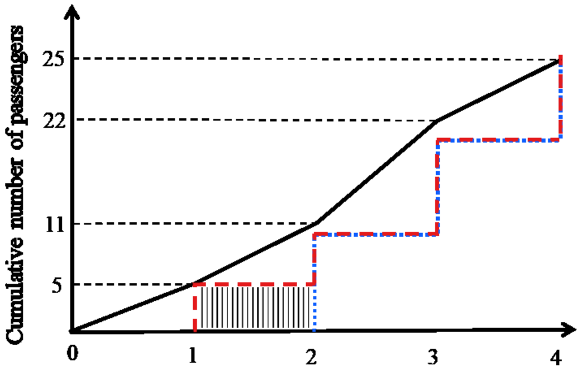

Fig. 3: An illustrative example

3. Theoretical property analysis

This section investigates some theoretical properties of the optimal solution to the above-defined problem, which will serve as the cornerstone for developing efficient algorithms in the following sections.

3.1. Review of Newell's CA model

We first review Newell's classical CA model to illustrate the necessity to explore the theoretical properties of the investigated problem. Note that only unsaturated traffic is considered in Newell's CA model. Because in unsaturated traffic, passengers waiting at the station can be cleared at each dispatch, their waiting cost can be formulated as

$$w \int_0^T (A(t) - D(t))dt = \sum_{k \in \mathcal{K}} \frac{1}{2} w (t_k - t_{k-1})^2 a(t'),$$

where $t' \in [t_{k-1}, t_k)$. Replacing $(t_k - t_{k-1})$ with a continuous function $h(t)$ and $a(t')$ with $a(t)$, respectively, Newell approximated the waiting cost as

$$w \int_0^T (A(t) - D(t))dt \approx \sum_{k \in \mathcal{K}} \int_{t_{k-1}}^{t_k} \frac{w h(t) a(t)}{2} dt.$$

However, this approximation formulation does not apply to cases where oversaturated traffic is present. Since the arrival demand is greater than the maximum vehicle capacity when the system is oversaturated, there might be passengers left behind after a dispatch, which changes the exact formulation of the passenger waiting cost. Consider the simplest situation with a linear $A(t)$ as **Error! Reference source not found.** shows. With the abovementioned formulation, the red shaded area that is related to the queued passengers (or oversaturated traffic) after each dispatch is completely omitted. Indeed, to formulate the waiting cost, we have to introduce a new variable q_k to capture the number of passengers left in the queue after the k-th dispatch as follows:

$$w\int_0^T (A(t)-D(t))dt = \sum_{k\in\mathcal{K}\setminus\{1\}} \left(\frac{1}{2}(t_k - t_{k-1})^2 a(t') + q_k(t_k - t_{k-1})\right).$$

An approximation to the first term in this formulation can be easily obtained from the classical CA model. Nonetheless, the second term is much more difficult to approximate since variable q_k actually couples decisions in relatively distant time neighborhoods. More specifically, the current queuing state q_k may affect future queuing states $\{q_{k'}, \forall k' > k\}$ and the more significant the oversaturated traffic, the stronger and more long-lasting this effect is. This breaks the very fundamental premise of CA: decisions in a local neighborhood do not much affect other distant neighborhoods, or the effect of local decisions attenuates across distance. Thus, the oversaturated traffic poses a great challenge to formulating the investigated problem into the CA model.

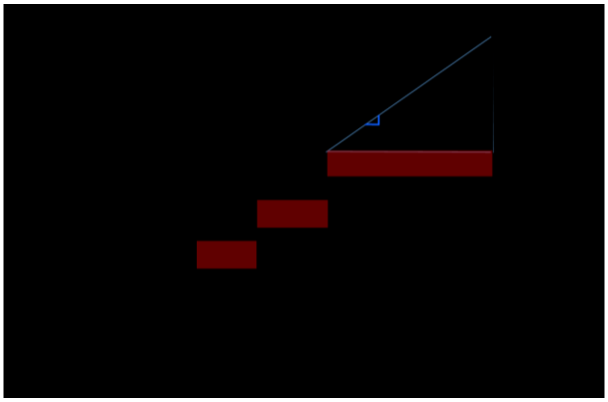

Fig. 4: An illustrative example with linear demand.

As mentioned previously, the consideration of time-varying vehicle capacity also challenges a direct application of the classical CA framework. As can be seen from the above formulation, the passenger waiting cost in Newell's CA model is not dependent on the vehicle capacity since passengers are cleared at each dispatch. However, this does not work when multiple vehicle formations are considered. Obviously, the amount of passengers left behind the station and thus the passenger waiting costs are different if vehicles with different capacities are dispatched. Further, in Newell's model, the vehicle dispatch cost is a given input parameter rather than a decision variable as in the investigated problem. Thus, it cannot be used to search for the optimal formations of the vehicle at each dispatch.

3.2. Theoretical property analysis

This subsection analyzes the theoretical properties to the investigated problems. The proofs of the following lemmas and theorems are available in Appendix A. .

Lemma 1. *For f_i satisfying concave property* (1), *we obtain* $f_{i_1} + f_{i_2} + \cdots + f_{i_n} \geq f_{i_1+i_2+\cdots+i_n}, \forall\, i_1, i_2, \cdots, i_n, i_1 + i_2 + \cdots + i_n \in \mathcal{I}, n \in \mathbb{Z}^+$.

Lemma 2. *For f_i satisfying concave property* (1), *we obtain $f_{i_1} + f_{i_4} \leq f_{i_2} + f_{i_3}, \forall i_1 \leq i_2 \leq i_3 \leq i_4 \in \mathcal{I}$ and $i_2 + i_3 = i_1 + i_4$.*

With this, we will specify some optimal solution properties in the following theorems.

Theorem 1. *In an optimal dispatch solution $\{t_k, i_k\}_{\forall k\in\mathcal{K}}$ to problem* (2) ~ (8) *with arrival curve $A(t)$ and departure curve $D(t)$, if $A(t_k) - D(t_{k-1}) \geq Ic$, then $i_k = I, \forall k \in \mathcal{K}$.*

Theorem 2. *In an optimal dispatch solution* $\{t_k, i_k\}_{\forall k \in \mathcal{K}}$ *to problem* (2) ~ (8) *with arrival curve* $A(t)$ *and departure curve* $D(t)$*, if* $A(t_k) - D(t_{k-1}) > Ic$*, then* $t_k - t_{k-1} = \underline{h}, \forall k \in \mathcal{K}$.

Theorems 1 and 2 indicate that in an optimal solution, the best policy is always to dispatch vehicles in formation I with $\underline{h}$ when the number of passengers who want to board is greater than the longest vehicle's capacity. Note that these findings are well aligned with those in Newell (1982) and Daganzo (1997) that the maximum service rate shall be applied to serve a system under oversaturation. To more rigorously analyze this, for any given arrival curve $A(t)$, we define the oversaturated and unsaturated periods as follows.

Definition 1. For a given arrival curve $A(t)$, the *oversaturated period set* denoted as $\mathcal{T}^S := \{t | t \in [\mathcal{O}(z), \mathcal{E}(z)), \forall z \in [1, |\mathcal{O}|]\}$ and the *unsaturated period set* as $\mathcal{T}^U := [0, T] \backslash \mathcal{T}^S$, where $\mathcal{O}(\text{z})$ is the starting moment and $\mathcal{E}(\text{z})$ the ending moment of the z-th oversaturated period, respectively. $\mathcal{O}$ and $\mathcal{E}$ are obtained with Algorithm 1.

Algorithm 1. Oversaturation Detection with Continuous Inputs

```
Input: $T$; $I$; $c$; $\underline{h}$; $a(t), A(t), \forall t \in [0, T]$
$\mathcal{O} \leftarrow \emptyset$
$\mathcal{E} \leftarrow \emptyset$
$SSP \leftarrow 0$
$z \leftarrow 0$
while $z = |\mathcal{O}|$
    $z \leftarrow z + 1$
    $\mathcal{O}(z) \leftarrow \underset{t \in (SSP, T]}{\text{arginf}} \left( \int_{\max\{t-\underline{h}, 0\}}^{t} a(t)\, dt > Ic \right)$
    $\mathcal{E}(z) \leftarrow \underset{t \in (\mathcal{O}(z), T]}{\text{arginf}} \left( \frac{A(t) - A(\mathcal{O}(z))}{t - \mathcal{O}(z)} \leq \frac{Ic}{\underline{h}} \right)$
    $SSP \leftarrow \mathcal{E}(z)$
end while
Output: $\mathcal{O}$; $\mathcal{E}$
```

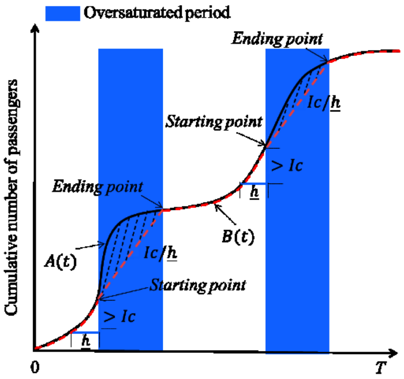

Fig. 5: Oversaturated and unsaturated periods

Note that $T \in \mathcal{T}^U$ since no feasible solution will otherwise exist satisfying constraints (7). Further, since unsaturated and oversaturated periods appear in an alternating pattern, thus for simplicity, we can list the unsaturated periods as $\mathcal{T}^U = \{[\mathcal{E}(0), \mathcal{O}(1)), [\mathcal{E}(1), \mathcal{O}(2)), \cdots, [\mathcal{E}(|\mathcal{O}|), \mathcal{O}(|\mathcal{O}| + 1)]\}$ where $\mathcal{E}(0) = 0$ and $\mathcal{O}(|\mathcal{O}| + 1) = T$. With this definition, we see that the starting moment of an oversaturated period, $\mathcal{O}(z)$, is the first $t \in [0, T]$ during an unsaturated period when the cumulative arrival demand from $\max\{t - \underline{h}, 0\}$ to t is identical to Ic. Further, the ending moment of an oversaturated period, $\mathcal{E}(z)$, is characterized as the first $t \in [0, T]$ when the queue that starts at $\mathcal{O}(z)$ vanishes. With this definition, Theorems 1 and 2 imply that for any dispatch within an oversaturated period, a vehicle in formation I is dispatched with headway $\underline{h}$. Following the way Newell (1982) and Daganzo (1997) decomposed the oversaturated delay from unsaturated delay by drawing a straight line with an even service rate, we can revise arrival curve $A(t)$ in the following way to simplify the problem.

Definition 2. For a given arrival curve $A(t)$ at the origin station, the corresponding virtual arrival demand curve is defined as

$$B(t) := \begin{cases} A(t) & t \in \mathcal{T}^U \\ \frac{Ic}{\underline{h}}\left(t - \mathcal{O}(z)\right) + A\left(\mathcal{O}(z)\right) & t \in [\mathcal{O}(z), \mathcal{E}(z)), \forall z \in [1, |\mathcal{O}|] \end{cases}$$

and thus the virtual arrival demand rate $B'(t) \leq Ic/\underline{h}, \forall t \in [0, T]$ (the right derivative is used at points that are not differentiable). Obviously, $A(t) \geq B(t), \forall t \in [0, T]$. This is illustrated as the red dashed curve in Fig. 5.

Lemma 3. *Denote problem* (2) ~ (8) *as the original problem (OP), and the same problem where the original arrival curve* $A(t)$ *is replaced with virtual arrival curve* $B(t)$ *as the revised unsaturated problem (RUP). Then the feasible regions of the OP and RUP are the same. For any feasible solution* $\mathbf{s} := \{t_k, i_k\}_{\forall k \in \mathcal{K}}$*, OP and RUP have the same dispatch curve, and the objective values of OP and RUP, respectively denoted as* $OP(\mathbf{s})$ *and* $RUP(\mathbf{s})$*, are always separated by a constant difference:*

$$OP(\mathbf{s}) - RUP(\mathbf{s}) = W(A) \coloneqq w \int_0^T \big(A(t) - B(t)\big)\, dt.$$

Lemma 3 directly leads to the following result.

Theorem 3. *Problems OP and RUP have the same optimal solution(s). For an optimal solution* $\mathbf{s}$ *to both problems,* $OP(\mathbf{s}) - RUP(\mathbf{s}) = W(A)$.

Theorem 3 implies that instead of solving OP, we can just solve RUP with virtual arrival curve $B(t)$. Note that RUP has constantly unsaturated arrival demands since $B'(t) \le \frac{Ic}{\underline{h}}, \forall t \in [0, T]$, which can dramatically reduce the complexity of the investigated problem. Further, note that oversaturated arrival demand $A(t)$ always leads to oversaturated waiting cost $W(A)$. This also sheds light on the demand management side, e.g., managing the arrival demand to approach the unsaturated pattern $B(t)$ (which explains why we call $B(t)$ the virtual arrival curve), which however is out of this paper's scope. Actually, since $B(t)$ is unsaturated throughout the time horizon, the theorem below shows that the corresponding optimal departure curve $D(t)$ is not far separated from $B(t)$.

Theorem 4. *An optimal dispatch solution* $\{t_k, i_k\}_{\forall k \in \mathcal{K}}$ *to RUP satisfies: (i) if* $t_k - t_{k-1} > \underline{h}$*, then* $B(t_k) - D(t_k) = 0, \forall k \in \mathcal{K}$*; and (ii) if* $t_k - t_{k-1} = \underline{h}$*, then* $B(t_k) - D(t_k) \in [0, c), \forall k \in \mathcal{K}$*, where* $D(t)$ *is the corresponding departure curve.*

Interestingly, the property of having minimum queue at every dispatch (which approaches to 0 as $c \to 0$) echoes the finding in Newell (2002) that the queue in the major approach needs to be always cleared before changing the signal. This alludes to certain structural connection between these two seemingly unrelated problems that is worth future investigation. These elegant properties simplify the problem structure and largely reduce the domain where an optimal solution may appear for OP. They will be used to decompose the problem in the CA model in the following section.

4. Continuum approximation

This section presents a CA model that aims to shed macroscopic analytical insights and tackle large-scale problems. A CA approach can approximate a local neighborhood in the searching space with a continuous function whose parameters are homogenous when properties of the searching space change slowly (Li et al., 2016). If the unit-time or unit-area cost around this neighborhood is mainly determined by the settings in its vicinity but less dependent on distant neighborhoods, then the optimal solution to this separable continuous function can well approximate the optimal solution of this neighborhood. In our problem, the total cost in a neighborhood is largely determined by the vehicle capacity and the dispatch headway around this neighborhood and $A'(t)$ varies relatively smoothly in each neighborhood (e.g., with a time period comparable to the optimal dispatch headway). Thus, we can well approximate the original model with CA.

4.1. Model formulation

Theorem 3 shows that instead of solving OP, we can just solve RUP with virtual arrival curve $B(t)$ where only unsaturated traffic is present. Hence, we can formulate the CA model based on the RUP.

Given an optimal solution $\mathbf{s} \coloneqq \{t_k, i_k\}_{\forall k \in \mathcal{K}}$ to RUP with virtual cumulative arrival $B(t)$ and departure $D(t)$, we define the dispatch headway at time $t \in [0, T]$ as $\hat{h}(t) \coloneqq t_k - t_{k-1}, \text{s.t.}\, t \in (t_{k-1} t_k], \exists k \in \mathcal{K}$ and the vehicle formation at time $t \in [0, T]$ as $i(t) \coloneqq i_k, \text{s.t.}\, t \in (t_{k-1}, t_k], \exists k \in \mathcal{K}$. For the convenience of the notation, define $\hat{h}(0) \coloneqq t_1 - t_0$ and $i(0) \coloneqq 1$. Then, the optimal objective value of RUP can be formulated as

$$RUP(\mathbf{s}) = \sum_{k \in \mathcal{K}} f_{i_k} + w \int_0^T \big(B(t) - D(t)\big) dt = \int_0^T \left(\frac{f_{i(t)}}{\hat{h}(t)} + w\big(B(t) - D(t)\big) \right) dt, \tag{9}$$

s.t. (3) ~ (8) （where $A(t)$ is replaced with $B(t)$） .

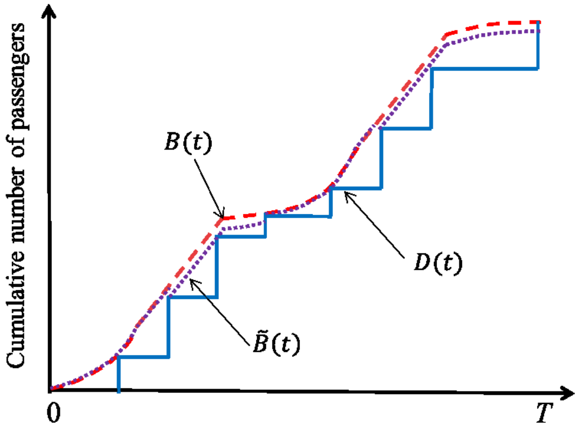

Fig. 6: Revised virtual arrival demand curve

In order to approximate the OP, we revise the virtual arrival demand curve as $\tilde{B}(t) \coloneqq B(t) - \big(B(t_{k-1}) - D(t_{k-1})\big), \forall t \in (t_{k-1}, t_k], k \in \mathcal{K}$ (see the dotted line in Fig. 6). Then we reformulate (9) as

$$RUP(\mathbf{s}) = \int_0^T \left(\frac{f_{i(t)}}{\hat{h}(t)} + w\left(\tilde{B}(t) - D(t)\right)\right) dt + \; w\int_0^T \left(B(t) - \tilde{B}(t)\right) dt \tag{10}$$

$$= \sum_{k\in\mathcal{K}} \int_{t_{k-1}}^{t_k} \left(\frac{f_{i(t)}}{\hat{h}(t)} + w\left(\tilde{B}(t) - D(t)\right)\right) dt + \; w\int_0^T \left(B(t) - \tilde{B}(t)\right) dt.$$

From Theorem 4, we learn that $B(t) - \tilde{B}(t) < c, \forall t \in [0, T]$ and thus $\int_0^T \left(B(t) - \tilde{B}(t)\right) dt < cT$. Note that cT is a relatively small value compared with $RUP(\mathbf{s})$. Further, numerical experiments with exact solution methods show that $B(t) - \tilde{B}(t)$ is 0 at most of the time (see Section 5.1.2). Therefore, we can drop $w\int_0^T \left(B(t) - \tilde{B}(t)\right) dt$ from (10) and obtain

$$RUP(\mathbf{s}) \approx \sum_{k\in\mathcal{K}} \left[\int_{t_{k-1}}^{t_k} \left(\frac{f_{i(t)}}{\hat{h}(t)}\right) dt + w\int_{t_{k-1}}^{t_k} \left(\left(\tilde{B}(t) - D(t)\right)\right) dt\right]. \tag{11}$$

Note that in a specific neighborhood $[t_{k-1}, t_k], k \in \mathcal{K}$, $\tilde{B}(t)$ is a continuous function. With $\tilde{B}'(t) = \frac{\tilde{B}(t)-\tilde{B}(t_{k-1})}{t-t_{k-1}}, \forall t \in [t_{k-1}, t_k]$, we have $\int_{t_{k-1}}^{t_k} \left(\tilde{B}(t) - D(t)\right) dt = \int_{t_{k-1}}^{t_k} \left(\tilde{B}(t) - \tilde{B}(t_{k-1})\right) dt = \int_{t_{k-1}}^{t_k} (t - t_{k-1})\tilde{B}'(t)dt = \int_{t_{k-1}}^{t_k} (t - t_{k-1})B'(t)dt$, which yields

$$RUP(\mathbf{s}) \approx \sum_{k\in\mathcal{K}} \left[\int_{t_{k-1}}^{t_k} \left(\frac{f_{i(t)}}{\hat{h}(t)}\right) dt + wB'(t)\int_{t_{k-1}}^{t_k} (t - t_{k-1}) dt\right]$$

$$\approx \sum_{k\in\mathcal{K}} \left[\int_{t_{k-1}}^{t_k} \left(\frac{f_{i(t)}}{\hat{h}(t)}\right) dt + wB'(t)\frac{(t_k - t_{k-1})^2}{2}\right] \tag{12}$$

$$\approx \sum_{k\in\mathcal{K}} \left[\int_{t_{k-1}}^{t_k} \left(\frac{f_{i(t)}}{\hat{h}(t)}\right) dt + wB^{\prime}(t)\int_{t_{k-1}}^{t_k} \left(\frac{t_k - t_{k-1}}{2}\right) dt\right]$$

$$\approx \int_0^T \left(\frac{f_{i(t)}}{\hat{h}(t)} + \frac{wB'(t)\hat{h}(t)}{2}\right) dt.$$

Note that in Eq.

(12), $B'(t)$ can be pulled out of the integral since $B'(t)$ varies relatively slowly in each local time neighborhood so that we can regard it approximately as a constant. Also note that $\hat{h}(t)$ is essentially a step function, to solve which is akin to determining $t_k, \forall k \in \mathcal{K}$ themselves. To simplify this, we approximate $\hat{h}(t)$ with a smooth function $h(t)$ and then the RUP can be decomposed across the operational horizon as the following unit-time RUP

$$\min_{h(t),i(t)\in\mathcal{I}} c_t\big(i(t),h(t)\big) \coloneqq \frac{f_{i(t)}}{h(t)} + \frac{wB'(t)h(t)}{2}, \forall t \in [0,T]. \tag{13}$$

Further, based on Theorem 4 and the fact that $B(t) - \tilde{B}(t)$ is 0 at most of the time, we can view that $B(t) - \tilde{B}(t) = 0$ approximately holds $\forall t \in [0,T]$ in the optimal solution(s). That is to say, the queue length reduces to 0 for almost all dispatches across the operational horizon in the optimal solution(s). With this, $h(t) < \frac{i(t)c}{B'(t)}$ must approximately hold in the optimal solution(s) (and otherwise $B(t) - \tilde{B}(t) = 0, \forall t \in [0,T]$ would be violated; i.e., the queue lengths after some dispatches are not 0), which together with (3) yields the following constraints

$$\underline{h} \le h(t) < \frac{i(t)c}{B'(t)}, \forall i(t) \in \mathcal{I}, t \in [0,T]. \tag{14}$$

$$\underline{h} < \frac{i(t)c}{B'(t)}, \forall i(t) \in \mathcal{I}, t \in [0,T]. \tag{15}$$

Note that the above unit-time RUP at each time point t, i.e., (13), (14) and (15), has only two decision variables and three constraints so it is much simpler than both the original formulation and the discrete model. By solving the unit-time RUP $\forall t \in [0,T]$, we can apply the optimal solution $c^*(t) \coloneqq \min_{h(t),i(t)\in\mathcal{I}} c_t\big(i(t),h(t)\big)$ to RUP to obtain the objective value of OP as

$$OP(\mathbf{s}) \approx RUP(\mathbf{s}) + W(A) \approx \int_0^T c^*(t)dt + W(A). \tag{16}$$

4.2. Analytical solution

This section presents a closed-form analytical optimal solution to the proposed CA model. As mentioned in the previous section, the OP can be solved by solving the unit-time RUP across the operational horizon. Thus, the following analysis primarily focuses on the unit-time RUP.

This solution approach for Problem (13) at each time $t \in [0,T]$ has two steps. The first step is to solve $h(t)$ for a given feasible $i(t) \in \mathcal{I}(t)$ where $\mathcal{I}(t) \coloneqq \{i | i \in \mathcal{I}, ic > \underline{h}B'(t)\}$ is the set of vehicle formations at time $t \in [0,T]$ feasible to Constraints (15). The second step solves the full problem by jointly optimizing $h(t)$ and $i(t)$.

Now we describe the first step for a given $i(t) \in \mathcal{I}(t)$. Without considering Constraints (14), note that since $\frac{\partial c_t(i(t),h(t))}{\partial h(t)} = -\frac{f_{i(t)}}{(h(t))^2} + \frac{wB'(t)}{2}$, function $c_t\big(i(t),h(t)\big)$ for a fixed $i(t)$ strictly decreases with $h(t) \in \left(0, \sqrt{\frac{2f_{i(t)}}{wB'(t)}}\right)$ and strictly increases with $h(t) \in \left(\sqrt{\frac{2f_{i(t)}}{wB'(t)}}, \infty\right)$. With this property, we will solve the optimal $h(t)$ for given $i(t)$, denoted by $h^*_{i(t)}(t)$, after adding back Constraints (14). Solving the left hand side and right hand side of Constraints (14) yield $a(t) \le \frac{2f_{i(t)}}{w\underline{h}^2}$ and $a(t) \le \frac{w[i(t)c]^2}{2f_{i(t)}}$, respectively. Depending on the relationship between $\frac{2f_{i(t)}}{w\underline{h}^2}$ and $\frac{w[i(t)c]^2}{2f_{i(t)}}$, the analytical solution to the unit-time RUP can be divided into the following two cases.

Case 1. When $0 < \underline{h} \le \frac{2f_{i(t)}}{wi(t)c}$, *i.e.* $\frac{w[i(t)c]^2}{2f_i(t)} \le \frac{2f_{i(t)}}{w\underline{h}^2}$.

If $0 \le B'(t) \le \frac{w[i(t)c]^2}{2f_i(t)}$, we can obtain $\underline{h} \le \sqrt{\frac{2f_{i(t)}}{wB'(t)}} \le \frac{i(t)c}{B'(t)}$ and thus $h^*_{i(t)}(t) = \sqrt{\frac{2f_{i(t)}}{wB'(t)}}$. If $\frac{w[i(t)c]^2}{2f_i(t)} < B'(t) < \frac{2f_{i(t)}}{w\underline{h}^2}$, we can obtain that $\sqrt{\frac{2f_{i(t)}}{wB'(t)}} > \frac{i(t)c}{B'(t)}$ and thus $h^*_{i(t)}(t) = \frac{i(t)c}{B'(t)}$. If $B'(t) \ge \frac{2f_{i(t)}}{w\underline{h}^2}$, based on $0 < \underline{h} \le \frac{2f_{i(t)}}{wi(t)c}$, we can obtain $\underline{h} \geq \frac{i(t)c}{B'(t)}$, which contradicts to (15). Thus, $B'(t) \geq \frac{2f_{i(t)}}{w\underline{h}^2}$ cannot happen if $\underline{h} \leq \frac{2f_{i(t)}}{wi(t)c}$. Thus, in an optimal solution to the unit-time RUP for a given feasible $i(t) \in \mathcal{I}(t)$, when $0 < \underline{h} \leq \frac{2f_{i(t)}}{wi(t)c}$, the optimal solution $h(t)$ is

$$h^*_{i(t)}(t) = \begin{cases} \sqrt{\frac{2f_{i(t)}}{wB'(t)}}, & \text{if } 0 \leq B'(t) \leq \frac{w[i(t)c]^2}{2f_i(t)} \\ \frac{i(t)c}{B'(t)}, & \text{if } B'(t) > \frac{w[i(t)c]^2}{2f_i(t)} \end{cases}, \forall\, i(t) \in \mathcal{I}(t), t \in [0, T]. \tag{17}$$

Case 2. When $\underline{h} > \frac{2f_{i(t)}}{wi(t)c}$, i.e., $\frac{w[i(t)c]^2}{2f_i(t)} > \frac{2f_{i(t)}}{w\underline{h}^2}$.

If $0 \leq B'(t) \leq \frac{2f_{i(t)}}{w\underline{h}^2}$, we can obtain $\underline{h} \leq \sqrt{\frac{2f_{i(t)}}{wB'(t)}} \leq \frac{i(t)c}{B'(t)}$ and thus $h^*_{i(t)}(t) = \sqrt{\frac{2f_{i(t)}}{wB'(t)}}$. If $\frac{2f_{i(t)}}{w\underline{h}^2} < B'(t) < \frac{w[i(t)c]^2}{2f_i(t)}$, we can obtain $\sqrt{\frac{2f_{i(t)}}{wB'(t)}} < \underline{h}$ and thus $h^*_{i(t)}(t) = \underline{h}$. If $B'(t) \geq \frac{w[i(t)c]^2}{2f_i(t)}$, based on $\underline{h} > \frac{2f_{i(t)}}{wi(t)c}$, we can obtain $\underline{h} > \frac{i(t)c}{B'(t)}$, which contradicts to (15). Thus, $B'(t) \geq \frac{w[i(t)c]^2}{2f_i(t)}$ cannot happen if $\underline{h} > \frac{2f_{i(t)}}{wi(t)c}$. Thus, in an optimal solution to the unit-time RUP for a given feasible $i(t) \in \mathcal{I}(t)$, when $\underline{h} > \frac{2f_{i(t)}}{wi(t)c}$, then the optimal solution $h(t)$ is

$$h^*_{i(t)}(t) = \begin{cases} \sqrt{\frac{2f_{i(t)}}{wB'(t)}}, & \text{if } 0 \leq B'(t) \leq \frac{2f_{i(t)}}{w\underline{h}^2} \\ \underline{h}, & \text{if } B'(t) > \frac{2f_{i(t)}}{w\underline{h}^2} \end{cases}, \forall\, i(t) \in \mathcal{I}(t), t \in [0, T]. \tag{18}$$

With these results, we are ready to move to the next step on the joint optimization. Because the cardinality of $\mathcal{I}(t)$ is limited in real-world cases, we can plug (17), (18) into the unit-time RUP and then solve it simply by exhaustive enumeration, i.e.,

$$i^*(t), h^*(t) = \underset{i(t)\in\mathcal{I}(t)}{\operatorname{argmin}} \left\{ c_t\left(i(t), h^*_{i(t)}(t)\right) \right\}, \forall t \in [0, T], \tag{19}$$

which directly leads to the optimal unit-time cost $c^*(t) = c_t\big(i^*(t), h^*(t)\big), \forall t \in [0, T]$. Finally, after solving the RUP's for all neighborhoods $t \in [0, T]$, we can plug $c^*(t), \forall t \in [0, T]$ into (16) to obtain an estimated value of $OP(\mathbf{s})$.

4.3. Discretization

The analytical solution provides us with a closed-form expression of the optimal solution to the original problem. Yet the continuous solutions $h^*(t), i^*(t)$ cannot be directly applied to the discrete vehicle scheduling problem. Therefore, a discretization approach is needed to convert $h^*(t)$ into discrete time points for each dispatch $t^*_k, \forall k \in \mathcal{K}$ and $i^*(t)$ into the discrete vehicle capacity $i^*_k, \forall k \in \mathcal{K}$. Daganzo (2005) proposed a systematic method that can find an approximate step function to the continuous headway function generated by the CA model efficiently. The idea of Daganzo's method is to find a fixed headway such that area under the $h^*(t)$ curve and that above the curve is the same. However, due to the potential computation difficulties in Daganzo's method, e.g. evaluating the areas below and above the curve, finding points on the vertical axis to draw horizontal segments, we propose a method to discretize $h^*(t)$ by using the definition of headway, i.e. $h^*(t^*_k) = t^*_k - t^*_{k-1}$, iteratively for $k = K, K-1, \cdots 1$, as shown in Fig. 7. Specifically,

Algorithm 2. Discretization

```
Input: $h^*(t), i^*(t), \forall t \in [0, T], \underline{h}$
$K \leftarrow 0$
$t^*_K \leftarrow T$
while $t^*_K \geq \underline{h}$
    $K \leftarrow K + 1$
    $t^*_K \leftarrow \underset{t\in[0,t^*_{K-1})}{\operatorname{argsup}} (h^*(t) = t^*_{K-1} - t)$
    $i^*_{K-1} \leftarrow Round\left(\frac{\int_{t^*_K}^{t^*_{K-1}} i^*(t)dt}{t^*_{K-1} - t^*_K}\right)$
end while
$i^*_K \leftarrow \frac{\int_0^{t^*_K} i^*(t)dt}{t^*_K}$
Reverse $(i^*_k, t^*_k) = (i^*_{K-k}, t^*_{K-k}), \forall k \in \mathcal{K} := \{0,1, \cdots, K\}$
Output: $t^*_k, i^*_k, \forall k \in \mathcal{K}$
```

given t_k^*, we need to determine $h^*(t_k^*)$ and t_{k-1}^* such that $h^*(t_k^*) = t_k^* - t_{k-1}^*$. This can obviously be realized by drawing a 45° line from $(t = t_k^*, h^*(t) = 0)$ backward in time and find its intersection with $h^*(t)$. Thus, we can discretize $h^*(t)$ by drawing 45° lines recursively for $k = K, K-1, \cdots, 1$. With the discretized $h^*(t)$, $i^*(t)$ can then be discretized with a weighted average method. The basic framework of this approach is shown as the pseudocode in Algorithm 2.

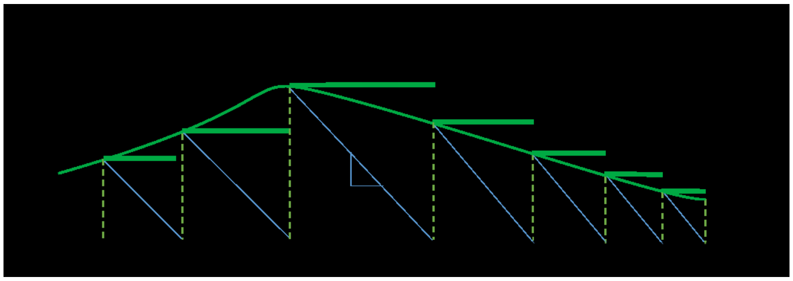

Fig. 7: Discretization of $h^*(t)$

Since the investigated problem requires that all passengers must be transported at the end of the operational horizon, this algorithm starts from the last time point T, i.e. $t_K^* = T$. From this point $(t = T, h^*(t) = 0)$, we draw a 45° line backward in time and find its intersection with $h^*(t)$, the abscissa of which is the time point of the previous dispatch, t_{K-1}^*. Then we compute the weighted average of $i^*(t), \forall t \in [t_{k-1}, t_k]$ and round it to an integer, which results in the vehicle formation for the K^{th} dispatch, i.e. i_K^*. This step locates t_{K-1}^* given t_K^* and determines i_K^*. We repeat this procedure from t_{K-1}^* to locate the previous dispatch times and formations until reaching a point whose distance to the origin is no greater than the minimum headway. Note that when implementing this algorithm, the index starts from 0 to K, so at the end we need to reverse the indexes to obtain the correct solution. Once the discrete time point $t_k^*, \forall k \in \mathcal{K}$ and vehicle formation $i_k^*, \forall k \in \mathcal{K}$ are obtained, we can plug them into objective function (2) to obtain the total system cost with respect to the discrete design.

5. Numerical experiments

This section presents two sets of numerical examples. All the experiments are run in MATLAB 2017b on a DELL Studio PC with 3.60 GHz of Intel Core i7-7700 CPU and 16 GB of RAM in a Windows environment. The first set of experiments is built on smart card data collected from Batong line in the Beijing Metro System, China where strong temporal demand fluctuations and oversaturated traffic can be observed. The objectives of these experiments are to compare the computation performances of the proposed solution methods and demonstrate the effectiveness of the time-varying capacity design. The second set of experiments investigates modular autonomous vehicle (MAV) by simulating a shuttle system with MAV information from Next Future Transportation Inc. (http://www.next-future-mobility.com/) and future travel demand data in Tampa Bay Area, USA (obtained from Gannett Fleming, Inc.). With these numerical experiments, we demonstrate how future autonomous transportation technology can benefit from time-varying capacity design and reveal managerial insights into optimal system operators.

5.1. Case study 1: Batong Line in Beijing Metro System

We first explain how we set up the numerical experiments by transforming a corridor system in Beijing Metro to a shuttle system that our model can address. As shown in Fig. 8, Batong line is a bi-directional line with a total length of 18.964 km and 13 stations that are numbered sequentially from 1 to 13. Since both directions show strong temporal demand fluctuations and an evident oversaturated period in each direction over a day, without loss of generality, we selected the direction with a morning oversaturated period (i.e., the direction from Stations 13 to 1) for our experiments. The maximum passenger flow section in this direction is 4-3, so we treat stations 4-13 as a virtual origin while stations 1-3 and all stations on other lines in the metro network a virtual destination. Note that all stations on other lines in the network are considered since the destinations of a large portion of demand emanating from stations 4-13 are out of Batong line. In this way, this metro line is converted into a one-to-one shuttle system that essentially represents the bottleneck of the metro line. Then, we count the number of passengers that will cross the maximum passenger flow section (i.e. passengers whose origins are one of stations 4-13 and destinations are one of the rest of stations in the network) per 0.1 minute and obtain the time-dependent cumulative arrival demand curve and the arrival demand rate curve over the entire operational horizon as Fig. 9. Note that in this case study we are solving an optimal schedule to the converted shuttle system (as if trips not passing through the bottleneck do not exist) rather than to the original corridor system. Such an analysis (i.e. operational design for multiple stop systems based on simply demands passing through the maximum load points) has been applied in Salzborn (1972) and Daganzo (1997). Further, though they are not optimal, solutions from the converted shuttle system suffice to provide an approximate optimal or at least a feasible solution to the optimal operation of the original corridor system. Therefore, results in this case study can still provide managerial insights to system operators.

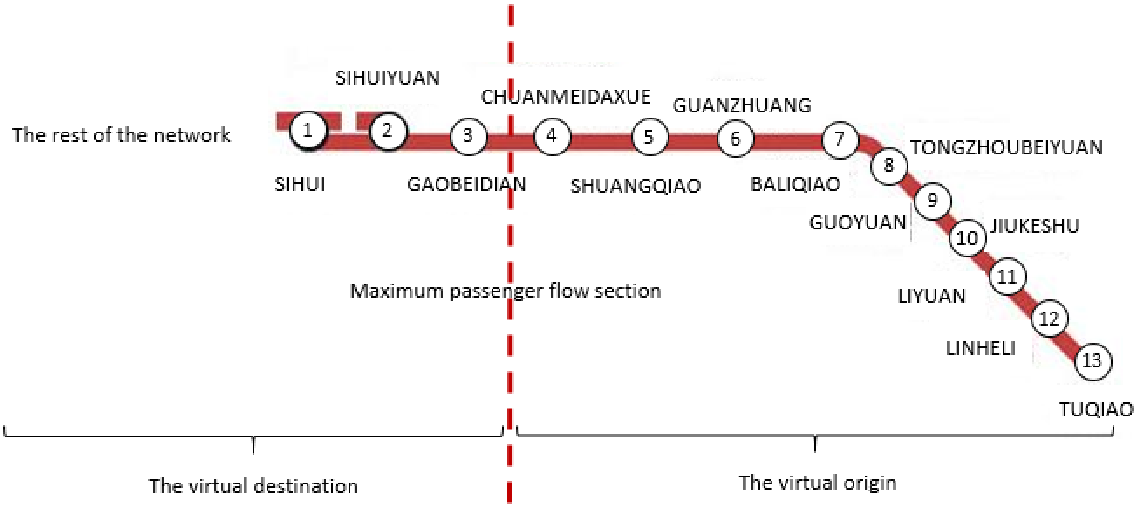

Fig. 8: Batong Line and its corresponding shuttle system and the converted shuttle system

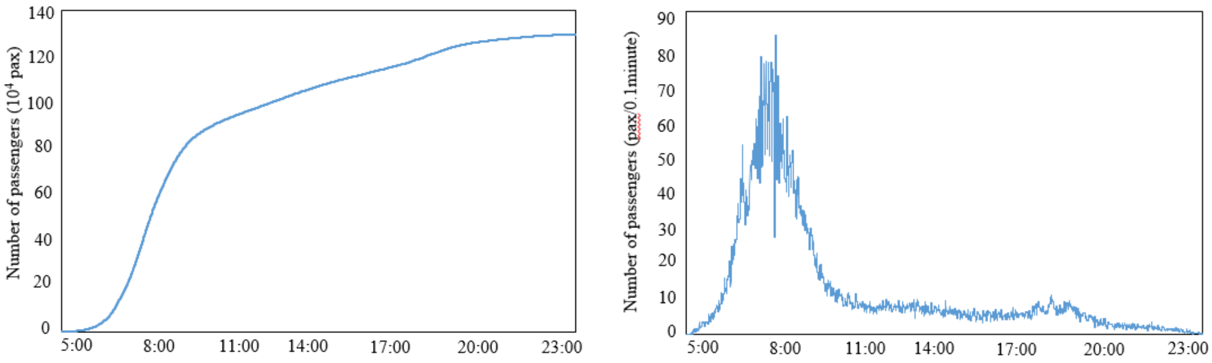

Fig. 9: Cumulative arrival demand (left) and arrival demand rate (right) over an operational day

The trains serving on this line are composed of 6 carriages, each with a capacity of 226 pax (i.e., passengers). Hence, we set $\mathcal{I} = [1,2,3,4,5,6]$ and $c = 226$ pax/carriage. The minimum dispatch headway is set as $\underline{h} = 3$ minutes. We adopt the monthly average salary per capita to compute the waiting cost per passenger per unit time. Beijing Municipal Human Resources and Social Security Bureau reports that the average monthly salary per capita is \$1096.19 in 2017 (http://www.bjrbj.gov.cn/). Provided that one works 22 days a month and 8 hours per day, the hourly average salary per capital is around \$6.35 and thus $w = 0.11$ \$/min. The dispatch energy cost function we adopt in this case is $f_i = C^F + C^V(i)^\alpha, \forall i \in \mathcal{I}$ and we obtain $\alpha = 0.5$, C^F=2.049 \$, and C^V=5.56 \$ through a calibration process that is explained in Appendix B. .

5.1.1. Computation performances

We first design 21 instances with three operational horizons, i.e., 3 hours (7:00-10:00), 9 hours (5:00-14:00), and 18 hours (5:00-23:00), and 7 discretization intervals, i.e., 0.1, 0.2, 0.3, 0.5, 1.0, 1.5, and 3.0 minutes to investigate the computation performance of the proposed CA method. Fig. 10 plots the computation times of CA of these instances, which are indexed as 1-7 for the instances with the 3-hour operational horizon and the discretization interval decreasing from 3 through 0.1 minutes, 8-14 for the instances with the 9-hour operational horizon and the discretization interval decreasing from 3 through 0.1 minutes, and 15-21 for the instances with the 18-hour operational horizon and the discretization interval decreasing from 3 through 0.1 minutes. We see that the CA method can solve all instances in less than 1 second, as opposed to over thousands of seconds from the discrete methods (see Table 4 in Chen et al., (2018) for the detailed computation time comparison). Further, CA's computational time does not increase too fast with the instance size. Actually, since each neighborhood is just solved analytically in a constant time, the computational time increases only linearly with the number of total neighborhoods, which depends on the discrete size of a neighborhood and the total time horizon. The linear time complexity of CA is much more computationally-friendly than discrete models whose solution time often exponentially increases with the instance size. Further, comparison with the exact solutions from the counterpart discrete models (see Table 3 in Chen et al., (2018)) shows that the CA methods can produce satisfactory approximations of the optimal objective value across all instances, with the largest optimality gap of 0.63% from CA-I (CA with integral) and 1.11% from CA-D (CA with discretization) across all these instances. Overall, CA is a very efficient solution approach without much approximation error for the investigated problem, and it is particularly suitable for extremely large scale instances (which would not be solved otherwise) or real-time applications (which may require sub-second response time) in engineering contexts striving for a reasonably good (yet not necessarily the exact optimal) solution with limited computational resources.

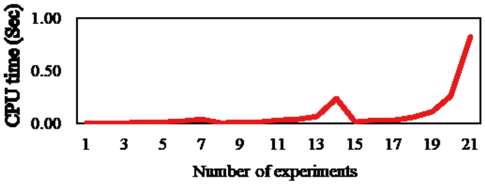

Fig. 10: Computation performance of CA.

5.1.2. Result demonstration

In this subsection, we present some numerical results to verify the analytical theorems in Section 3.2 and further analyze the accuracy of the CA model. We select 2 instances with a discretization interval of 3 minutes from the previous subsection and plot their $A(t)$, $B(t)$ and $D(t)$ obtained from G-NSR (the discrete model solved by an existing commercial solver, Gurobi) on the left in Fig. 11, with a zoom in of the oversaturated periods. As can be seen from these figures, there is an oversaturated period (i.e. where $A(t) > B(t), t \in [0, T]$) and trains with 6 modular units are dispatched with a headway of 3 minutes within this period, which verifies Theorems 1 and 2. Further, we solve these 2 instances with G-NSR using $A(t)$ and $B(t)$ as the demand; i.e., we solve the original problem and the corresponding revised unsaturated problem for these instances. The OP and RUP result in the same optimal solution and the objective values of the RUP (i.e., 12346 and 23741) and those of the corresponding OP (i.e., 95708 and107120) are separated by a constant value 83375, which verifies Theorem 3. Besides, we plot the differences between the cumulative arrival and departure across the operational horizon obtained from three solution methods on the right in Fig. 11, respectively. As can be seen from these figures, when we solve the instances with G-NSR and using $B(t)$ as the demand, $B(t) - D(t)$ drops to 0 whenever a train is dispatched, which verifies Theorem 4. It also demonstrates that dropping $w \int_{t=0}^{T} \left(B(t) - \tilde{B}(t)\right) dt$ in the CA model will not significantly affect the results. Besides, there are not significant differences between the blue lines and red lines in these figures, which again demonstrates that the CA model can offer near-optimal solutions to the original problem with high accuracy.

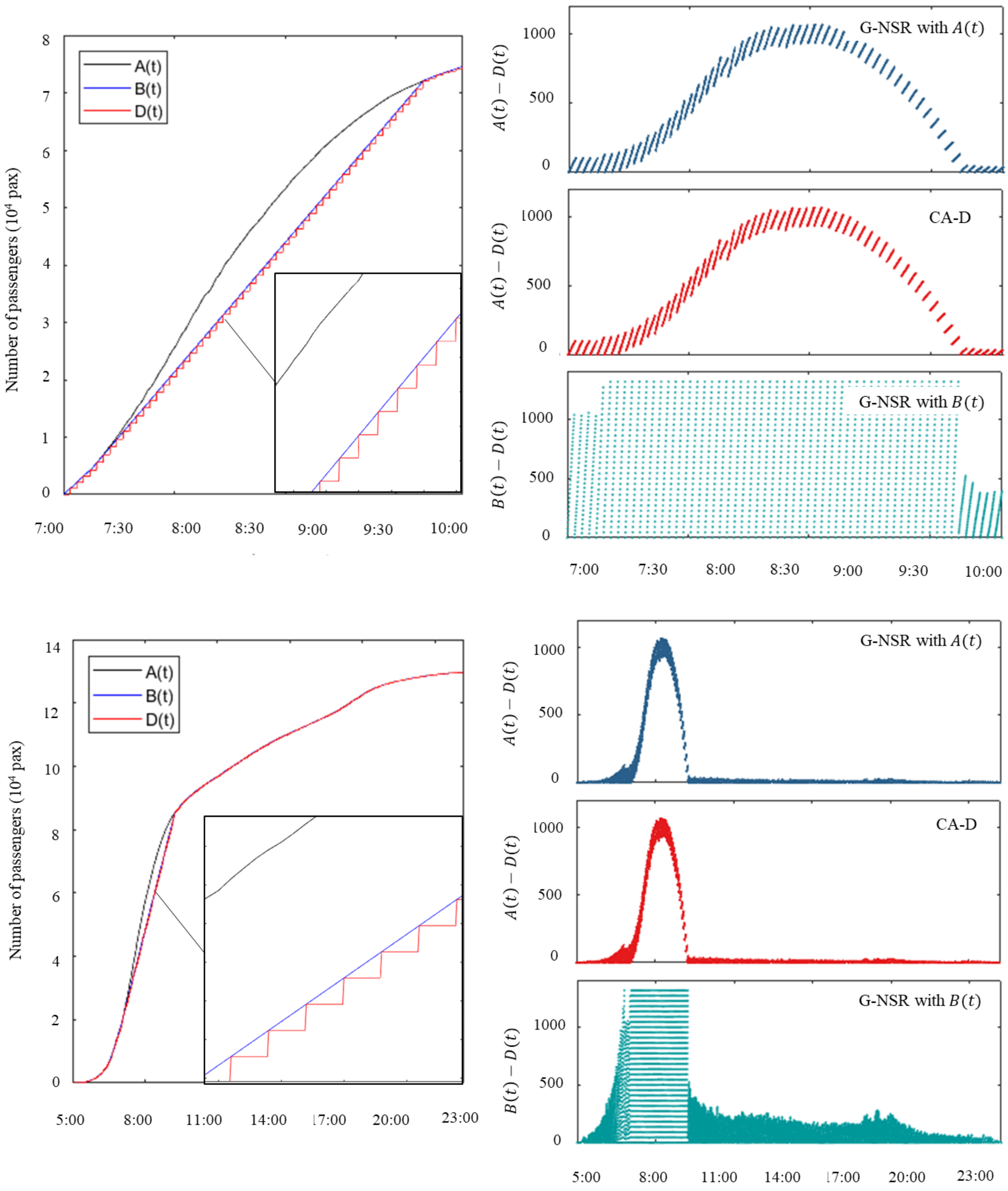

Fig. 11: Arrival and departure curves of the (a) 3-hour and (b) 18 hour instances with a discretization interval of 3 mins

Next, to provide a demonstration of the accuracy of the discretization method proposed in Section 4.3. We select 2 instances with a discretization interval of 0.1 minutes from the previous section and plot the continuous dispatch headways and vehicle formations as well as their corresponding discrete values obtained from Algorithm 2 from top to bottom in Fig. 12. It can be observed from these figures that the discretized values show a similar trend as the continuous functions and are very close to the continuous values most of the time. Although some local fluctuations can be seen in some neighborhoods (e.g. some gaps of $h^*(t)$ exist at the beginning of the operational horizon in the left figures in Fig. 12 (b), (c)), these errors can actually be offset by the inherent mechanism of the method itself and thus will not affect the near-optimal solution evidently, which is why the CA solutions after discretization is very close to the exact solution approaches.

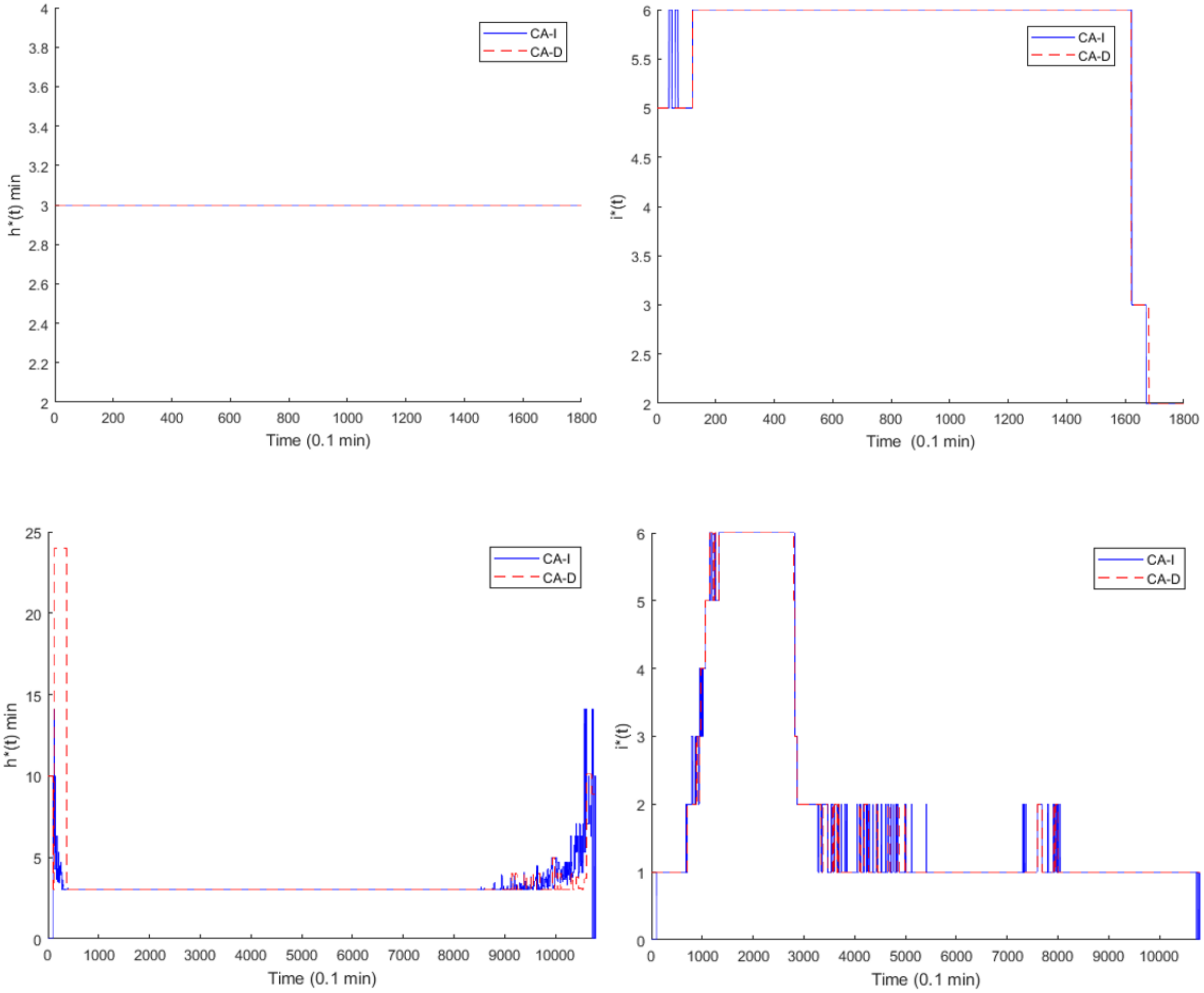

Fig. 12: Optimal dispatch headway (left) and vehicle formation (right) from CA-I and CA-D of the (a) 3-hour and (b) 18 hour instance with a discretization interval of 0.1 minutes

5.1.3. Effectiveness of time-varying train capacity adjustment under different demand scenarios

In this subsection, we assess the effectiveness of time-varying train capacity adjustment (TCA) under various demand scenarios. We benchmark the system performance with TCA against that in a benchmark system without TCA where only trains with 6 modular units can be dispatched. The arrival demand profile of the 18-hour operational horizon with a discretization interval of 0.1 minutes is used as the base case. To illustrate how different demand levels impact the effectiveness of the TCA, we test several demand scenarios. Thus, we multiplied the demand data in the base scenario by 0.25, 0.5, 1.0 and 1.5 to obtain 4 demand datasets, labeled sequentially from Scenario 1 to 4 in the following analysis. We use average load percentage (ALP), total energy cost (TEC) and total waiting cost (TWC) as system performance metrics. ALP is defined as $\frac{1}{K}\sum_{k\in\mathcal{K}}\frac{d_k}{i_k c}$ while TEC and TWC refer to the first and second terms in objective function (2), respectively, and TC is the sum of this two terms. Note that a larger value of ALP imply better system performance while TWC, TEC and TC act in the opposite way. Further, to quantitatively compare the system performance, we define a gap measure as $\frac{V_{TCA}-V_{BEN}}{V_{BEN}}$, where V_{TCA} denotes the system performance value (ALP, TWC or TEC) with TCA while V_{BEN} denotes the corresponding value of the benchmark system without TCA.

The system performance metrics of all demand scenarios are summarized in Table 1. We can see that the ALP gaps are positive while the TEC gaps are negative across all demand scenarios. It indicates that although time-varying train capacity adjustment introduces extra costs on assembling and dissembling trains, it still improves the train capacity utilization and reduces the total energy cost. Further, while in some instances the small positive gaps of TWC reveal that the total waiting cost is slightly increased, the total system cost still experience a reduction (i.e. the TC gaps are negative). Thus, in general, time-varying train capacity design can improve the overall system performances. Further, the improvements are more significant when the passenger demand is relatively low since the maximum ALP and minimum TC are both observed in the lowest demand scenario. The main reason is that, as proved in Section 3, the best policy is always to dispatch the trains in formation *I* in relatively high demand scenarios so there is not much possibility to adjust train capacities according to the time-varying demand. In contrast, a low demand scenario provides us with sufficient flexibilities to time-varying train capacity design.

Table 1. System performance metrics of different demand scenarios

Demand scenario	With TCA			Without TCA			Gap			
	ALP (%)	TWC (10^3 \$)	TEC (10^3 \$)	ALP (%)	TWC (10^3 \$)	TEC (10^3 \$)	ALP (%)	TWC (%)	TEC (%)	TC (%)
1	31.63	5.97	3.34	7.81	6.90	5.45	305.19	-13.44	-38.85	-24.66
2	47.66	10.28	3.66	14.48	10.12	5.88	229.18	1.55	-37.77	-12.90
3	67.78	104.08	4.08	28.44	103.93	5.99	138.33	0.14	-31.84	-1.60
4	73.16	680.55	4.59	42.28	680.78	6.04	73.04	-0.03	-24.06	-0.25

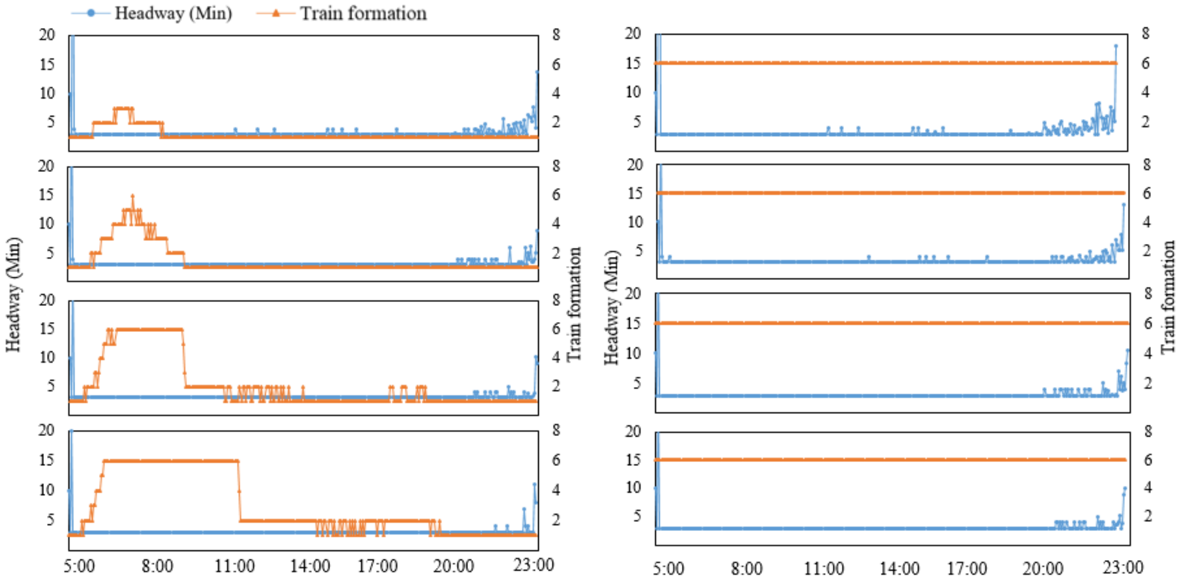

Fig. 13 Optimal dispatch headway and train formation with (left) or without (right) time-varying capacity design of Batong Line for demand scenarios 1 to 4 (from top to bottom)

To explore variations of the dispatch headway and train capacity under various demand levels, we plot the temporal headways and train formations in all demand scenarios by ranking them in a descending order from top to bottom in Fig. 13. As can be seen from these figures, the optimal headways obtained with and without TCA show similar patterns over the operational horizon. Moreover, the differences between them are not evident in most cases, which indicates that the capability to improve system performance by simply adjusting the dispatch headway is limited. In contrast, train capacities show great variations with demand fluctuations when TCA is allowed. Indeed, it is not necessary to dispatch trains in formation *I* during unsaturated periods, especially in Scenarios 1 and 2 where oversaturated periods are not present. Since the oversaturated period lasts longer in high demand scenarios, the flexibility to adjust train capacity is restricted. This, again, justifies that the effectiveness of TCA is stronger when the demand is relatively low.

5.2. Case study 2: Future modular autonomous vehicle services in Tampa

In this section, we investigate a hypothetical shuttle transportation system with modular autonomous vehicles (MAV) serving between downtown Tampa and Palm River-Clair Mel (see Fig. 14) in 2040. The travel demand over the designated operational horizon (i.e., 6:00 a.m. to 24:00 p.m., see Fig. 15) is predicted with an activity-based travel demand simulator, Daysim 2.0 (Bownman, 2018). To address the temporal demand fluctuations, pods (i.e., a single AV unit) are joint and detached dynamically during operation. According to information from the official website of Next Future Transportation Inc. (2018) and pre-experiments, we set $\mathcal{I} = [1,2,3,4,5]$ and $c = 6$ pax/pod. Based on the bus schedules of a public transit service provider in Tampa (HART, 2018), we adopt the minimum dispatch headway in their existing schedules for this case study and thus set $\underline{h} = 12$ minutes. Since no empirical data about the energy consumption of the MAV's are available, we estimate the energy cost function as follows:

1. Compute the electricity consumption $\forall i \in \mathcal{I}$. First, compute the longitudinal tractive force using Eq. (1) in (Yi and Shirk, 2018). Second, we divide the route into 5 segments based on their maximum speed limits and then compute the energy consumption on each segment by multiplying the longitudinal tractive forces on that segment and its length. Third, we sum up the electricity consumption along the routes to obtain the electricity consumption for the entire route.
2. Estimate the dispatch energy cost function through regression analysis. Eq. (1) in Yi and Shirk (2018) describes a linear relationship between the longitudinal tractive force and vehicle weight. We adopted it here since the vehicle weight is the only variable in this case study. Therefore, we apply linear regression to estimate the dispatch energy cost function.

Following the abovementioned steps, we obtain $\alpha = 1$, $C^F = 1.912$ \$ and $C^V = 3.540$ \$. Besides, based on the household income information in 2017 in Tampa (Statistical Atlas, 2017), the unit time waiting cost per passenger is set as $w = 0.8$ \$/min**.**

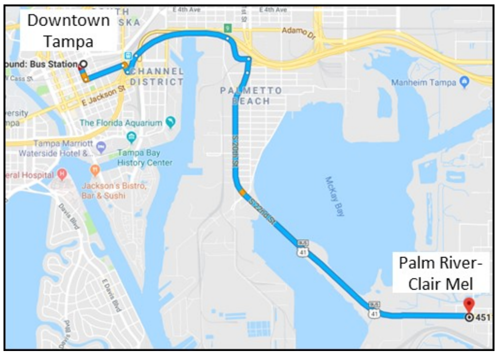

Fig. 14: Route information (source: Google Map)

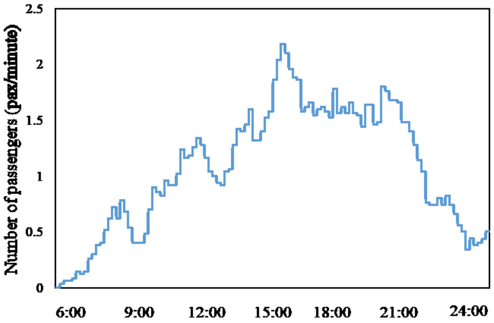

Fig. 15: Arrival demand rate

We assess the effectiveness of time-varying AV capacity design (TCA) under different demand scenarios by using system operations without TCA as a benchmark. In the benchmark system, only vehicles consisting of 5 units can be dispatched. Eight demand scenarios are considered, which are obtained by multiplying the demand profile shown in Fig. 15 by 0.1, 0.3, 0.6, 0.9, 1.0, 2.0, 3.0, 4.0. The first 5 scenarios correspond to different market penetration rates (or modal share) of the AV services while the last three consider further mature markets where more trips are induced by the sophisticated AV services. The same performance metrics proposed in Section 5.1.3 are used here. The optimal design of dispatch headway and vehicle capacity for different demand scenarios are plotted in Fig. 16 from top to bottom. The variations of performance metrics are shown in Fig. 17. As can be seen from Fig. 16, the demand in the studied area is so low that even when the market penetration rate of the AV service reaches 100%, only vehicles in formation 1 are dispatched most of the time when TCA is allowed. Further, different from rail rapid transit systems, the minimum dispatch headway in a bus transit system is a relatively large value, so there is not much space to improve the system performance through changing the dispatch headways. As a result, the optimal solutions of the first five scenarios without TCA are basically the same, which justifies the necessity to reduce the system cost through time-varying capacity design. When additional trips are induced, larger vehicles are needed due to the oversaturated traffic. In these scenarios, both operational strategies tend to select the minimum dispatch headway so the total waiting costs are the same. Yet by adjusting vehicle capacities based on the time-dependent demand, TCA can not only increase the vehicle utilization rate but reduce the total energy costs (Fig. 17).

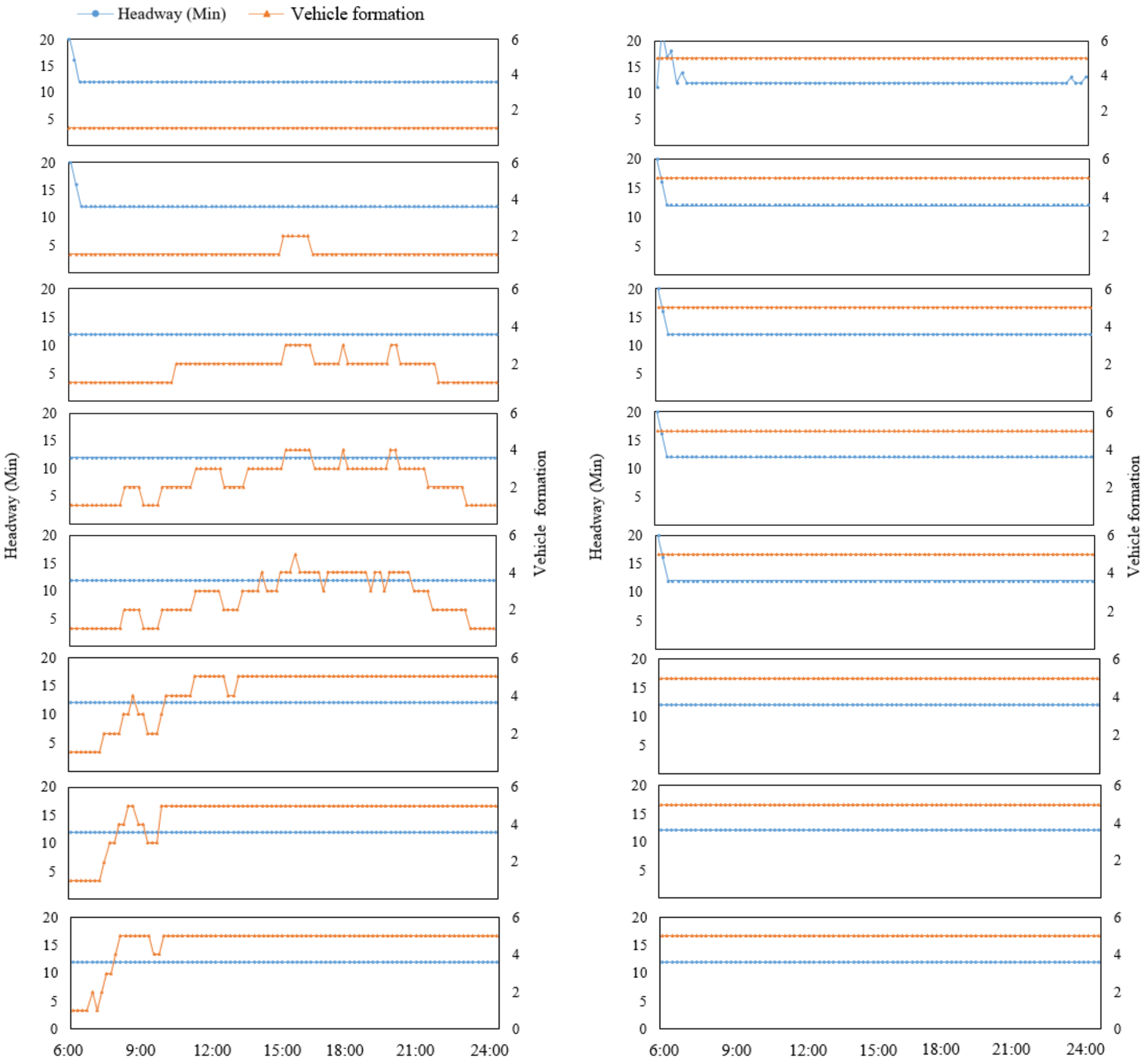

Fig. 16 Optimal dispatch headway and vehicle formation with (left) or without (right) TCA of the future MVA services for demand scenarios 1 to 8 from top to bottom

6. Conclusions

This paper investigates the joint design problem of dispatch headway and capacity for a shuttle system under oversaturated traffic. By proving that the optimal solution is to dispatch vehicles with the maximum capacity and the minimum headway within the oversaturated period, we propose a virtual arrival demand curve to investigate some theoretical properties of the RUP. Based on the theoretical property that the original problem can be solved by just solving its corresponding RUP, a macroscopic CA model is proposed to efficiently solve the problem with reasonable approximation accuracy in real-world contexts. This model decomposes the operational horizon into finite small neighborhoods with homogeneous properties such that the headway can be approximated with a continuous function.

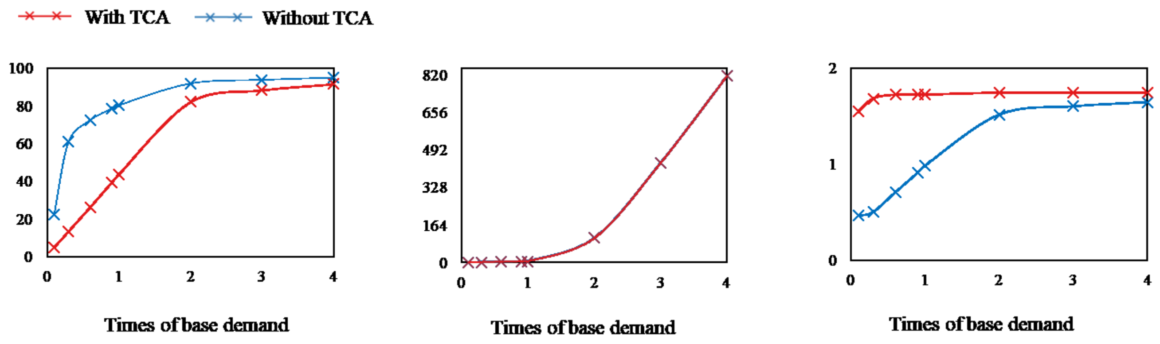

Fig. 17: Performance metrics with (left) or without (right) TCA under various demand scenarios

Then integrating the solutions of all local neighborhoods yields a near-optimum design to the original problem. Numerical experiments based on real-world traffic data collected from Batong line in Beijing Subway System and Tampa Bay Area are conducted to assess the computation performances of the proposed solution approach and verify the theoretical properties. The main findings are summarized as follows.

1. Compared with the exact solutions, the CA model offers near-optimal solutions with excellent approximation accuracy in almost no time. Therefore, the proposed modeling method holds a promising future in solving both large-scale traditional transit schedule planning problem (e.g. rail timetable design) and real-time scheduling for next-generation transportation systems (e.g. modular autonomous vehicle provided by Next Future Transportation Inc.).
2. The analytical theoretical properties are verified by the numerical experiments. Unsurprisingly, it is optimal to dispatch vehicles with the maximum capacity and the minimum headway during oversaturated periods. The joint design problem can actually be solved by simply solving its revised unsaturated problem, whose optimal solution ensures that the queue length right after each dispatch is less than the unit vehicle capacity. These results shed important managerial insights and more importantly, can be used to design efficient algorithms to find the exact solutions to the joint design problem.
3. Incorporating time-varying capacities adjustment into transit scheduling can improve the vehicle utilization rates and reduce the total cost (i.e. the total of energy cost and passenger waiting cost) for the shuttle system in both unsaturated and oversaturated traffic. The effectiveness of this innovative operational strategy is stronger when unsaturated traffic is present since vehicle capacities can be changed more flexibly.

Overall, this paper provides a methodological foundation for jointly design the dispatch headway and vehicle capacity in a shuttle system. The results in this paper can be applied to various shuttle transportation systems, such as subways, high-speed trains, bus rapid transits, modular autonomous vehicles, etc. The analytical theorems in this paper can help design efficient algorithms to solve the exact solution(s) to the joint design problem, which will be discussed in Chen et al. (2018). Also, although this study only considers the simplest one-to-one demand scenario, it constitutes a building block to develop methodologies for solving the problem in more complex settings (e.g., a corridor system with multiple OD pairs where a much more complex many to many demand scenario should be considered), which will be an interesting future research direction. Moreover, this paper solves a time-vary demand problem, addressing the dynamic design of vehicle dispatch time and capacity in more complex contexts would also be a challenging but meaningful research question. To tackle this problem, other optimization approaches such as simulation and the combination of CA tools together with advances in other related areas (discrete optimization, simulation, etc.) may be needed. Further, incorporating the fleet size management into the system dynamics is also an interesting topic. Finally, this study focuses on modular vehicles whose capacity can be adjusted at any stations in an UMTS without additional operational and delay costs. It is interesting to investigate how time-varying capacity operations can be achieved in existing UMTS where such operations incur costs and must be done in certain places in the network.

Acknowledgements

This research is supported by the U.S. National Science Foundation through Grants CMMI #1541130 and CMMI #1638355. Comments and help provided by Dr. Carlos Daganzo on CA methods approaching oversaturated problems are sincerely appreciated. Comments and data provided by Mr. Christopher Cochran from Hillsborough Area Regional

Transit Authority, Dr. Chunyu Lu from Gannett Fleming and Mr. Tommaso Gecchelin from NEXT Future Transportation Inc. are gratefully appreciated.

Appendix A. Proofs of lemmas and theorems in Section 3.2

Appendix A. includes the proofs for the lemmas and theorems in Section 3.2.

Lemma 1. *For f_i satisfying concave property* (1)*, we obtain* $f_{i_1} + f_{i_2} + \cdots + f_{i_n} \geq f_{i_1+i_2+\cdots+i_n}, \forall\, i_1, i_2, \cdots, i_n, i_1 + i_2 + \cdots + i_n \in \mathcal{I}, n \in \mathbb{Z}^+$.

Proof: Property (1) yields $\frac{i}{j} f_i + \frac{j-i}{j} f_{i+j} \leq f_j, \forall i \leq j, i + j \in \mathcal{I}$, which yields $f_i + f_j \geq \frac{i+j}{j} f_i + \frac{j-i}{j} f_{i+j}$. Further, $\frac{i}{i+j} f_{i+j} = \frac{i}{i+j} f_{i+j} + \frac{j}{i+j} f_0 \leq f_i$, which together with the previous equation yields $f_i + f_j \geq f_{i+j}$ .Then this lemma can be easily proved by iteratively applying this relationship; i.e., $f_{i_1} + f_{i_2} + \cdots + f_{i_n} \geq f_{i_1+i_2} + f_{i_3} + \cdots + f_{i_n} \geq \cdots \geq f_{i_1+i_2+\cdots+i_n}$. This completes the proof. □

Lemma 2. *For f_i satisfying concave property* (1)*, we obtain* $f_{i_1} + f_{i_4} \leq f_{i_2} + f_{i_3}, \forall i_1 \leq i_2 \leq i_3 \leq i_4 \in \mathcal{I}$ *and* $i_2 + i_3 = i_1 + i_4$.

Proof: Based on property (1), $\frac{i_4-i_2}{i_4-i_1} f_{i_1} + \frac{i_2-i_1}{i_4-i_1} f_{i_4} \leq f_{i_2}$ and $\frac{i_4-i_3}{i_4-i_1} f_{i_1} + \frac{i_3-i_1}{i_4-i_1} f_{i_4} \leq f_{i_3}$. This yields $\frac{2i_4-i_2-i_3}{i_4-i_1} f_{i_1} + \frac{i_3+i_2-2i_1}{i_4-i_1} f_{i_4} \leq f_{i_2} + f_{i_3}$. Since $i_2 + i_3 = i_1 + i_4$, the pervious equation yields $f_{i_1} + f_{i_4} \leq f_{i_2} + f_{i_3}$. This completes the proof. □

Theorem 1. *In an optimal dispatch solution* $\{t_k, i_k\}_{\forall k \in \mathcal{K}}$ *to problem* (2) ~ (8) *with arrival curve $A(t)$ and departure curve $D(t)$, if $A(t_k) - D(t_{k-1}) \geq Ic$, then $i_k = I, \forall k \in \mathcal{K}$.*

Proof. This theorem will be proven by contradiction. If the condition in this theorem does not hold, then in this optimal solution, there exists a $\bar{k} \in \mathcal{K}$ with $i_{\bar{k}} < I$ and $A(t_{\bar{k}}) - D(t_{\bar{k}-1}) \geq Ic$. Then find m such that $\sum_{m'=0}^{m} i_{\bar{k}+m'} < I$ and $\sum_{m'=0}^{m} i_{\bar{k}+m'} \geq I$, as Case 1 in Fig. 18 shows. Note that since $i_{\bar{k}} < I$, then $m \geq 1$. Further, as Case 2 in Fig. 18 shows, construct an alternate solution $\{t_k, i'_k\}_{\forall k \in \mathcal{K} \setminus \{\bar{k}+1, \cdots, \bar{k}+m-1\}}$ where $i'_{\bar{k}} = I$, $i'_{\bar{k}+m} = \sum_{m'=0}^{m} i_{\bar{k}+m'} - I$ (i.e., $I + i'_{\bar{k}+m} = \sum_{m'=0}^{m} i_{\bar{k}+m'}$), and $i'_k = i_k$ for all other k indexes. Note that if $\sum_{m'=0}^{m} i_{\bar{k}+m'} = I$, then $i'_{\bar{k}+m} = 0$, which means no vehicles are dispatched at time $t_{\bar{k}+m}$ in the alternate solution. Denote the cumulative departure curves of Case 2 as $D'(t)$. Then the difference in the waiting cost between the optimal and alternate solutions is proportional to the shaded area with vertical lines in Fig. 18, formulated as

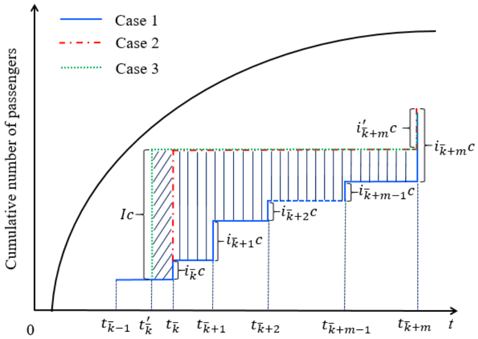

Fig. 18: An illustrative example

$$\int_{t_{\bar{k}}}^{t_{\bar{k}+m}} \big(A(t) - D(t)\big)\, wdt - \int_{t_{\bar{k}}}^{t_{\bar{k}+m}} \big(A(t) - D'(t)\big) wdt = \int_{t_{\bar{k}}}^{t_{\bar{k}+m}} \big(D'(t) - D(t)\big)\, wdt$$
$$= w \sum_{m'=0}^{m-1} \Big((t_{\bar{k}+m'+1} - t_{\bar{k}+m'}) \big(Ic - \sum_{m''=0}^{m'} i_{\bar{k}+m''} \big) \Big) > 0\,.$$

Then we investigate the difference in the energy cost between the optimal and the alternate solution $\sum_{m'=0}^{m} f_{i_{k+m'}} - f_I - f_{i'_{\bar{k}+m}}$. Lemma 1 indicates that $\sum_{m'=0}^{m-1} f_{i_{k+m'}} \geq f_{\bar{\iota}_k}$ where $\bar{\iota}_k \coloneqq \sum_{m'=0}^{m-1} i_{\bar{k}+m'}$. Next, Lemma 2 indicates that $f_{\bar{\iota}_k} + f_{i_{k+m}} \geq f_I + f_{i'_{\bar{k}+m}}$. This indicates $\sum_{m'=0}^{m} f_{i_{k+m'}} - f_I - f_{i'_{\bar{k}+m}} \geq 0$. With this, the objective value of the optimal solution is always strictly greater than that of the alternate solution, which is a contradiction. This completes the proof.□

Theorem 2. *In an optimal dispatch solution* $\{t_k, i_k\}_{\forall k \in \mathcal{K}}$ *to problem* (2) ~ (8) *with arrival curve $A(t)$ and departure curve $D(t)$, if $A(t_k) - D(t_{k-1}) > Ic$, then $t_k - t_{k-1} = \underline{h}, \forall k \in \mathcal{K}$.*

Proof. This theorem will be proven by contradiction. If the condition in this theorem does not hold, then in this optimal solution, there exists a $\bar{k} \in \mathcal{K}$ with $t_{\bar{k}} - t_{\bar{k}-1} > \underline{h}$ and $A(t_{\bar{k}}) - D(t_{\bar{k}-1}) > Ic$, as Case 2 in Fig. 18 shows. With this, as Case 3 in Fig. 18 shows, we can construct an alternate feasible solution $\{t'_k, i'_k\}_{\forall k \in \mathcal{K} \setminus \{\bar{k}+1,\cdots,\bar{k}+m-1\}}$ such that $t'_{\bar{k}} < t_{\bar{k}}$, $t'_k = t_k$ for all other k indexes and $i'_k = i_k$ for all k indexes. Obviously, the energy cost remains the same since the number of dispatches are exactly the same. As to the waiting cost, we only need to make a comparison in the interval $[t_{\bar{k}-1}, t_{\bar{k}}]$. Denote the cumulative departure curve of Case 3 as $D''(t)$, and we can obtain

$$\int_{t_{\bar{k}-1}}^{t_{\bar{k}}} \left(A(t) - D'(t)\right) - \int_{t_{\bar{k}-1}}^{t_{\bar{k}}} (A(t) - D''(t))\, w dt = (t_{\bar{k}} - t_{\bar{k}}{}')\, Ic > 0.$$

Thus, the objective value of the optimal solution is greater than that of the alternate solution, which is a contradiction. This completes the proof. □

Lemma 3. *Denote problem* (2) ~ (8) *as the original problem (OP), and the same problem where the original arrival curve* $A(t)$ *is replaced with virtual arrival curve* $B(t)$ *as the revised unsaturated problem (RUP). Then the feasible regions of the OP and RUP are the same. For any feasible solution* $\mathbf{s} \coloneqq \{t_k, i_k\}_{\forall k \in \mathcal{K}}$, *OP and RUP have the same dispatch curve, and the objective values of OP and RUP, respectively denoted as* $OP(\mathbf{s})$ *and* $RUP(\mathbf{s})$, *are always separated by a constant difference:*

$$OP(\mathbf{s}) - RUP(\mathbf{s}) = W(A) \coloneqq w \int_0^T \left(A(t) - B(t)\right) dt.$$

Proof. First, given a feasible dispatch solution $\mathbf{s} \coloneqq \{t_k, i_k\}_{\forall k \in \mathcal{K}}$, to OP, we will use induction to show that the OP and RUP have the same departure curve. Let $D(t)$ and $D^B(t)$ denote the departure curves corresponding to arrival curves $A(t)$ and $B(t)$, respectively.

Base case: When $z = 1$, $A(t) = B(t)$, thus $D(t) = D^B(t), \forall t \in [\mathcal{E}(0), \mathcal{O}(1))$ for the same dispatch solution.

Induction step: Assume $D(t) = D^B(t), \forall t \in [\mathcal{E}(0), \mathcal{O}(z))$ for a $0 \le z \le |\mathcal{O}|$. Then for the oversaturated period $[\mathcal{O}(z), \mathcal{E}(z))$, the facts that $D^B(t) \le B(t)$ and $B'(t) = Ic/\underline{h}$ yield that at any dispatch point $t_k \in [\mathcal{O}(z), \mathcal{E}(z))$, if exists, then $B(t_k) - D^B(t) \ge Ic$. Thus $D^B(t)$ always grows $i_k c$ at t_k. Since $A(t) \ge B(t)$, similarly, we obtain that $D(t)$ also grows to $i_k c$ at t_k. In this way, $D(t) = D^B(t), \forall t \in [\mathcal{E}(0), \mathcal{E}(z))$. Further, since $A(t) = B(t)$ during unsaturated period $[\mathcal{E}(z), \mathcal{O}(z+1))$ (or $[\mathcal{E}(|\mathcal{O}|), \mathcal{O}(|\mathcal{O}|)]$ for the last unsaturated period), then we obtain $D(t) = D^B(t), \forall t \in [\mathcal{E}(0), \mathcal{O}(z+1))$.

The above induction proves that a feasible solution to OP is feasible to RUP and yields the same departure curve. With the same induction, we can show that a feasible solution to RUP is feasible to OP and yields the same departure curve as well. With this, the formulation of objective function (2) yields $OP(\mathbf{s}) - RUP(\mathbf{s}) = w \int_0^T (A(t) - B(t))\, dt$. □

Theorem 4. *An optimal dispatch solution* $\{t_k, i_k\}_{\forall k \in \mathcal{K}}$ *to RUP satisfies: (i) if* $t_k - t_{k-1} > \underline{h}$, *then* $B(t_k) - D(t_k) = 0, \forall k \in \mathcal{K}$; *and (ii) if* $t_k - t_{k-1} = \underline{h}$, *then* $B(t_k) - D(t_k) \in [0, c), \forall k \in \mathcal{K}$, *where* $D(t)$ *is the corresponding departure curve.*

Proof. We first prove $B(t_k) - D(t_k) \in [0, c), \forall k \in \mathcal{K}$ holds for both cases by induction.

Base case: First investigate the base case with $k = 1$. When $B(t_1) - D(t_1) > 0$, if $i_1 = I$, then Theorem 2 indicates that $t_1 \le \underline{h}$, which contradicts to $B(t)$ being unsaturated. Therefore, $B(t_1) - D(t_1) = 0$ if $i_1 = I$. Otherwise, if $i_1 < I$ and $B(t_1) - D(t_1) \ge c$, then we can raise the first dispatch to $i_1 + 1$ and drop the second dispatch to $i_2 - 1$. Then with similar analysis in Theorem 1, we can show that this will always strictly reduce the objective value, which obviously contradicts to this solution being optimal. Therefore $B(t_1) - D(t_1) \in [0, c)$ holds for the base case.

Induction step: Assume that for a $k \in \mathcal{K} \setminus \{K\}$, $B(t_k) - D(t_k) \in [0, c)$. Then at time t_{k+1}, when $B(t_{k+1}) - D(t_{k+1}) > 0$, if $i_k = I$, Theorem 2 indicates that $t_{k+1} - t_k = \underline{h}$, and thus $B(t_{k+1}) - D(t_{k+1}) = [B(t_k) - D(t_k)] + [B(t_{k+1}) - B(t_k)] - Ic < c + [B(t_{k+1}) - B(t_k)] - Ic$ (with the induction assumption) $< c$ (since $B(t)$ is unsaturated any time). Next, consider the case when $i_{k+1} < I$. If $B(t_{k+1}) - D(t_{k+1}) \ge c$, then $k + 1 < |\mathcal{K}|\}$ due to constraints (6). Then with similar analysis in Theorem 1, we can show that raising i_{k+1} to $i_{k+1} + 1$ and dropping i_{k+2} to $i_{k+2} - 1$ will always strictly reduce the objective value, which obviously contradicts to this solution being optimal. Therefore, $B(t_{k+1}) - D(t_{k+1}) \in [0, c)$ also holds.

The above induction shows that $B(t_k) - D(t_k) \in [0, c)$ holds for both cases in an optimal solution. Next we further prove that $B(t_k) - D(t_k) = 0, \forall k \in \mathcal{K}$ holds if $t_k - t_{k-1} > \underline{h}$ by contradiction. If the condition does not hold, then in this optimal solution, there exists a $\bar{k} \in \mathcal{K}$ with $t_{\bar{k}} - t_{\bar{k}-1} > \underline{h}$ and $B(t_{\bar{k}}) - D(t_{\bar{k}}) > 0$. Then, we can construct an alternate solution $\{t'_k, i'_k\}_{\forall k \in \mathcal{K}}$ such that $t'_{\bar{k}} < t_{\bar{k}}$, $t'_k = t_k$ for all other k indexes and $i'_k = i_k$ for all k indexes, and there are two cases. If $t'_{\bar{k}} - t'_{\bar{k}-1} > \underline{h}$ can be satisfied, as Fig. 19 (a) shows, we can always find a $t'_{\bar{k}}$ such that $B(t'_{\bar{k}}) - D(t'_{\bar{k}}) = 0$. With similar analysis in Theorem 2, we can show that this will always strictly reduce the objective value, which obviously contradicts to this solution being optimal. Otherwise, if $t'_{\bar{k}} - t'_{\bar{k}-1} = \underline{h}$, as Fig. 19 (b) shows, this reduces to the first case and we have shown that $B(t_k) - D(t_k) \in [0, c), \forall k \in \mathcal{K}$ holds if $t_k - t_{k-1} = \underline{h}$.

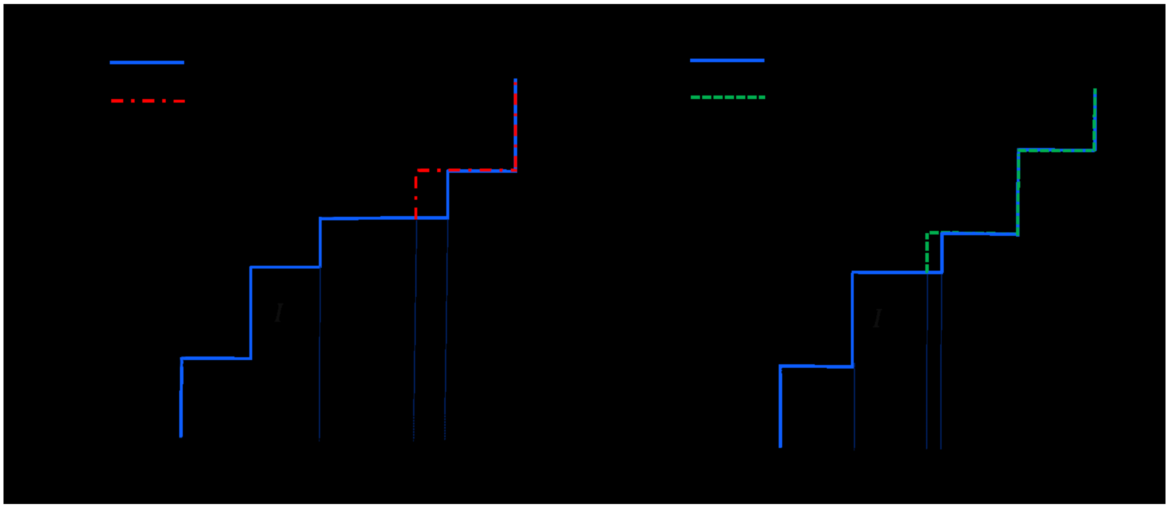

Fig. 19: An illustrative example

This completes the proof. □

Appendix B. Calibration of the energy cost function in Case study 1

Since there are not empirical data from the Beijing Metro System to calibrate the parameters in this function, we combine the data obtained from a field experiment in Shenzhen Subway System (Jian, 2017) and the traction energy consumption data from Batong line to estimate these parameters. The field experiment was conducted on Line 7 in Shenzhen Subway System to test the energy consumption of different train lengths (3 and 6 carriages) under three load conditions, denoted as AW0, AW2 and AW3. The unit-carriage traction and auxiliary energy consumption are shown in Table 2. Besides, the traction energy consumption of Batong line is 160.99 kWh/(100 carriage km) and the electricity price is 0.13 \$/(kWh). With these data, we estimate the parameters as follows.

(1) Compute the average traction and auxiliary energy consumption over three load conditions of the 3- and 6-carriage trains on Line 7, respectively.

(2) Divide the traction energy of Batong line by the average traction energy of Line 7, which results in the traction ratio. Then for both train formations, multiply the average traction energy of Line 7 by the traction ratio, which results in the unit-carriage traction energy of Batong line. Compute the unit-carriage auxiliary energy of Batong line for both train formations in a similar way.

(3) Sum up the unit-carriage traction and auxiliary energy consumption of Batong Line for both train formations respectively, which results in the unit-carriage energy consumption of Batong Line for both train formations.

(4) Estimate the total energy consumption of both train formations for Batong line by multiplying the number of carriages and its unit-carriage energy consumption.

(5) Apply the energy consumption and the corresponding carriage number into the dispatch energy cost function,

which yields two equations with 3 parameters (i.e. C^F, C^V, α). Arbitrarily set a value to α and we can solve these equations. Here, we set $\alpha = 0.5$ and obtain C^F=2.049 $, C^V=5.56 $.

Table 2 Results of the field experiment in Line 7, Shenzhen Subway System

Energy consumption (kWh/carriage)	3-carraige		6-carriage	
	Traction energy	Auxiliary energy	Traction energy	Auxiliary energy
AW0	1.77	0.40	2.28	0.43
AW2	2.12	0.60	2.92	0.83
AW3	2.71	0.73	3.13	1.03
Average	2.20	0.58	2.78	0.76

References

Albrecht, T. 2009. Automated timetable design for demand-oriented service on suburban railways. Public Transport, 1, 5-20.

Ackerman, E. 2016. nuTonomy to test world's fully autonomous taxi service in Singapore this year [Online]. IEEE Spectrum. Available: https://ac.els-cdn.com/S2352146517303009/1-s2.0-S2352146517303009-main.pdf?_tid=cf1f5970-94c3-45ee-8c6c-2c590788dd4f&acdnat=1523207874_27c1bc8acdb7a5bd861a99342a37db00 [Accessed 08-April 2018].

Adamski, A., & Turnau, A. 1998. Simulation support tool for real-time dispatching control in public transport. Transportation Research Part A: Policy and Practice, 32(2), 73-87.

Ansari, S., Başdere, M., Li, X., Ouyang, Y., & Smilowitz, K. 2017. Advancements in continuous approximation models for logistics and transportation systems: 1996–2016. Transportation Research Part B: Methodological.

Barnett, A., & Kleitman, D. J. 1973. Optimal scheduling policies for some simple transportation systems. Transportation Science, 7(1), 85-99.

Bownman, L. J. 2018. Available: http://jbowman.net/ [Accessed 08-April 2018].

Chang, S. K., & Schonfeld, P. M. 1991. Multiple period optimization of bus transit systems. Transportation Research Part B: Methodological, 25(6), 453-478.

Ceder, A. A. 2011. Public-transport vehicle scheduling with multi vehicle type. Transportation Research Part C: Emerging Technologies, 19(3), 485-497.

Chen, H., Gu, W., Cassidy, M. J., & Daganzo, C. F. 2015. Optimal transit service atop ring-radial and grid street networks: a continuum approximation design method and comparisons. Transportation Research Part B: Methodological, 81, 755-774.

Chen Z., Li, X. & Zhou X. 2018. Operational design for shuttle systems under oversaturated traffic Part II: Discrete Modeling Method. Available:

https://www.researchgate.net/publication/325794307_Operational_Design_for_Shuttle_Systems_with_Modular_Vehicles_under_Oversaturated_Traffic_Part_II_Discrete_Modeling_Method.

Cohen, M. A., & Moon, S. (991. An integrated plant loading model with economies of scale and scope. European Journal of Operational Research, 50(3), 266-279.

Daganzo, C. F. 1997. Fundamentals of transportation and traffic operations (Vol. 30). Oxford: Pergamon.

Daganzo, C. F. 2005. Logistics systems analysis. Springer Science & Business Media.

Daganzo, C. F. 2009. A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. Transportation Research Part B: Methodological, 43(10), 913-921.

Daganzo, C. F. 2010. Structure of competitive transit networks. Transportation Research Part B: Methodological, 44(4), 434-446.

Estrada, M., Roca-Riu, M., Badia, H., Robusté, F., & Daganzo, C. F. 2011. Design and implementation of efficient transit networks: procedure, case study and validity test. Transportation Research Part A: Policy and Practice, 45(9), 935-950.

Fan, W., Mei, Y., & Gu, W. 2018. Optimal design of intersecting bimodal transit networks in a grid city. Transportation Research Part B: Methodological, 111, 203-226.

Freyss, M., Giesen, R., & Muñoz, J. C. 2013. Continuous approximation for skip-stop operation in rail transit. Procedia-Social and Behavioral Sciences, 80, 186-210.

Gao, Y., Kroon, L., Schmidt, M., & Yang, L. 2016. Rescheduling a metro line in an over-crowded situation after disruptions. Transportation Research Part B: Methodological, 93, 425-449.

Guo, Q.-W., Chow, J. Y. & Schonfeld, P. 2017. Stochastic dynamic switching in fixed and flexible transit services as market entry-exit real options. Transportation Research Part C: Emerging Technologies.

Guo, X., Sun, H., Wu, J., Jin, J., Zhou, J., & Gao, Z. 2017. Multiperiod-based timetable optimization for metro transit networks. Transportation Research Part B: Methodological, 96, 46-67.

HART. 2018. Maps and schedules [online]. HART. Available: http://www.gohart.org/Pages/maps-schedules.aspx [accessed 12-June 2018].

Hassold, S., & Ceder, A. 2012. Multiobjective approach to creating bus timetables with multiple vehicle types. *Transportation Research Record: Journal of the Transportation Research Board*, (2276), 56-62.

Hassold, S., & Ceder, A. A. 2014. Public transport vehicle scheduling featuring multiple vehicle types. *Transportation Research Part B: Methodological*, *67*, 129-143.

Holmberg, K., & Tuy, H. 1999. A production-transportation problem with stochastic demand and concave production costs. Mathematical programming, 85(1), 157-179.

Huang, Y., Yang, L., Tang, T., Gao, Z. & Cao, F. 2017. Joint train scheduling optimization with service quality and energy efficiency in urban rail transit networks. Energy, 138, 1124-1147.

Hurdle, V. F. 1973a. Minimum Cost Schedules for a Public Transportation Route—I. Theory. Transportation Science, 7(2), 109-137.

Hurdle, V. F. 1973b. Minimum Cost Schedules for a Public Transportation Route II. Examples. Transportation Science, 7(2), 138-157.

Hurdle, V. F. 1973c. Minimum cost locations for parallel public transit lines. Transportation Science, 7(4), 340-350.

Jian, L. 2017. Available measures to energy savings in urban rail transit operations. Shenzhen Subway.

Lambert, F. 2017. A new 'trackless electric train' (aka a bus) starts testing in China [Online]. Electrek. Available: https://electrek.co/2017/10/30/trackless-electric-train-china/ [Accessed 08-April 2018].

Li, X., Ma, J., Cui, J., Ghiasi, A. & Zhou, F. 2016. Design framework of large-scale one-way electric vehicle sharing systems: A continuum approximation model. Transportation Research Part B: Methodological, 88, 21-45.

Lin, Z. & Kwan, R. S. 2016. A branch-and-price approach for solving the train unit scheduling problem. Transportation Research Part B: Methodological, 94, 97-120.

Newell, G. F. 1971. Dispatching policies for a transportation route. Transportation Science, 5, 91-105.

Newell, G. F. 1974. Control of pairing of vehicles on a public transportation route, two vehicles, one control point. Transportation Science, 8(3), 248-264.

Newell, G.F. 1982. Applications of queueing theory (Second edition). Chapman and Hall.

Niu, H. & Zhou, X. 2013. Optimizing urban rail timetable under time-dependent demand and oversaturated conditions. Transportation Research Part C: Emerging Technologies, 36, 212-230.

Niu, H., Zhou, X. & Gao, R. 2015. Train scheduling for minimizing passenger waiting time with time-dependent demand and skip-stop patterns: Nonlinear integer programming models with linear constraints. Transportation Research Part B: Methodological, 76, 117-135.

Ouyang, Y., Nourbakhsh, S. M., & Cassidy, M. J. 2014. Continuum approximation approach to bus network design under spatially heterogeneous demand. Transportation Research Part B: Methodological, 68, 333-344.

Rail, H.-S. 2017. Flexible grouped high-speed rail will be tested this year so train lengths are no longer limited to 8 or 16 carriages [Online]. Finance. [Accessed 08-April 2018].

Salzborn, F. J. 1972. Optimum bus scheduling. Transportation Science, 6(2), 137-148.

Salzborn, F. J. M. 1980. Scheduling bus systems with interchanges. Transportation Science, 14(3), 211-231.

Sheffi, Y., & Sugiyama, M. 1982. Optimal bus scheduling on a single route. Transport, 60, 68.

Shi, J., Yang, L., Yang, J., & Gao, Z. 2018. Service-oriented train timetabling with collaborative passenger flow control on an oversaturated metro line: An integer linear optimization approach. Transportation Research Part B: Methodological, 110, 26-59.

Statistical Atlas. 2017. Household income in Tampa, Florida [online]. Available: https://statisticalatlas.com/place/Florida/Tampa/Household-Income [Accessed 12-June 2018].

Sun, L., Jin, J. G., Lee, D.-H., Axhausen, K. W. & Erath A. 2014. Demand-driven timetable design for metro services. Transportation Research Part C: Emerging Technologies, 46, 284-299.

Tarek, F. 2018. Dubai tests autonomous pods in drive for smart city [Online]. Reuters. Available: https://www.reuters.com/article/us-emirates-transportation-autonomous/dubai-tests-autonomous-pods-in-drive-for-smart-city-idUSKCN1GD5G6 [Accessed 08-April 2018].

Wang, Y., Tang, T., Ning, B., Van Den Boom, T. J. & de Schutter, B. 2015. Passenger-demands-oriented train scheduling for an urban rail transit network. Transportation Research Part C: Emerging Technologies, 60, 1-23.

Wirasinghe, S.C., 1990.Re-Examination of Newell's dispatching policy and extension to a public bus route with many to many time-varying demand, in Transportation and Traffic Theory (Ed. Koshi, M.), Elsevier, 1, 363-377

Xu, Y., Jia, B., Ghiasi, A. & Li, X. 2017. Train routing and timetabling problem for heterogeneous train traffic with switchable scheduling rules. Transportation Research Part C: Emerging Technologies, 84, 196-218.

Yang, X., Chen, A., Ning, B., & Tang, T. 2016. A stochastic model for the integrated optimization on metro timetable and speed profile with uncertain train mass. Transportation Research Part B: Methodological, 91, 424-445.

Yi, Z. & Shirk, M. 2018. Data-driven optimal charging decision making for connected and automated electric vehicles: A personal usage scenario. Transportation Research Part C: Emerging Technologies, 86, 37-58.

Yin, J., Yang, L., Tang, T., Gao, Z. & Ran, B. 2017. Dynamic passenger demand oriented metro train scheduling with energy-efficiency and waiting time minimization: Mixed-integer linear programming approaches. Transportation Research Part B: Methodological, 97, 182-213.

Zhao, N., Roberts, C., Hillmansen, S., Tian, Z., Weston, P. & Chen, L. 2017. An integrated metro operation optimization to minimize energy consumption. Transportation Research Part C: Emerging Technologies, 75, 168-182.

Zhou, W. & Teng, H. 2016. Simultaneous passenger train routing and timetabling using an efficient train-based Lagrangian relaxation decomposition. Transportation Research Part B: Methodological, 94, 409-439.