# On the design of an optimal flexible bus dispatching system with modular bus units: Using the three-dimensional Macroscopic Fundamental Diagram

Igor Dakic[a], Kaidi Yang[b,*], Monica Menendez[c], Joseph Y. J. Chow[d]

[a]*Traffic Engineering Group, Institute for Transport Planning and Systems, ETH Zurich, Switzerland*
[b]*Autonomous Systems Lab, Stanford University, USA*
[c]*Division of Engineering, New York University Abu Dhabi, United Arab Emirates*
[d]*C2SMART University Transportation Center, New York University, USA*

**Abstract**

This study proposes a novel *flexible bus dispatching system* in which a fleet of fully automated modular bus units, together with conventional buses, serves the passenger demand. These modular bus units can either operate individually or combined (forming larger modular buses with a higher passenger capacity). This provides enormous flexibility to manage the service frequencies and vehicle allocation, reducing thereby the operating cost and improving passenger mobility.

We develop an optimization model used to determine the optimal composition of modular bus units and the optimal service frequency at which the buses (both conventional and modular) should be dispatched across each bus line. We explicitly account for the dynamics of traffic congestion and complex interactions between the modes at the network level based on the recently proposed three-dimensional macroscopic fundamental diagram (3D-MFD). To the best of our knowledge, this is the first application of the 3D-MFD and modular bus units for the frequency setting problem in the domain of bus operations.

Numerical results show the improvements in the total system cost made by adjusting the number of combined modular bus units and their dispatching frequencies to the evolution of both, the car and the bus passenger demand. A comparison with the commonly used approach that considers only the bus system (neglecting the complex multimodal interactions and congestion propagation) reveals the value of the proposed modeling framework. Finally, a sensitivity analysis of the effect of the operating unit cost of modular bus units, the size of the bus network, and the size of the bus fleet sheds light on the robustness of the proposed optimization framework.

*Keywords:* Optimal bus dispatching system; Modular bus units; Frequency setting problem; 3D-MFD

## 1. Introduction

In the past few decades, urbanization has become a prevalent trend for many places around the globe. As cities continue to grow, they require substantial infrastructure resources, placing various burdens to society and economic development. To address the social, economic, and environmental challenges associated with rapid urbanization, it is crucial to provide a sustainable transportation system. This is because transportation connects people and goods for social and economic interactions (Krugman, 1991).

Public transport is considered to be a crucial aspect of a sustainable urban development, as it allows more passengers to efficiently travel across an urban area (Roca-Riu et al., 2020). Nevertheless, to be attractive and a suitable alternative for the car owners, public modes of transport need to provide a good level of service. This can be achieved in many different ways, e.g. through a proper network design, frequency setting, timetable design, fleet assignment, and crew assignment. Due to their high complexity, the optimization of each of these stages is usually performed

*Corresponding author. Tel.: +1 650 885 1186
*Email addresses:* i.dakic@ivt.baug.ethz.ch (Igor Dakic), kaidi.yang@stanford.edu (Kaidi Yang), monica.menendez@nyu.edu (Monica Menendez), joseph.chow@nyu.edu (Joseph Y. J. Chow)

sequentially and is considered as a separate problem, influencing the decisions taken at the subsequent stage (Ceder, 2007; Desaulniers and Hickman, 2007). The decisions can be made for different planning horizons, depending on whether they are strategical, tactical, or operational (Martínez et al., 2014).

One of the prominent research topics in the public transport community, more specifically in the domain of bus operations, is the frequency setting problem. In this problem, the time intervals between consecutive buses need to be determined in a way that they can satisfy the total passenger demand generated during the planning horizon without a disproportionate increase in operator costs. As such, the problem belongs to the group of tactical problems, related to improving the level of service and reducing the operating cost of buses. Methods used to find the optimal bus frequency are usually based on minimizing the passenger waiting time, transfer time, total passenger cost, total operator cost, or a combination of them. Existing studies on this topic, up to now, rarely consider the impact of traffic conditions on the optimal bus frequency or the effects that the optimal bus frequency might have on traffic (Dakic et al., 2020).

To ensure that the determined bus frequency actually leads to an efficient bus service, it is crucial to account for the impact of traffic conditions, especially in case of mixed traffic (Loder et al., 2019a). With the emerging concepts related to the macroscopic modeling of large-scale networks, we can now study urban traffic dynamics in a parsimonious way. These concepts are mostly based on the Macroscopic Fundamental Diagram (MFD) (Daganzo, 2007; Geroliminis and Daganzo, 2008), also known as the Network Fundamental Diagram (Mahmassani et al., 2013). The MFD has been shown to be a useful and elegant tool to determine the current state of traffic in an urban transportation network. It links the vehicle accumulation and the travel production with a well-defined and reproducible curve. Although initially the MFDs were studied within the context of uni-modal (i.e. car) traffic, recent research efforts have resulted in the development of a three-dimensional MFD (3D-MFD), applicable for bi-modal traffic (Geroliminis et al., 2014; Loder et al., 2017, 2019b; Dakic et al., 2020). The 3D-MFD offers new ways to analyze complex interactions between different transportation modes, i.e. cars and buses (Dakic and Menendez, 2018), and has been mainly used to develop efficient perimeter control schemes for bi-modal urban networks (Ampountolas et al., 2017; Chiabaut et al., 2018; He et al., 2019; Dakic et al., 2019; Yang et al., 2019). The potential to apply the 3D-MFD for the purpose of capturing traffic conditions for the frequency setting problem has not been explored whatsoever.

When determining the optimal bus frequency, one of the most constraining parameters is the available vehicle fleet. The vehicle fleet can consist not only of conventional buses, but also of modular bus units that can either operate individually or combined together (forming thereby a single modular bus of a higher passenger capacity). In railway systems, for example, the vehicle fleet corresponds to a stock of cars and locomotives (commonly referred to as train units) that, when combined, form a single train. The number of train units assigned to a train is, in most cases, determined according to the predefined dispatched frequency of trains and the level of passenger demand (Cordeau et al., 2001; Lingaya et al., 2002; Fioole et al., 2006; Cacchiani et al., 2010). Only recently did researchers investigate the potential for combining the allocation of train units and the optimization of the dispatched frequency (Cadarso et al., 2013; Yue et al., 2017; Wang et al., 2018; Chen et al., 2019).

Recent advances in vehicle technology have opened new opportunities to apply similar concepts (i.e. combining and splitting of vehicle units) also for the bus systems (Guo et al., 2018; Chen et al., 2019, 2020). Such concept, herein called *flexible bus dispatching system*, can be especially useful for public transport operators, as it offers new perspectives and enormous flexibility to better manage the allocation of the vehicle resources and reduce the operating cost. For example, rather than dispatching a bus with a high passenger capacity in case of low passenger demand, an operator can send a group of few combined modular and fully automated bus units that, overall, has lower passenger capacity (increasing thereby the passenger occupancy) and lower operating cost. Furthermore, as the bus units are fully automated, there is no cost for assigning bus drivers to them.

Building on the knowledge of the 3D-MFD and the advanced automated vehicle technology, in this study we propose such a flexible bus dispatching system and provide a framework on how to model it. In particular, we develop an optimization model to determine the optimal number of combined automated bus units and the optimal frequency at which the units should be dispatched across different bus lines, while accounting for the traffic dynamics at the network level. Our goal is to investigate whether such an optimization model can maintain (or even reduce) the operating cost, while increasing the level of service, and vice-versa. To the best of our knowledge, this is the first application of the 3D-MFD and modular bus units for the frequency setting problem in the domain of bus operations.

Overall, the contributions of this research are fourfold. First, we propose a flexible bus dispatching system that allows to dynamically combine and split modular and fully automated bus units for the purpose of improving the

operation of the bus system. Second, we apply the 3D-MFD concept for the frequency setting problem to capture the complex multimodal interactions and congestion propagation, accounting for both vehicle and passenger dynamics. As such, the proposed modeling framework can also be applied to the existing bus dispatching system consisting only of conventional buses, addressing the limitations of the current state of the art regarding the modeling of traffic dynamics. Third, we develop an optimization framework that jointly determines the optimal number of combined modular bus units and the optimal frequency at which the units, both conventional and modular, should be dispatched across different bus lines. The proposed optimization model accounts for the effects of the dispatching policy on the operation and travel cost of both modes. Fourth, we study the effect of the operating unit cost of modular bus units, the size of the bus network, and the size of the bus fleet, testing the robustness of the proposed optimization framework.

The remainder of this paper is organized as follows. Section 2 provides an overview of the current state of the art. Section 3 describes in detail the proposed methodological framework used to determine the optimal number of combined automated bus units and the optimal frequency at which the units should be dispatched. In Section 4, we quantify the improvements acquired by the proposed system in comparison to that with only conventional buses, and show the value of the proposed modeling approach based on the 3D-MFD. In addition, we conduct a sensitivity analysis testing the robustness of the proposed optimization framework. Concluding remarks and future research directions are given in Section 5.

## 2. Literature review

While a comprehensive literature review on the frequency setting problem can be found in Ibarra-Rojas et al. (2015), for the reader's convenience, here we review some of the most relevant studies, along with the recent works that have been published since then.

Early studies on the optimal bus frequency were based on analytical models (Newell, 1971; Salzborn, 1972; Schéele, 1980; Han and Wilson, 1982) or heuristic methods (Furth and Wilson, 1981), and were usually formulated with the aid of graph theory, with nodes and arcs representing bus stops and street segments, respectively.

Vuchic et al. (1978) introduced the maximum load section-based method to determine the service frequency that can meet the passenger demand. Following a similar approach, Ceder and Wilson (1986) established four alternative methods based on the passenger count data. The objectives were twofold: to minimize the fleet size and to meet the passenger demand. The minimum fleet size was also used as the objective function by Salzborn (1972), who derived a continuous model for the optimal bus frequency in case of both, single-bus routes (Salzborn, 1972) and multiple-bus routes (Salzborn, 1980). Hadas and Shnaiderman (2012) and Li et al. (2013) studied the frequency setting problem with stochasticity in the demand, arrival times, boarding/alighting times, and travel times. Using a non-linear program to formulate the problem, Verbas and Mahmassani (2013) and Verbas et al. (2015) provided an optimal allocation of resources over space and time, while minimizing the weighted sum of ridership and wait time savings. Recent work by Gkiotsalitis and Cats (2018) determined the optimal bus frequency based on a travel time variability parameter.

Overall, the previous studies have rarely addressed the frequency setting problem under varying traffic conditions, especially in case of mixed traffic. To the best of our knowledge, the only attempt made in this direction can be found in Zhang and Liu (2019), who proposed a responsive bus dispatching strategy based on a doubly dynamical approach (as in Liu and Geroliminis, 2017). Although that study investigated a time-dependent bus dispatching problem in a bi-modal network, it only considered a simple network with two main directions, without analyzing possible effects of the dispatched frequency on the passenger dynamics across different bus stops, and how the available vehicle fleet should be optimally distributed across different bus lines during the planning horizon. Moreover, no research has been conducted regarding the concept of combining and splitting modular vehicle units along fixed bus lines. While some studies can be found in the scientific literature for railway systems (Cordeau et al., 2001; Lingaya et al., 2002; Fioole et al., 2006; Cacchiani et al., 2010; Cadarso et al., 2013; Yue et al., 2017; Cats and Haverkamp, 2018; Wang et al., 2018) and more recently for the metro system (Chen et al., 2019, 2020), strategies developed for railway systems are not readily applicable to bus systems. This is because the railway system represents a closed environment, where trains do not interact with other modes. For the bus system, however, we need to consider the complex interactions between buses and cars.

This study aims to fill these research gaps by: (i) proposing a novel, flexible bus dispatching system, in which the bus fleet consists of modular and fully automated bus units that can either be combined together or operate individually; and (ii) jointly optimizing the allocation of the modular bus units and their dispatching frequencies across

different bus lines. To this end, we develop a macroscopic modeling framework to account for the traffic dynamics and interactions between buses and cars, based on the 3D-MFD.

## 3. Methodological framework

In this section, we present all the elements of the methodological framework, including the network representation based on the 3D-MFD (Section 3.1), the proposed optimization model (Section 3.2), and the macroscopic modeling approach from both vehicular (Sections 3.3) and passenger (Section 3.4) perspective. For readers' convenience, Table 1 provides the list of the most important notation used in this paper.

### *3.1. System description and network representation*

We consider here a bi-modal urban network consisting of buses and cars denoted as $b$ and $c$, respectively. A bus operator manages a set of bus lines $\mathcal{L}$, indexed by $l$, with two types of buses: (i) conventional buses denoted as $r$; and (ii) modular buses denoted as $m$. The bus type is indexed by $i \in \mathcal{I} = \{r, m\}$. Each modular bus can include $u \in \mathcal{U}_{m,l} = \{1, ..., |\mathcal{U}_{m,l}|\}$ modular bus units as shown in Fig. 1, where $|\mathcal{U}_{m,l}|$ is the maximum number of modular bus units that can be combined as a single modular bus on line $l$. Note that the modular bus units are fully automated, thus there is no cost for assigning bus drivers to them (similarly to the Next Future Mobility system[1]). To simplify the notation, each conventional bus is considered as an individual conventional bus unit, i.e. $|\mathcal{U}_{r,l}| = 1,\ \forall l \in \mathcal{L}$ (see Fig. 1). In addition, all units of the same type $i$ are considered to be equivalent in terms of operational characteristics (e.g. energy consumption, vehicle dynamics, etc.) and passenger capacity $C_i$, such that the capacity of the conventional bus unit is $C_r$ and the capacity of the modular bus unit is $C_m \leq C_r$. This, in turn, implies that the capacity of the conventional bus is also $C_r$, whereas the capacity of a modular bus consisting of $u$ modular bus units is $u \cdot C_m$. Finally, we assume that each conventional bus is equivalent to $\zeta \geq 1$ modular bus units (in terms of capacity), i.e. $C_r = \zeta \cdot C_m$. The size of the bus fleet $F$ can therefore be expressed either in terms of the equivalent number of conventional ($F_r + F_m/\zeta$) or modular ($F_r \cdot \zeta + F_m$) bus units. This gives the following formulation for the penetration rate of each type of bus units, $p_i$, defined as the share of the number of bus units of a given type $i$ within the total number of equivalent bus units, i.e.:

$$p_i = \begin{cases} (F_r \cdot \zeta)/(F_r \cdot \zeta + F_m), & \text{if } i = r, \\ F_m/(F_r \cdot \zeta + F_m), & \text{if } i = m. \end{cases} \tag{1}$$

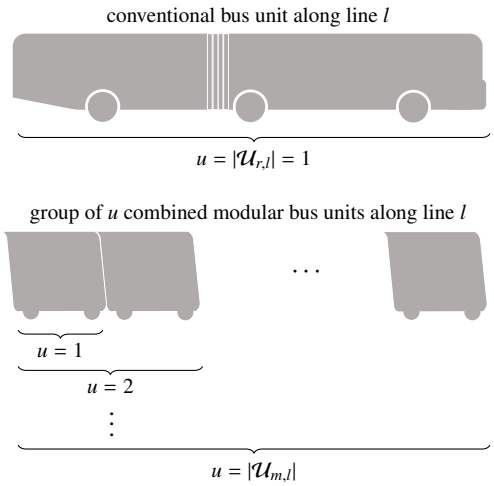


Figure 1: Schematic representation of the types of bus units that form the bus fleet in the proposed flexible bus dispatching system.

[1] www.next-future-mobility.com

Table 1: Nomenclature.

| | |
|---|---|
| $\mathcal{L}$ | set of bus lines indexed by $l$ |
| $\mathcal{S}_l$ | set of segments along line $l$ indexed by $s$ |
| $\mathcal{I}$ | set of types of bus units indexed by $i$: conventional unit ($i = r$); modular unit ($i = m$) |
| $\mathcal{T}$ | set of time intervals indexed by $t$ |
| $\tau$ | length of time interval |
| $\Psi_{b,\min}$ | minimum dispatching flow of buses of each type |
| $\Psi_{b,\max}$ | maximum dispatching flow of buses of each type |
| $F_i$ | total number of available bus units of type $i$ |
| $C_i$ | passenger capacity of a bus unit of type $i$ |
| $\pi_i$ | operating cost of a bus unit of type $i$ |
| $\eta_i$ | bus unit-car equivalent for a bus unit of type $i$ |
| $\lvert U_{i,l}\rvert$ | maximum number of bus units of type $i$ that can be combined along line $l$ |
| $\Omega$ | average value of time |
| $L_b$ | length of the bus network |
| $L_c$ | length of the car network |
| $\ell_b$ | average vehicular trip length for the bus mode |
| $\ell_c$ | average vehicular trip length for the car mode |
| $\ell_{b,\text{pax}}$ | average passenger trip length for the bus mode |
| $v_{\max}$ | free-flow speed of buses and cars |
| $v_b(t)$ | average bus speed in the network during interval $t$ |
| $v_c(t)$ | average car speed in the network during interval $t$ |
| $Q_b(t)$ | circulating flow for the bus mode given by the 3D-MFD during interval $t$ |
| $Q_c(t)$ | circulating flow for the car mode given by the 3D-MFD during interval $t$ |
| $G_b(t)$ | trip completion flow for the bus mode given by the 3D-MFD during interval $t$ |
| $G_c(t)$ | trip completion flow for the car mode given by the 3D-MFD during interval $t$ |
| $\lambda_{c,\text{int}}(t)$ | internal (originating from within the network) car demand during interval $t$ |
| $\lambda_{c,\text{ext}}(t)$ | external (going into the network) car demand during interval $t$ |
| $\overline{N}_c$ | maximum car accumulation in the subnetwork with car only and mixed segments |
| $N_b(t)$ | bus accumulation in the entire network at the beginning of interval $t$ |
| $N_c(t)$ | car accumulation in the entire network at the beginning of interval $t$ |
| $N_{b,\text{mix}}(t)$ | bus accumulation in the subnetwork with mixed segments at the beginning of interval $t$ |
| $N_{b,\text{ded}}(t)$ | bus accumulation in the subnetwork with dedicated segments at the beginning of interval $t$ |
| $\ell_{b,l,s}$ | length of bus segment $s$ of line $l$ |
| $\ell'_{b,l,s}$ | average distance between the bus stops on $s$ of line $l$ |
| $\Theta_{i,l,s}(t)$ | dwell time of buses of type $i$ on segment $s$ of line $l$ during interval $t$ |
| $\Theta_{l,s}(t)$ | average dwell time on segment $s$ of line $l$ during interval $t$ |
| $v_{b,l,s}(t)$ | bus speed on segment $s$ of line $l$ during interval $t$ |
| $\varphi_{c,l,s}(t)$ | receiving flow of cars to segment $s$ of line $l$ during interval $t$ |
| $\psi_{b,i,l}(t)$ | dispatched flow of bus units of type $i$ along line $l$ during interval $t$ |
| $\Psi_{b,i,l}(t)$ | dispatched flow of buses of type $i$ along line $l$ during interval $t$ |
| $\phi_{b,i,l,s}(t)$ | transferring flow of bus units of type $i$ between segments $s$ and $s+1$ of line $l$ during interval $t$ |
| $\Phi_{b,i,l,s}(t)$ | transferring flow of buses of type $i$ between segments $s$ and $s+1$ of line $l$ during interval $t$ |
| $\overline{N}_{c,l,s}(t)$ | maximum car accumulation that can be accommodated on segment $s$ of line $l$ at the beginning of interval $t$ |
| $N_{c,l,s}(t)$ | accumulation of cars on segment $s$ of line $l$ at the beginning of interval $t$ |
| $N_{b,i,l,s}(t)$ | accumulation of buses of type $i$ on segment $s$ of line $l$ at the beginning of interval $t$ |
| $n_{b,i,l,s}(t)$ | accumulation of bus units of type $i$ on segment $s$ of line $l$ at the beginning of interval $t$ |
| $u_{i,l,s}(t)$ | average number of combined units of type $i$ on segment $s$ of line $l$ at the beginning of interval $t$ |
| $B_{l,s}(t)$ | total number of boarding passengers on segment $s$ of line $l$ during interval $t$ |
| $A_{l,s}(t)$ | total number of alighting passengers on segment $s$ of line $l$ during interval $t$ |
| $P_{l,s,s'}(t)$ | number of on-board passengers on segment $s$ with destination on segment $s'$ of line $l$ at the beginning of interval $t$ |
| $R_{l,s,s'}(t)$ | trip completion flow of bus passengers on segment $s$ with destination on segment $s'$ of line $l$ during interval $t$ |
| $D_{l,s,s'}(t)$ | number of boarding passengers on segment $s$ with destination on segment $s'$ of line $l$ during interval $t$ |
| $O_{l,s,s'}(t)$ | outflow of passengers between segments $s$ and $s+1$ with destination on segment $s'$ of line $l$ during interval $t$ |
| $\omega_{l,s,s'}(t)$ | number of unserved passengers on segment $s$ with destination on segment $s'$ of line $l$ at the beginning of interval $t$ |
| $\lambda_{l,s,s'}(t)$ | average arrival rate of bus passengers on segment $s$ with destination on segment $s'$ of line $l$ during interval $t$ |
| $\delta_{l,s,s'}(t)$ | share of passengers on segment $s$ with destination on segment $s'$ of line $l$ during interval $t$ |
| $\rho_{b,i,l,s,s'}(t)$ | occupancy of a bus unit of type $i$ of passengers on segment $s$ with destination is on segment $s'$ of line $l$ |
| $\rho_{b,l,s,s'}(t)$ | average bus occupancy of passengers on segment $s$ with destination is on segment $s'$ of line $l$ |

The bus operator is interested in dynamically determining the bus type and the dispatching frequency for each bus line to optimize its utility (defined in the Section 3.2) based on real-time or historical demand information. The decision should be made for each time interval $t \in \mathcal{T}$ of length $\tau$. We make the following assumption on the demand information to which the operator has access.

**Assumption 1** (Data provision)**.** We assume that the bus operator has access to demand and routing information of bus passengers at the bus stop level and car demand at the network level.

**Remark.** We make the following remarks regarding Assumption 1. First, the bus operator can predict detailed information (demand and routing) on bus passengers through historical data (e.g. from smart card data, surveillance cameras inside buses, or passenger surveys) or real-time data (e.g. from smart phones). Second, detailed information on car traffic (such as OD pairs and route choices of car users) is typically not available to the bus operator due to privacy concerns (De Montjoye et al., 2013) and the lack of comprehensive monitoring devices. Thus, only aggregated traffic volumes at the network level are available for the car mode.

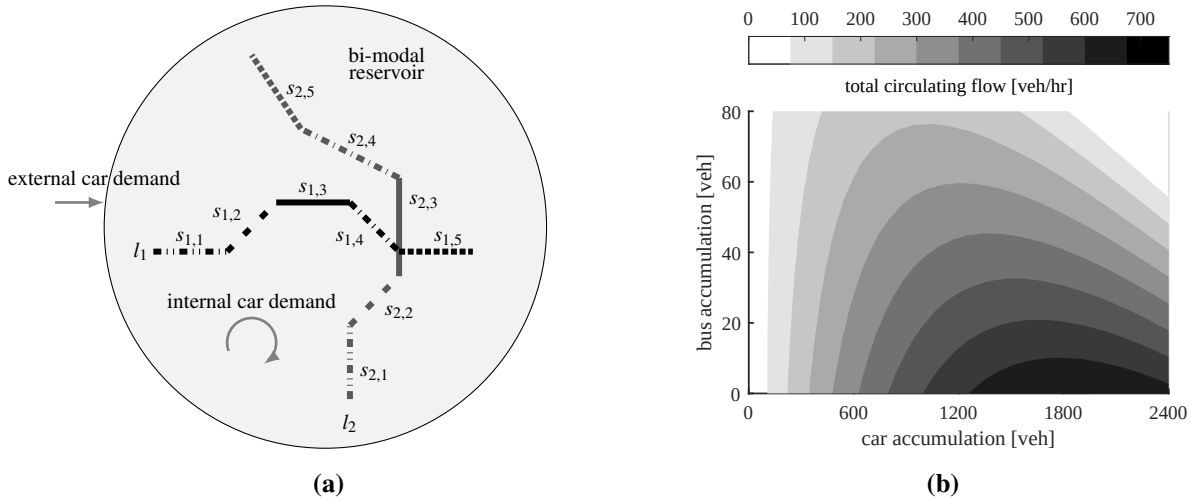


Figure 2: Schematic representation of the bi-modal reservoir model (a) used to model traffic dynamics based on the 3D-MFD (b).

Given limited data provision for the car mode (as per Assumption 1), we abstract the given bi-modal urban network into a macroscopic bi-modal reservoir system (Fig. 2a). This level of abstraction allows us to model the traffic dynamics at the network level using the 3D-MFD (Fig. 2b). The 3D-MFD can capture the complex multimodal interactions, which include the impact of bus operations (e.g. accumulation, dispatching frequency, dwell times at bus stops along mixed lanes, transit signal priority along dedicated bus lanes, etc.) on the speed of cars, as well as the impact of car accumulation on the operation of buses (e.g. reduced speed along mixed lanes in congested traffic conditions). Moreover, the 3D-MFD implicitly accounts for the OD pairs and route choices of car users. Recently, researchers have demonstrated using both simulation (Paipuri and Leclercq, 2020a) and empirical data (Paipuri and Leclercq, 2020b) that the 3D-MFD model can be successfully applied as a valid and accurate modeling tool to asses various traffic management strategies, such as the frequency setting problem. In this paper, we apply the 3D-MFD concept to model the vehicular dynamics for the car mode at the network level. Without loss of generality, we assume that the congestion in the network is homogeneous, i.e. it exhibits a well-defined 3D-MFD, as otherwise the network can be partitioned into several homogeneous regions using well-established partitioning algorithms (see e.g. Saeedmanesh and Geroliminis, 2017; Ambühl et al., 2019). Note, however, that such network-level modeling cannot be directly applied to the bus mode, as the lane allocation might not remain constant along the whole bus line. This, in turn, may affect the operating regime, i.e. the speed of the bus system along a given line. To account for such effects, we model the vehicular dynamics for the bus mode at the segment level. We do this by splitting each bus line $l$ into multiple segments, indexed by $s \in \mathcal{S}_l = \{1, ..., |\mathcal{S}_l|\}$ (Fig. 2a), with similar length $\ell_{b,l,s}$, where $|\mathcal{S}_l|$ is the

total number of segments along line $l$. These segments are grouped into two categories, i.e. per lane allocation type (mixed or dedicated): $\mathcal{S}_l = \mathcal{S}_{l,\text{mix}} \cup \mathcal{S}_{l,\text{ded}}$. Note that, for any given segment, we assign the lane allocation that is more prevalent along its length. Following the rationale of the cell transmission model (CTM) (Daganzo, 1994), we set the length of the smallest segment across all bus lines, $\min_{l\in\mathcal{L},s\in\mathcal{S}_l} \ell_{b,l,s}$, to be such that it satisfies the following condition: $\tau \cdot v_{\max} \leq \min_{l\in\mathcal{L},s\in\mathcal{S}_l}\{\ell_{b,l,s}\}$, where $v_{\max}$ denotes the free-flow speed of both modes[2]. It is also important to note that the proposed approach for modeling traffic dynamics for the bus mode is CTM-inspired, i.e. it is not the CTM, and is only used to account for varying lane allocation along a bus line. That being said, the modal interactions could potentially be captured better if a detailed CTM is used for both modes. However, this would require a detailed information on OD pairs and route choices of car users, which is not available for the considered problem, as stated in Assumption 1.

In the following, we formulate the optimization framework used to determine the optimal number of combined modular bus units and the optimal frequency at which the units should be dispatched.

### 3.2. Mathematical formulation of the optimization framework

In this subsection, we present the optimization framework for maximizing the efficiency of the proposed bus dispatching system. The decision variables for the considered optimization problem include: (i) the dispatching flow of buses $\boldsymbol{\Psi} = [\Psi_{b,i,l}(t) : l \in \mathcal{L}, i \in \mathcal{I}, t \in \mathcal{T}]$, defined as the number of dispatching buses (both conventional and modular) per unit time; and (ii) the dispatching flow of bus units $\boldsymbol{\psi} = [\psi_{b,i,l}(t) : l \in \mathcal{L}, i \in \mathcal{I}, t \in \mathcal{T}]$, defined as the number of dispatching bus units (both conventional and modular) per unit time, across all bus lines and for each time interval. Recall that, for conventional buses, $\psi_{b,r,l}(t) = \Psi_{b,r,l}(t)$, as each conventional bus is considered as an individual conventional bus unit. This, however, may not be the case for modular buses, given that each modular bus may include multiple modular bus units, i.e. $\psi_{b,m,l}(t) \geq \Psi_{b,m,l}(t)$. This way, we can obtain the average number of modular bus units contained in a dispatching modular bus on any given line $l$ during time interval $t$ as $u_{m,l}(t) = \psi_{b,m,l}(t)/\Psi_{b,m,l}(t)$. The objective function is to minimize the total system cost $Z(\boldsymbol{\Psi}, \boldsymbol{\psi})$, which includes both the operator, $Z_O(\boldsymbol{\Psi}, \boldsymbol{\psi})$, and the user cost, $Z_U(\boldsymbol{\Psi}, \boldsymbol{\psi})$ (the specific forms of $Z_O(\boldsymbol{\Psi}, \boldsymbol{\psi})$ and $Z_U(\boldsymbol{\Psi}, \boldsymbol{\psi})$ will be given in Eq. 8 and Eq. 9, respectively). Mathematically, the optimization problem can be formulated as:

$$\min_{\boldsymbol{\Psi},\boldsymbol{\psi}} Z(\boldsymbol{\Psi}, \boldsymbol{\psi}) = Z_O(\boldsymbol{\Psi}, \boldsymbol{\psi}) + Z_U(\boldsymbol{\Psi}, \boldsymbol{\psi}), \tag{2}$$

$$\text{s.t. } \psi_{b,i,l}(t) \geq 0, \quad \forall l \in \mathcal{L},\ t \in \mathcal{T},\ i \in \mathcal{I}, \tag{3}$$

$$\Psi_{b,\min} \leq \Psi_{b,i,l}(t) \leq \Psi_{b,\max}, \quad \forall l \in \mathcal{L},\ t \in \mathcal{T},\ i \in \mathcal{I}, \tag{4}$$

$$\Psi_{b,i,l}(t) - \psi_{b,i,l}(t) \leq 0, \quad \forall l \in \mathcal{L},\ t \in \mathcal{T},\ i \in \mathcal{I}, \tag{5}$$

$$|\mathcal{U}_{i,l}| \cdot \Psi_{b,i,l}(t) - \psi_{b,i,l}(t) \geq 0, \quad \forall l \in \mathcal{L},\ t \in \mathcal{T},\ i \in \mathcal{I}, \tag{6}$$

$$\underbrace{\sum_{l\in\mathcal{L}} \psi_{b,i,l}(t) \cdot \tau}_{\textit{term 1}} + \underbrace{\sum_{l\in\mathcal{L}} \sum_{s\in\mathcal{S}_l} n_{b,i,l,s}(t)}_{\textit{term 2}} - \underbrace{\sum_{l\in\mathcal{L}} \frac{n_{b,i,l,|\mathcal{S}_l|}(t)}{N_b(t)} \cdot G_b(N_b(t), N_c(t)) \cdot \tau}_{\textit{term 3}} \leq F_i, \quad \forall t \in \mathcal{T},\ i \in \mathcal{I}, \tag{7}$$

where $N_b(t)$ and $N_c(t)$ stand for the total bus and car accumulation, respectively, at the beginning of time interval $t$; $n_{b,i,l,s}(t)$ is the accumulation of bus units of type $i$ on segment $s$ of line $l$ at the beginning of time interval $t$; $G_b(\cdot)$ is the trip completion flow for the bus mode given by the 3D-MFD during time interval $t$.

In terms of the constraints, Eq. 3 ensures that the dispatching flows of bus units $\boldsymbol{\psi}$ need to be non-negative. Eq. 4 sets bounds for the dispatching flows of buses, where thresholds $\Psi_{b,\min}$ and $\Psi_{b,\max}$ are determined by the bus operator as the minimum and maximum bus frequencies, respectively, based on the operator budget, restrictions imposed by traffic authorities, demand patterns, etc. These thresholds could further ensure that the resulting frequencies fall within the range of frequencies observed in the given 3D-MFD, such that the 3D-MFD would be invariant with respect to the dispatching frequencies. Eq. 5 represents that the number of dispatching buses along any bus line cannot exceed the number of dispatching bus units. Eq. 6 ensures that the number of combined modular units does not exceed the maximum value for a given line. Eq. 7 represents that the total number of bus units required for the operation during

[2]We assume that buses and cars have the same free-flow speed, as described later in Assumption 4.

the planning horizon cannot exceed the vehicle fleet in case of both conventional and modular bus units, where *term 1* calculates the total number of dispatching bus units of type $i$ across all bus lines during time interval $t$, *term 2* denotes the total accumulation of bus units of type $i$ across all bus lines at the beginning of time interval $t$, and *term 3* computes the total number of bus units of type $i$ reaching their destination/terminal station across all bus lines during time interval $t$.

The operator cost $Z_O(\boldsymbol{\Psi}, \boldsymbol{\psi})$ is calculated in Eq. 8 based on the number of bus units of each type operating during the planning horizon and the corresponding unit cost $\pi_i$. Here, $\pi_i$ includes the unit cost of assigning bus drivers to conventional bus units, as well as the unit cost incurred by the energy consumption and maintenance of both conventional and modular bus units.

$$Z_O(\boldsymbol{\Psi}, \boldsymbol{\psi}) = \sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{I}} \sum_{l\in\mathcal{L}} \sum_{s\in\mathcal{S}_l} n_{b,i,l,s}(t) \cdot \pi_i \cdot \tau. \tag{8}$$

On the other hand, the user cost $Z_U(\boldsymbol{\Psi}, \boldsymbol{\psi})$ is related to the total time traveled in the system (including the waiting time), computed across both modes and converted into an equivalent monetary cost with a parameter $\Omega$ denoting the average value of time (Eq. 9).

$$Z_U(\boldsymbol{\Psi}, \boldsymbol{\psi}) = \Omega \cdot \underbrace{\sum_{t\in\mathcal{T}} \sum_{l\in\mathcal{L}} \sum_{s\in\mathcal{S}_l} \sum_{s'\in\mathcal{S}_l} (P_{l,s,s'}(t) + \omega_{l,s,s'}(t)) \cdot \tau}_{\textit{term 1}} + \Omega \cdot \underbrace{\sum_{t\in\mathcal{T}} \sum_{l\in\mathcal{L}} \sum_{s\in\mathcal{S}_l} \sum_{s'\in\mathcal{S}_l} \frac{1}{2} \cdot \frac{\lambda_{l,s,s'}(t) \cdot \tau \cdot |\mathcal{I}|}{\sum_{i\in\mathcal{I}} \Phi_{b,i,l,s}(t)}}_{\textit{term 2}} + \Omega \cdot \underbrace{\sum_{t\in\mathcal{T}} \rho_c \cdot N_c(t) \cdot \tau}_{\textit{term 3}}, \tag{9}$$

where $P_l(t)$ and $\omega_l(t)$ stand for the total number of on-board passengers and the total number of passengers who cannot board the bus, respectively, on line $l$ at the beginning of time interval $t$; $\lambda_l(t)$ is the average arrival rate of bus passengers on line $l$ during time interval $t$; $\rho_c$ denotes the average car occupancy ($\rho_c \approx 1$). In Eq. 9, *term 1* calculates the time traveled by bus users and the waiting time of passengers who cannot board the bus; *term 2* represents a proxy for the passenger waiting time due to the bus headway defined as half of the average bus headway across all types of bus units, $(\sum_{i\in\mathcal{I}} \Phi_{b,i,l}(t)/|\mathcal{I}|)^{-1}$; and *term 3* denotes the time traveled by car users. This way, the user cost function captures the influence of bus operations on the travel cost of both modes.

Notice that the formulated optimization problem requires input variables from both vehicular and passenger perspective. These variables are estimated based on the 3D-MFD. The input parameters to the 3D-MFD model are the topological layout of bus lines (including the lane allocation, the number of bus stops, as well as the distance between the stops) and both, the car and the bus passenger demand level. In the following, we describe the proposed modeling framework based on the 3D-MFD.

We make the following assumptions for the methodological framework, in terms of passenger preferences, the automated modular vehicle technology, and the effects of bus operations on car traffic.

**Assumption 2** (Passenger preferences)**.** We make the following assumptions regarding the passenger preferences: (i) the demand information of each mode, as well as the routing information of bus passengers, is exogenously given, i.e. we do not consider a dynamic change in mode choice, route choice, or departure time as a function of the dispatching policy; and (ii) bus passengers do not have a preference on which type of units they choose to board.

**Remark.** We make the following remarks regarding Assumption 2. First, we aim to develop short-term dispatching strategies to improve the overall operator and passenger performance, rather than making long-term planning decisions (e.g. adding bus lines, changing the spatial distribution of bus lanes, changing the bus network structure, etc.) that might affect the demand. This is similar to other concepts related to improving bus operations such as (bi-modal) perimeter control (Ampountolas et al., 2017; Chiabaut et al., 2018; He et al., 2019; Dakic et al., 2019; Yang et al., 2019), anti-bus bunching methods (Eberlein et al., 2001; Hickman, 2001; Daganzo, 2009; Xuan et al., 2011; Muñoz et al., 2013; Sirmatel and Geroliminis, 2018), or other bus dispatching-related studies (Szeto and Wu, 2011; Hadas and Shnaiderman, 2012; Gkiotsalitis and Cats, 2018), which assume that there is no instantaneous mode, route, or departure time shift due to the implemented control strategy. Second, we assume that bus passengers do not have a preference on which type of units they choose to board, which applies to a common scenario where passengers board whichever bus is first available to them. Nevertheless, the proposed methodological framework can be extended to account for passenger preference over types of units, which, however, is out of the scope for this paper.

**Assumption 3** (Automated modular vehicle technology)**.** We make the following assumptions regarding the considered modular vehicle technology: (i) the composition of modular bus units is determined only at the terminal station, and that no combining/splitting of modular units occurs along a bus line once a modular bus is dispatched; (ii) the size of a modular bus never exceeds the size of the conventional bus, i.e. the value of $|\mathcal{U}_{m,l}|$ is chosen such that the length of a modular bus containing the maximum number of combined modular units corresponds to the length of the conventional bus; and (iii) automated modular bus units have similar vehicle dynamics as conventional buses.

**Remark.** We make the following remarks regarding Assumption 3. First, although the en-route combining and splitting of modular bus units is technologically feasible, this requires accurate prediction of bus travel times. However, the bus travel times can be highly stochastic on mixed lanes, and thus we assume that the composition of modular bus units is determined only at the terminal station. Moreover, such en-route combining and splitting of units is only helpful if the units were to take different routes or skip stops; and both of these strategies are out of the scope for this paper. Second, the main benefit of modular bus units is that we can dispatch buses with lower capacity in scenarios where demand is lower to reduce operational costs, and thus it is not beneficial to consider the operational regime where the size of modular buses is larger than that of conventional buses. Additionally, this constraint ensures that the combined modular bus units never exceed the physical length of the current bus stops. Third, given the assumption of similar vehicle dynamics of modular bus units and conventional buses, we do not consider the impact of reduced reaction times (enabled by the automated vehicle technology) on the shape of the 3D-MFD.

**Assumption 4** (Effects of bus operations on car traffic)**.** We assume that, along mixed lanes, buses drive at the speed of cars while cruising, i.e. both modes have the same speed either in the free-flow or congested traffic conditions (as in e.g. Daganzo, 2010; Estrada et al., 2011; Zheng and Geroliminis, 2013; Badia et al., 2016; Dakic and Menendez, 2018; Zhang and Liu, 2019). Consequently, buses do not act as moving bottleneck along mixed lanes, i.e. they affect car traffic only due to their stops (i.e. dwell time). Other effects of bus operations on car traffic, such as the effects of the vehicle size along mixed lanes and transit signal priority along dedicated bus lanes, are incorporated into the parameters of the 3D-MFD model.

**Remark.** We make the following remarks regarding Assumption 4. First, the 3D-MFD implicitly assumes that the vehicle size of buses is comparable to that of conventional buses. Note, however, that the size of a modular bus never exceeds the size of the conventional bus (as per Assumption 4). Hence, employing the 3D-MFD could over-estimate the negative impact of modular buses on car traffic. That being said, in this paper, we are taking a conservative approach that might under-estimate the benefits of the modular vehicle technology. Second, although buses and cars do not directly interact on dedicated bus lanes, bus operations can still affect car traffic through transit signal priority. These effects are, nevertheless, captured in the 3D-MFD.

### *3.3. Modeling vehicular dynamics*

To model car traffic at the network level while capturing the interactions between the modes based on the 3D-MFD, we follow the approach by Loder et al. (2017) and Paipuri and Leclercq (2020a), and assume a linear model between the car speed and the accumulation of both modes:

$$v_c(t) = \alpha_{b,\text{mix}} \cdot N_{b,\text{mix}}(t) + \alpha_{b,\text{ded}} \cdot N_{b,\text{ded}}(t) + \alpha_c \cdot N_c(t) + \beta, \tag{10}$$

where $\alpha_{b,\text{mix}}$, $\alpha_{b,\text{ded}}$, and $\alpha_c$ are parameters that capture the marginal effect of each mode on the average car speed; $\beta$ characterizes the effects of the bus network topology; $N_{b,\text{mix}}(t)$ and $N_{b,\text{ded}}(t)$ stand for the bus accumulation in the subnetwork with mixed and dedicated bus lanes, respectively, at the beginning of time interval $t$, such that $N_{b,\text{mix}}(t) + N_{b,\text{ded}}(t) = N_b(t)$. Note that $\alpha_{b,\text{mix}}$ accounts for the effects of the vehicle size of conventional buses on car speed along mixed lanes, whereas $\alpha_{b,\text{ded}}$ captures the effects of transit signal priority along dedicated segments. These parameters can be estimated with real data (see Loder et al. (2017) for more details). In this paper, we assume that the parameters of the 3D-MFD model are exogenously given.

Due to varying lane allocation along a bus line, we model the bus speed at the segment level as a function of the average (across both types of units) dwell time on a given segment, $\Theta_{l,s}(t)$, i.e.:

$$v_{b,\ell,s}(t) = \begin{cases} (1/v_c(t) + \theta'/\ell'_{b,l,s} + \Theta_{l,s}(t))^{-1}, & \text{if } s \in \mathcal{S}_{l,\text{mix}}, \\ (1/v_{\max} + \theta'/\ell'_{b,l,s} + \Theta_{l,s}(t))^{-1}, & \text{if } s \in \mathcal{S}_{l,\text{ded}}, \end{cases} \tag{11}$$

where $\theta'$ denotes the time lost per stop due to required door operations and deceleration/acceleration maneuvers; $\ell'_{b,l,s}$ is the average distance between the bus stops on segment $s$ of line $l$. Notice from Eq. 11 that, for dedicated bus segments, the bus speed is modeled independently of other vehicles, i.e. based on the free-flow speed of buses. On the other hand, given Assumption 4 that buses drive at the speed of cars while cruising, we model the bus speed along mixed segments based on the average car speed given by the 3D-MFD. Before we formulate the average dwell time $\Theta_{l,s}(t)$ used in Eq. 11, let us first introduce some notation.

Let $B_{l,s}(t)$ be the total number of boarding passengers and $A_{l,s}(t)$ the total number of alighting passengers on segment $s$ of line $l$ during time interval $t$. These variables are estimated using the passenger dynamics model, as described in the following subsection. Denote by $N_{b,i,l,s}(t)$ and $n_{b,i,l,s}(t)$ the accumulation of buses and the accumulation of bus units, respectively, of type $i$ on segment $s$ of line $l$ at the beginning of time interval $t$. The dwell time of buses of type $i$ on segment $s$ can be computed as the ratio of the total number of boarding/alighting passengers across all buses of type $i$ on a given segment $s$ (the term inside parenthesis in Eq. 12) and the corresponding bus accumulation:

$$\Theta_{i,l,s}(t) = \frac{1}{N_{b,i,l,s}(t)} \cdot \left( \theta \cdot \max\{B_{l,s}(t), A_{l,s}(t)\} \cdot \frac{n_{b,i,l,s}(t) \cdot C_i}{\sum_{i \in \mathcal{I}} n_{b,i,l,s}(t) \cdot C_i} \right), \tag{12}$$

where $\theta$ denotes the time added per boarding/alighting passenger. Note that Eq. 12 assumes a simultaneous boarding and alighting process. Alternatively, one can use a dwell time model based on a sequential boarding and alighting process. Also note that, due to Assumption 2, the bus occupancy can be uniformly distributed across both types of units. Therefore, the total number of passengers boarding/alighting buses of type $i$ is computed to be proportional to the total capacity of buses of a given type (the fraction inside parenthesis in Eq. 12).

The average dwell time on a given segment $s$ can then be computed as an average of the dwell times of both types of buses, weighted by their respective bus accumulations (Eq. 13).

$$\Theta_{l,s}(t) = \frac{\sum_{i \in \mathcal{I}} N_{b,i,l,s}(t) \cdot \Theta_{i,l,s}(t)}{\sum_{i \in \mathcal{I}} N_{b,i,l,s}(t)} = \frac{\theta \cdot \max\{B_{l,s}(t), A_{l,s}(t)\}}{\sum_{i \in \mathcal{I}} N_{b,i,l,s}(t)}. \tag{13}$$

We can also relate the total bus accumulation in each subnetwork (with mixed and dedicated bus segments) to the bus accumulation across all segments of the same type and across both types of buses (Eq. 14).

$$N_{b,\text{mix}}(t) = \sum_{i \in \mathcal{I}} \sum_{l \in \mathcal{L}} \sum_{s \in \mathcal{S}_{l,\text{mix}}} N_{b,i,l,s}(t), \tag{14a}$$

$$N_{b,\text{ded}}(t) = \sum_{i \in \mathcal{I}} \sum_{l \in \mathcal{L}} \sum_{s \in \mathcal{S}_{l,\text{ded}}} N_{b,i,l,s}(t). \tag{14b}$$

Now that we have introduced all the notation and defined all modeling variables, we can formulate the mass conservation equations used to model the evolution of the vehicle accumulation for each mode. Given the assumption of the homogeneous traffic conditions across the network, which assumes that the average car speed in mixed lanes is similar to the average car speed in car only lanes (as in e.g. Zheng and Geroliminis, 2013; Dakic and Menendez, 2018; Zhang and Liu, 2019), the system dynamics for the car mode at the network level can be modeled as:

$$N_c(t+1) = N_c(t) + \lambda_{c,\text{int}}(t) \cdot \tau + \lambda_{c,\text{ext}}(t) \cdot \tau - G_c(N_b(t), N_c(t)) \cdot \tau, \tag{15}$$

where $\lambda_{c,\text{int}}(t)$ and $\lambda_{c,\text{ext}}(t)$ stand for the internal (originating from within the network) and external (incoming to the network) car demand, respectively, during time interval $t$; and $G_c(\cdot)$ is the trip completion flow for the car mode given by the 3D-MFD during time interval $t$. The trip completion flow for the bus and the car mode includes both the internal outflow (reaching destination inside the network) and the external outflow (exiting the network). Geroliminis and Daganzo (2008) showed that $G_b(\cdot)$ and $G_c(\cdot)$ are proportional to the total circulating flows, i.e. $Q_b(\cdot) = v_b(t) \cdot N_b(t)/L_b$ and $Q_c(\cdot) = v_c(t) \cdot N_c(t)/L_c$, respectively, with a factor that represents the ratio of the network length for a given mode ($L_c$ for cars; $L_b$ for buses) and the average vehicular trip length ($\bar{\ell}_c$ for cars; $\bar{\ell}_b$ for buses):

$$G_b(N_b(t), N_c(t)) = Q_b(N_b(t), N_c(t)) \cdot L_b/\bar{\ell}_b = v_b(t) \cdot N_b(t)/\bar{\ell}_b, \tag{16a}$$

$$G_c(N_b(t), N_c(t)) = Q_c(N_b(t), N_c(t)) \cdot L_c/\bar{\ell}_c = v_c(t) \cdot N_c(t)/\bar{\ell}_c, \tag{16b}$$

where $v_b(t)$ is the average bus speed in the network (Eq. 17). It is worth mentioning that the average vehicular trip length for the bus mode is defined as the average length of a bus line in one direction of travel.

$$v_b(t) = \frac{\sum_{i\in\mathcal{I}}\sum_{l\in\mathcal{L}}\sum_{s\in\mathcal{S}_l} N_{b,i,l,s}(t) \cdot v_{b,l,s}(t)}{\sum_{i\in\mathcal{I}}\sum_{l\in\mathcal{L}}\sum_{s\in\mathcal{S}_l} N_{b,i,l,s}(t)}. \tag{17}$$

On the other hand, due to different operational regimes of buses across different lane allocations along a bus line, as stated before, the system dynamics for the bus mode is modeled at the segment level:

$$N_{b,i,l,s}(t+1) = \begin{cases} N_{b,i,l,s}(t) + \Psi_{b,i,l}(t) \cdot \tau - \Phi_{b,i,l,s}(t) \cdot \tau, & \text{if } s = 1, \\ N_{b,i,l,s}(t) + \Phi_{b,i,l,s-1}(t) \cdot \tau - \Phi_{b,i,l,s}(t) \cdot \tau, & \text{if } 1 < s < |\mathcal{S}_l|, \\ N_{b,i,l,s}(t) + \Phi_{b,i,l,s-1}(t) \cdot \tau - (N_{b,i,l,s}(t)/N_b(t)) \cdot G_b(N_b(t), N_c(t)) \cdot \tau, & \text{if } s = |\mathcal{S}_l|, \end{cases} \tag{18a}$$

$$n_{b,i,l,s}(t+1) = \begin{cases} n_{b,i,l,s}(t) + \psi_{b,i,l}(t) \cdot \tau - \phi_{b,i,l,s}(t) \cdot \tau, & \text{if } s = 1, \\ n_{b,i,l,s}(t) + \phi_{b,i,l,s-1}(t) \cdot \tau - \phi_{b,i,l,s}(t) \cdot \tau, & \text{if } 1 < s < |\mathcal{S}_l|, \\ n_{b,i,l,s}(t) + \phi_{b,i,l,s-1}(t) \cdot \tau - (n_{b,i,l,s}(t)/N_b(t)) \cdot G_b(N_b(t), N_c(t)) \cdot \tau, & \text{if } s = |\mathcal{S}_l|, \end{cases} \tag{18b}$$

where $\Phi_{b,i,l,s}(t)$ and $\phi_{b,i,l,s}(t)$ denote the transferring flow of buses and the transferring flow of bus units, respectively, of type $i$ between segments $s$ and $s+1$ of line $l$ during time interval $t$. These transferring flows are computed using a CTM-inspired approach, as the minimum of the sending flow from segment $s$ of line $l$ (the first term in Eq. 19) and the receiving flow to segment $s+1$ of the same line during a given interval $t$ (the second term in Eq. 19). Note that, for simplicity of presentation, Eq. 19 assumes that the bus lines do not overlap as otherwise the sending and receiving flows of buses/bus units should be proportionally distributed on a given overlapping segment based on the number of buses/bus units along the corresponding bus line.

$$\Phi_{b,i,l,s}(t) = \min\left\{v_{b,l,s}(t) \cdot \frac{N_{b,i,l,s}(t)}{\ell_{b,l,s}}, \frac{\varphi_{c,l,s+1}(t)}{\eta_i \cdot u_{i,l,s}(t)}\right\}, \tag{19a}$$

$$\phi_{b,i,l,s}(t) = \min\left\{v_{b,l,s}(t) \cdot \frac{n_{b,i,l,s}(t)}{\ell_{b,l,s}}, \frac{\varphi_{c,l,s+1}(t)}{\eta_i}\right\}, \tag{19b}$$

where $\varphi_{c,l,s}(t)$ denotes the receiving flow of cars to segment $s$ of line $l$ during time interval $t$; $\eta_i$ represents a parameter that quantifies by how much one bus unit of type $i$ reduces the car flow (i.e. a bus unit-car equivalent); $u_{m,l,s}(t) = n_{b,m,l,s}(t)/N_{b,m,l,s}(t)$ is the average number of modular bus units contained in a modular bus on segment $s$ of line $l$ at the beginning of time interval $t$. Recall that for conventional bus units $u_{r,l,s}(t) = 1$, i.e. $n_{b,r,l,s}(t) = N_{b,r,l,s}(t)$. Before we formulate $\varphi_{c,l,s}(t)$, let us first introduce some notation.

Let $\overline{N}_c$ be the maximum car accumulation in the subnetwork where cars are allowed to drive (including car only and mixed segments, i.e. $L_c$). Recall that in this study we assume homogeneous traffic conditions, i.e. the average car speed in mixed segments is similar to the average car speed in car only segments. This, in turn, implies that we can determine the total car accumulation in the subnetwork with mixed segments at the beginning of time interval $t$, $N_{c,\text{mix}}(t)$, by multiplying the total car accumulation in the network at the beginning of time interval $t$, $N_c(t)$, by a fraction that represents the ratio of the maximum car accumulation that can be accommodated in the subnetwork with mixed segments at the beginning of time interval $t$, $\overline{N}_{c,\text{mix}}(t)$, and the maximum car accumulation in the network, $\overline{N}_c$:

$$N_{c,\text{mix}}(t) = N_c(t) \cdot \frac{\overline{N}_{c,\text{mix}}(t)}{\overline{N}_c} = N_c(t) \cdot \frac{\sum_{l\in\mathcal{L}}\sum_{s\in\mathcal{S}_{l,\text{mix}}} \overline{N}_{c,l,s}(t)}{\overline{N}_c}, \tag{20}$$

where $\overline{N}_{c,l,s}(t)$ stands for the maximum car accumulation that can be accommodated on a given segment $s$ of line $l$ at the beginning of time interval $t$, defined as a function of the bus accumulation on that same segment (Eq. 21).

$$\overline{N}_{c,l,s}(t) = \overline{N}_c \cdot \frac{\ell_{b,l,s}}{L_c} - \sum_{i\in\mathcal{I}} n_{b,i,l,s}(t) \cdot \eta_i. \tag{21}$$

Using $\overline{N}_{c,l,s}(t)$, we can further distribute $N_{c,\text{mix}}(t)$ across the mixed segments according to the proportion of the remaining capacity, i.e:

$$N_{c,l,s}(t) = N_{c,\text{mix}}(t) \cdot \frac{\overline{N}_{c,l,s}(t)}{\sum_{l\in\mathcal{L}}\sum_{s\in\mathcal{S}_{l,\text{mix}}} \overline{N}_{c,l,s}(t)} = N_c(t) \cdot \frac{\overline{N}_{c,l,s}(t)}{\overline{N}_c}. \tag{22}$$

Similarly to the CTM, we can now compute the receiving flow of cars based on the maximum available car accumulation, $\overline{N}_{c,l,s}(t)$, and (only in case of a mixed segment) the number of cars present on a given segment at the beginning of time interval $t$, as given by Eq. 23, where $w$ is the backward wave speed.

$$\varphi_{c,l,s}(t) = \begin{cases} w \cdot (\overline{N}_{c,l,s}(t) - N_{c,l,s}(t))/\ell_{b,l,s}, & \text{if } s \in \mathcal{S}_{l,\text{mix}}, \\ w \cdot \overline{N}_{c,l,s}(t)/\ell_{b,l,s}, & \text{if } s \in \mathcal{S}_{l,\text{ded}}. \end{cases} \tag{23}$$

### *3.4. Modeling passenger dynamics*

In this section, we extend the modeling formulations to account for passenger occupancy dynamics. Taking into account that for the car mode, the number of passengers served can essentially be approximated with the number of served vehicles, i.e. the average car occupancy is $\rho_c \approx 1$, in the following we only describe the evolution of the total bus passengers over time:

$$P_{l,s,s'}(t+1) = \begin{cases} P_{l,s,s'}(t) + D_{\ell,s,s'}(t) - O_{\ell,s,s'}(t), & \text{if } s = 1, \\ P_{l,s,s'}(t) + D_{\ell,s,s'}(t) + O_{\ell,s-1,s'}(t) - O_{\ell,s,s'}(t), & \text{if } s > 1, \end{cases} \tag{24}$$

where $P_{l,s,s'}(t)$ is the total number of on-board passengers on segment $s$ with destination on segment $s'$ of line $l$ at the beginning of time interval $t$; $D_{l,s,s'}(t)$ represents the total number of boarding passengers on segment $s$ with destination on segment $s'$ of line $l$ during time interval $t$; and $O_{l,s,s'}(t)$ denotes the outflow (including the transferring and the trip completion flow, as elaborated further below) of bus passengers between segments $s$ and $s+1$ of line $l$ during time interval $t$, whose destination is on segment $s'$ of the same line (Eq. 25).

$$O_{\ell,s,s'}(t) = \begin{cases} \min\{P_{l,s,s'}(t), \sum_{i\in\mathcal{I}} \rho_{b,i,l,s,s'}(t) \cdot \phi_{b,i,l,s}(t) \cdot \tau\}, & \text{if } s \neq s' \wedge s < |\mathcal{S}_l| \\ \min\{P_{l,s,s'}(t), R_{l,s,s'}(N_b(t), N_c(t)) \cdot \tau\}, & \text{if } s = s' \vee s = |\mathcal{S}_l| \end{cases} \tag{25}$$

where $\rho_{b,i,l,s,s'}(t)$ and $R_{l,s,s'}(\cdot)$ stand for the occupancy of a bus unit of type $i$ and the trip completion flow, respectively, of bus passengers on segment $s$ of line $l$, whose destination is on segment $s'$ of the same line. Given Assumption 2 and the implied uniformly distributed bus occupancy across both types of units, we can obtain $\rho_{b,i,l,s,s'}(t)$ by multiplying the number of on-board passengers per unit capacity (the term inside parenthesis in Eq. 26) by the corresponding capacity of a given type of unit.

$$\rho_{b,i,l,s,s'}(t) = C_i \cdot \left( \frac{P_{l,s,s'}(t)}{\sum_{i\in\mathcal{I}} n_{b,i,l,s}(t) \cdot C_i} \right). \tag{26}$$

On the other hand, the trip completion flow $R_{l,s,s'}(\cdot)$ is computed by incorporating the average (across both types of units) bus occupancy of passengers on segment $s$ with destination is on segment $s'$ of line $l$, $\rho_{b,l,s,s'}(t)$, into Eq. 16b, and substituting the average vehicular distance for the bus mode with the average trip length of bus passengers, $\bar{\ell}_{b,\text{pax}}$, i.e.:

$$\begin{aligned} R_{l,s,s'}(N_b(t), N_c(t)) &= \rho_{b,l,s,s'}(t) \cdot Q_b(N_b(t), N_c(t)) \cdot L_b/\bar{\ell}_{b,\text{pax}} \\ &= \rho_{b,l,s,s'}(t) \cdot v_b(t) \cdot N_b(t)/\bar{\ell}_{b,\text{pax}}, \end{aligned} \tag{27}$$

where $\rho_{b,l,s,s'}(t)$ is given as an average of the occupancies of both types of buses (obtained by multiplying the average number of units contained in a bus of each type by the corresponding occupancy of a bus unit), weighted by their respective bus accumulations (Eq. 28).

$$\rho_{b,l,s,s'}(t) = \frac{\sum_{i\in\mathcal{I}} N_{b,i,l,s}(t) \cdot u_{i,l,s}(t) \cdot \rho_{b,i,l,s,s'}(t)}{\sum_{i\in\mathcal{I}} N_{b,i,l,s}(t)}. \tag{28}$$

Note that in case where passengers do not arrive at their destination (i.e. $s \neq s'$ and $s < |\mathcal{S}_l|$), the outflow $O_{\ell,s,s'}(t)$ represents the transferring flow from segment $s$ to segment $s+1$ of a given line $l$ during time interval $t$. Otherwise (i.e. $s = s'$ or $s = |\mathcal{S}_l|$), the outflow represents the trip completion flow of bus passengers given by the 3D-MFD.

The total number of boarding and alighting passengers on segment $s$ of line $l$ during time interval $t$ can now be determined as:

$$B_{l,s}(t) = \sum_{s' \in \mathcal{S}_l} D_{\ell,s,s'}(t), \tag{29}$$

$$A_{l,s}(t) = \sum_{s' \in \mathcal{S}_l} \mathbb{1}_{\{s=s' \ \vee \ s=|\mathcal{S}_l|\}} \cdot O_{\ell,s,s'}(t), \tag{30}$$

where $\mathbb{1}_{\{condition\}}$ is an indicator function that return the value of 1 if *condition* is satisfied. The total number of boarding passengers $D_{\ell,s,s'}(t)$ is bounded by the two parameters in Eq. 31: the total number of (accumulated) passengers that want to enter the bus, $D_{l,s,s',\text{acc}}$, and the total number of passengers that can enter the bus, $D_{l,s,s',\max}$:

$$D_{l,s,s'}(t) = \min\{D_{l,s,s',\text{acc}}(t), D_{l,s,s',\max}(t)\}. \tag{31}$$

Before we derive these two parameters, let us first introduce the evolution of the total number of passengers with destination on segment $s'$ of line $l$ who cannot board the bus on segment $s$ of the same line by the beginning of time interval $t$:

$$\omega_{l,s,s'}(t+1) = \omega_{l,s,s'}(t) + \lambda_{l,s,s'}(t) \cdot \tau - D_{l,s,s'}(t), \tag{32}$$

where $\lambda_{l,s,s'}(t)$ denotes the average arrival rate of bus passengers on segment $s$ with destination on segment $s'$ of line $l$ during time interval $t$.

Then, the total number of passengers that want to enter the bus unit on segment $s$ with destination on segment $s'$ of line $l$ during time interval $t$ can be obtained as follows:

$$D_{l,s,s',\text{acc}}(t) = \omega_{l,s,s'}(t) + \lambda_{l,s,s'}(t) \cdot \tau. \tag{33}$$

On the other hand, the total number of passengers with destination on segment $s'$ of line $l$ that can enter the bus on segment $s$ of the same line is determined by the total available capacity across both types of bus units operating along line $l$ during a given time interval:

$$D_{l,s,s',\max}(t) = \begin{cases} \delta_{l,s,s'}(t) \cdot \left( \sum_{i \in \mathcal{I}} n_{b,i,l,s}(t+1) \cdot C_i - \sum_{s' \in \mathcal{S}_l} (P_{l,s,s'}(t) - O_{l,s,s'}(t)) \right), & \text{if } s = 1, \\ \delta_{l,s,s'}(t) \cdot \left( \sum_{i \in \mathcal{I}} n_{b,i,l,s}(t+1) \cdot C_i - \sum_{s' \in \mathcal{S}_l} (P_{l,s,s'}(t) + O_{l,s-1,s'}(t) - O_{l,s,s'}(t)) \right), & \text{if } s > 1, \end{cases} \tag{34}$$

where $n_{b,i,l,s}(t+1)$ is calculated using Eq. 18, and $\delta_{l,s,s'}(t)$ represents the share of passengers on segment $s$ of line $l$ during time interval $t$ whose destination is on segment $s'$ of the same line (Eq. 35).

$$\delta_{l,s,s'}(t) = \frac{\lambda_{l,s,s'}(t) \cdot \tau + \omega_{l,s,s'}(t)}{\sum_{s' \in \mathcal{S}_l} (\lambda_{l,s,s'}(t) \cdot \tau + \omega_{l,s,s'}(t))}. \tag{35}$$

## 4. Numerical experiments and results

### *4.1. Case study and simulation scenarios*

Here we describe the simulation environment for testing the performance of the proposed flexible bus dispatching system. The considered traffic network is inspired by the City of Zurich, Switzerland, and is comprised of five bus lines with varying lane allocation. Dedicated bus segments are placed closer to the center of the network, whereas the mixed lane segments are installed closer to the periphery (Fig. 3). The simulated traffic conditions reflect a typical morning-peak period. The 3D-MFD and the demand profiles (both in terms of cars and passengers) during the 3 hour period (Fig. 4) are designed to mimic the aggregated traffic features of the city center of Zurich (Ambühl et al., 2017),

as this one exhibits a well-defined empirical 3D-MFD, as shown in Fig. 2b (see Loder et al., 2017, for more details). To obtain realistic traffic conditions, we impose lower and upper bounds to the conservation equations (both in terms of vehicles and passengers).

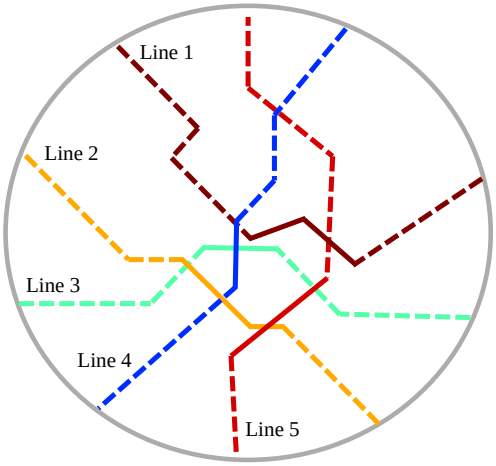


Figure 3: Schematic illustration of the studied network, where solid lines represent dedicated bus segments, whereas dashed lines represent mixed bus segments.

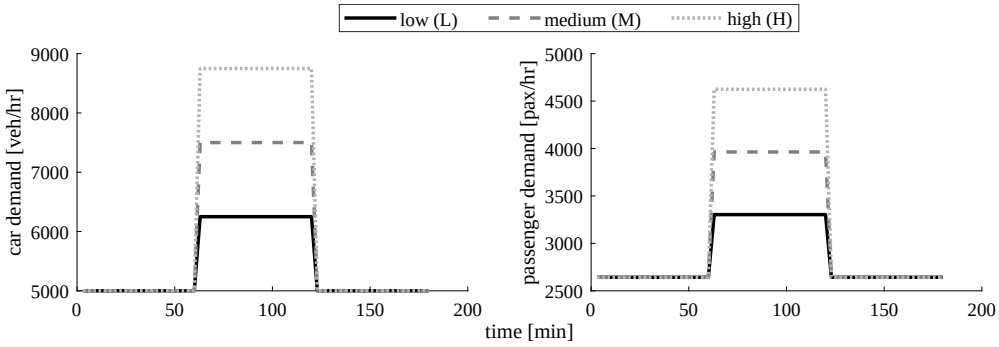


Figure 4: Simulation settings: (a) car demand profile; and (b) public transport passenger demand profile.

Tested traffic scenarios include three types of car demand (low, medium, and high) and three types of public transport passenger demand (low, medium, and high) (see Fig. 4). For each scenario, we vary the penetration rate of modular bus units: $p_m \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 1\}$. The penetration rate of $p_m = 0$ corresponds to the base case (i.e. the case corresponding to existing bus systems with only conventional buses), in which no modular units are incorporated into the bus fleet. In such a base case, the considered bus fleet includes $F_r = 50$ conventional bus units. In other cases with $p_m > 0$, the number of conventional buses depends on the penetration rate of modular bus units, i.e. $F_r = 50 \cdot (1 - p_m)$. We assume that the bus capacity is $C_r = 120$ pax/veh for conventional buses and $C_m = 20$ pax/veh for modular bus units, which gives $\zeta = 6$. We also use the value of $\zeta$ as the maximum number of modular bus units that can be combined across all bus lines, i.e. $|\mathcal{U}_{m,l}| = \zeta, \ \forall l \in \mathcal{L}$. The values of $C_m$ and $\zeta$ are determined based on the length of a modular bus unit (as in e.g. the Next Future Mobility system), such that the length of a modular bus containing the maximum number of combined modular units corresponds to the length of an articulated bus. The passenger cost per unit time is assumed to be 20 CHF/hr, comparable to the value of time in the city of Zurich (Axhausen et al., 2006). Finally, following recommendations by Sinner (2019), we approximate the unit cost $\pi_i$ for the bus operator by accounting for the driver cost (only in case of conventional buses), vehicle cost, operational cost (including the energy consumption and maintenance), overhead (e.g. administration and planning cost), and ticketing

cost. Using the values of the aforementioned elements of the operating unit cost for Switzerland, we obtain $\pi_r \approx 260$ CHF/hr and $\pi_m \approx 30$ CHF/hr.

The proposed optimization problem is solved using a conventional sequential quadratic programming algorithm (Boggs and Tolle, 1995), which allows to formulate the constraints explicitly, without using the penalty terms in the objective function. To find the optimal solution, we set the number of initial points to 50. Sensitivity analyses show that increasing the number of initial points beyond 50 leads to marginal improvements in the objective function (we omit the results here for brevity) for all scenarios with different demand levels and penetration rates of modular units. In the future, it might be possible to use more in depth sensitivity analysis methods (Ge et al., 2015; Ge and Menendez, 2017) to further improve the efficiency of the solution algorithm. Experiments are run on a 16-core Intel Xeon processor (3.19 GHz) with 256 GB RAM. The computation time for a given scenario and penetration rate of modular units is around 8.5 min for a 3-hour simulation.

### *4.2. Value of considering modular vehicle technology*

In this subsection, we demonstrate the value of considering the modular vehicle technology by comparing the total system cost (including the operator cost and the system delay) of the proposed flexible bus dispatching system with different penetration rates of modular units to that of the base case. The absolute (in CHF/hr) and relative (in %) improvements are shown in the form of bar plots (Fig. 5). Recall that the base case refers to the existing bus dispatching system consisting only of conventional buses, i.e. with $p_m = 0$. Note that even if the performance is the same, the cost for the base case described here is still not equivalent to the conventional bus cost. This is because we are only dealing with the frequency setting and not vehicle scheduling. The underlying vehicle scheduling problem (deadheading/assigning vehicles to trips or to reposition to merge with other vehicles) for modular bus units has a substantially different cost structure than that for conventional bus (likely much lower cost). Therefore, there are additional cost savings that are not covered here from that perspective. In Fig. 5, traffic scenarios are represented as XY, where X indicates the level of car demand and Y indicates the level of passenger demand for public transport: low (L), medium (M), or high (H). For example, LM stands for the scenario with low car and medium public transport passenger demand.

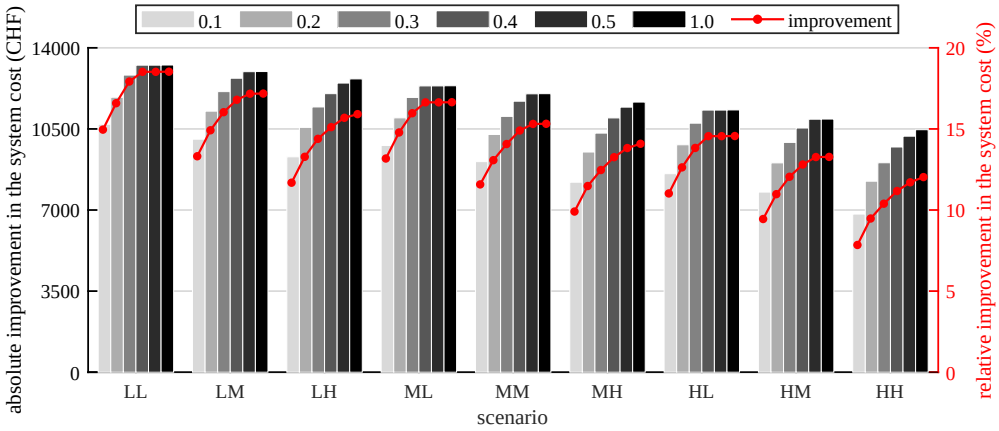


Figure 5: Comparison of the improvement in the total system cost for different penetration rates of modular units.

As demonstrated by Fig. 5, the proposed system substantially outperforms the base case with only conventional buses, especially for scenarios with lower car and public transport demand. Potential improvements can be up to 19%, depending on the level of car and passenger demand, as well as the penetration rate of modular bus units. As the level of demand (in particular for the car mode) increases, the improvement is smaller, but still significant. The reason for this is twofold. First, in case of low demand, the system sends a group of few combined (if not individual) modular units, optimizing thereby the utilization of the vehicle's capacity and reducing the operating cost. Second, due to a lower number of units being contained in a modular bus, the system has more units available in stock, allowing to dispatch more modular buses and reduce the passenger waiting time. This, on the other hand, has minimum impact

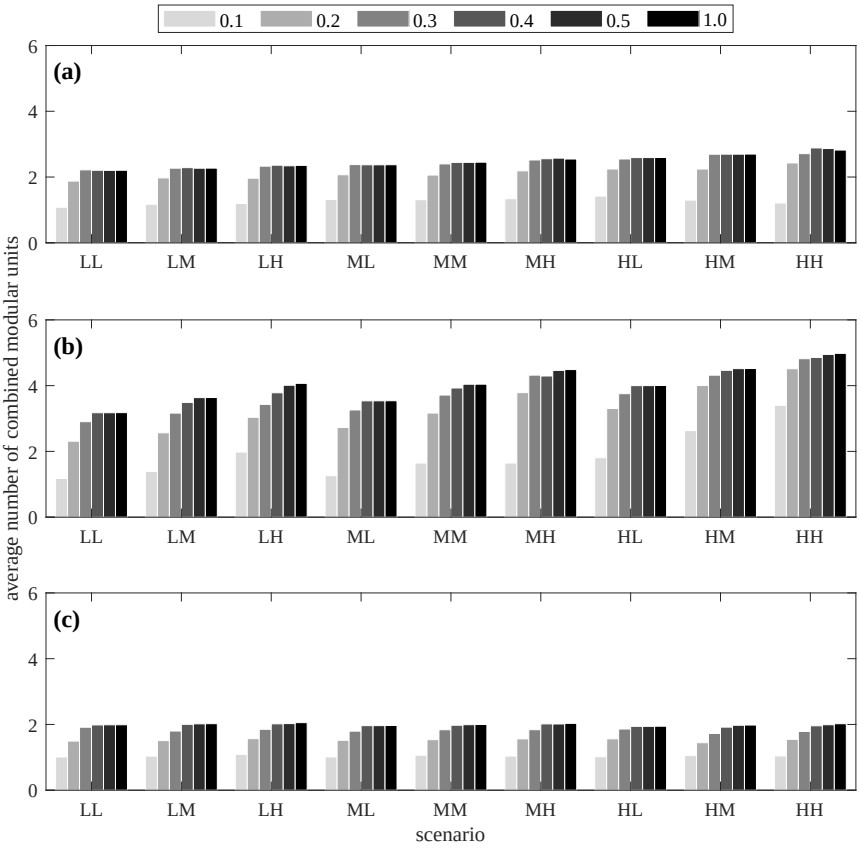


Figure 6: Evolution of the average number of units contained in a modular bus for different penetration rates of modular units: (a) before the peak; (b) during the peak; and (c) after the peak.

on car traffic in case of low car demand, due to lower interactions between the modes. That being said, dispatching more modular buses for higher car demand levels does not necessarily have to improve the system, as the impact on car traffic may be significant, resulting in lower overall system performance. This is illustrated in Fig. 6, where we observe that the average (across all bus lines) number of combined modular units increases as the car demand increases, especially during the peak period (Fig. 6b) and for higher penetration rates of modular units. Consequently, the improvements in the total system delay do not seem to be significant for congested traffic scenarios (Fig. 7), as the system has less flexibility in managing the allocation of the vehicle resources compared to other scenarios. In other words, in such congested cases, the proposed system tends to behave similarly to the conventional bus dispatching system, especially since we define that the capacity of the modular bus containing the maximum number of modular units is the same as that of the conventional bus. Nevertheless, the operator cost function still gets significantly improved (Fig. 8). Potential improvements in the operator cost function can be up to 58% for simulated scenarios. This shows that the proposed system provides more flexibility for dispatching buses to serve the passenger demand while reducing the operator cost compared to the conventional bus dispatching system, even for congested traffic conditions.

It is also worth mentioning that the minimum tested penetration rate of modular units (10%) exhibits a substantial reduction in the total cost. This suggests that the total system cost can be significantly reduced even when only a small fraction of vehicles within the bus fleet are modular, indicating the value of the modular vehicle technology in the early deployment stages. As the penetration rate of modular bus units increases, the marginal benefits of the modular vehicle technology decreases. Interestingly, there are only negligible improvements when the penetration rate is increased from 50% to 100%. This holds true for all tested scenarios (see Fig. 5).

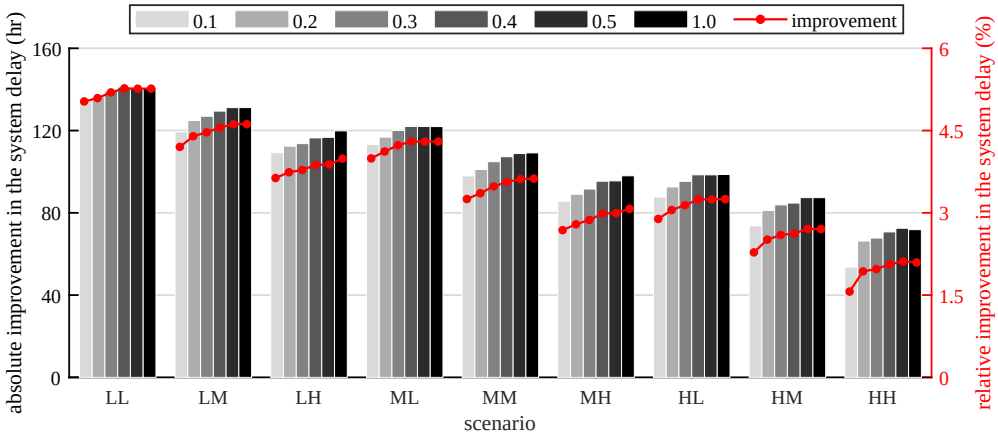


Figure 7: Comparison of the improvement in system delay for different penetration rates of modular units.

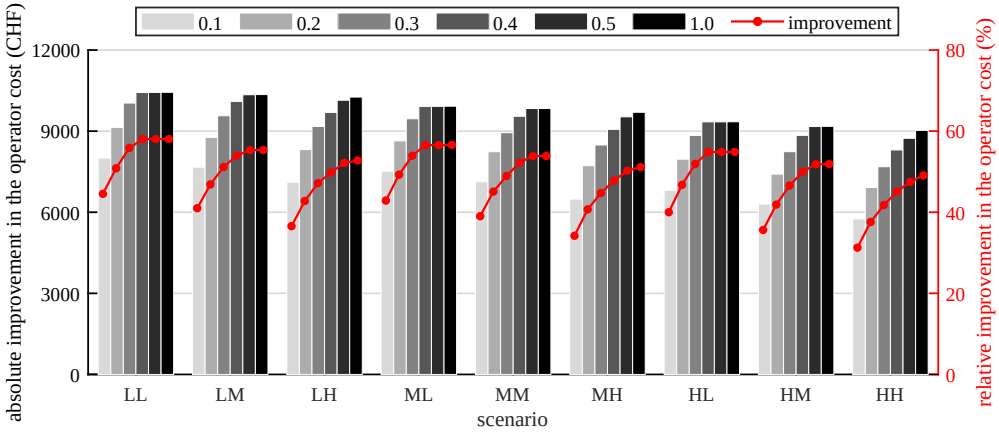


Figure 8: Comparison of the improvement in the operator cost for different penetration rates of modular units.

### *4.3. Value of considering multimodal interactions and congestion propagation using the 3D-MFD*

In this subsection, we quantify the value of integrating the 3D-MFD into the modeling framework. Recall that, to realistically model the proposed system, we employed the 3D-MFD to take into account factors such as the complex multimodal interactions and congestion propagation. These factors, however, are ignored in most scientific literature on the frequency setting problem, which typically assumes that the travel times (or equivalently, the bus speeds) are independent of the bus dispatching policy.

The value of employing the 3D-MFD is quantified by comparing the results of the proposed approach to those obtained when the optimal solution for the simplified problem (i.e. considering the bus system only) is incorporated into the proposed modeling framework. In other words, in such a simplified problem, we find the optimal bus dispatching policy ($\mathbf{\Psi}^*$ and $\boldsymbol{\psi}^*$) by considering neither the interactions with car traffic nor congestion propagation based on the 3D-MFD. This optimal bus dispatching policy is then used as an input into the proposed modeling framework to compute the total cost $Z^*$. The value of considering multimodal interactions and congestion propagation using the 3D-MFD, $\Delta Z$, is computed as the relative difference between the optimal total cost determined using the proposed optimization framework (i.e. from Section 4.2), $Z$, and the value of $Z^*$, i.e. $\Delta Z = [Z - Z^*]/Z^*$. The results of this comparison are shown in Fig. 9.

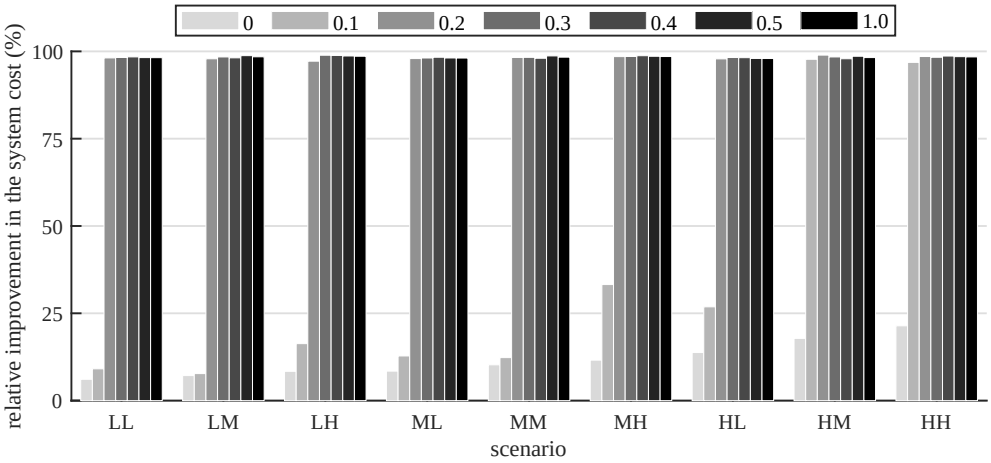


Figure 9: Comparison of the improvement in the system cost made by accounting for the complex multimodal interactions and congestion propagation for different penetration rates of modular units.

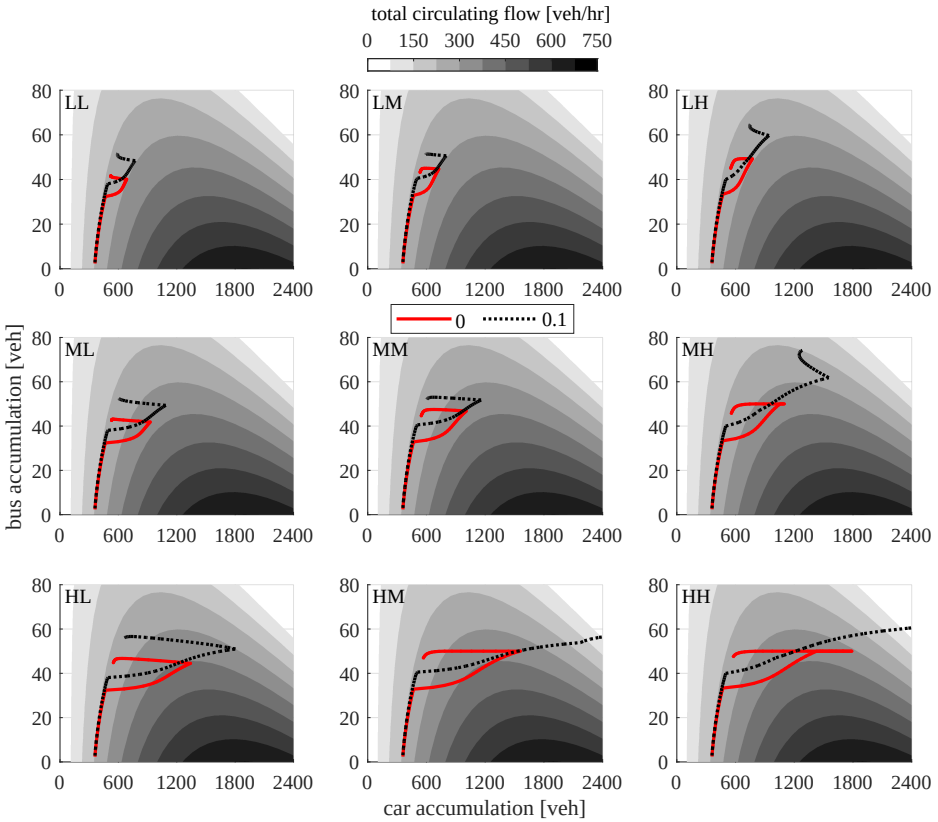


Figure 10: Evolution of the vehicle accumulation of both modes in the 3D-MFD plane for the tested traffic scenarios and two penetration rates of modular bus units ($p_m = 0$ and $p_m = 0.1$).

We observe that substantial improvements (up to 99%) in the total system cost can be achieved with a proper modeling of the complex multimodal interactions and congestion propagation. As expected, the lowest improvements are obtained for the scenarios with uncongested traffic conditions (i.e. low level of car demand) and zero penetration rate of modular units. The reason for this is twofold. First, in low car demand scenarios, the interactions between the

modes are minimized. As the level of interactions increases, the improvements increase. Second, for zero penetration rate of modular units, the system has fewer number of vehicles available to dispatch, reducing thereby the bus accumulation in the network and the associated negative effects on car traffic compared to the scenarios with non-zero penetration rates. This is illustrated in Fig. 10, which shows the evolution of both the bus and the car accumulation for the tested traffic scenarios when the optimal bus dispatching policy from the simplified problem ($\boldsymbol{\Psi}^*$ and $\boldsymbol{\psi}^*$) is incorporated into the proposed modeling framework. Note that, for brevity, we only show the results for the scenarios with the penetration rates of $p_m = 0$ and $p_m = 0.1$, as they exhibit the lowest improvements in Fig. 9. Notice from Fig. 10 that, without a proper modeling of multimodal interactions and congestion propagation based on the 3D-MFD, the vehicular accumulation of both modes significantly increases once the modular units are introduced to the system, especially for the congested traffic scenarios. This is because the system has more flexibility in adjusting the number of combined modular bus units to the passenger demand, resulting in higher dispatching flow of buses compared to the scenarios with zero penetration rate. Consequently, the total circulating flow in the network is substantially reduced or, in some cases, even equivalent to zero (see e.g. very congested traffic scenarios HM and HH in Fig. 10). In fact, the scenarios in which the total circulating flow reaches zero value are the cases when the improvements tend to 99%. This happens not only for HM and HL scenarios with the penetration rate of $p_m = 0.1$ in Fig.10, but also for all tested scenarios with the penetration rate of modular units that is $p_m > 0.1$, as shown in Fig. 9. These problems, however, do not appear in the proposed optimization framework, indicating its value in capturing the necessary factors for optimizing the performance of the whole network, while taking into account all transport modes.

### *4.4. Effects of the operating unit cost of modular bus units*

In the previous subsections, we used the value of $\pi_m \approx 30$ CHF/hr as the operating unit cost of modular bus units to quantify the performance of the proposed flexible bus dispatching system. This value was determined following the recommendations by Sinner (2019). Given such value of $\pi_m$, the operating cost of the modular bus consisting of the maximum number of modular units that can be combined along a bus line (in this paper set to $|\mathcal{U}_{m,l}| = \zeta = 6,\ \forall l \in \mathcal{L}$) is lower than that of the conventional (articulated) bus. This is realistic, considering that the modular bus units are fully automated, i.e. there is no cost for assigning bus drivers to them, which in some cities represents the dominant element in the total operating unit cost of conventional buses (Sinner, 2019).

Nevertheless, one can pose the following question: What happens to the performance of the proposed dispatching system if we increase the operating unit cost of modular bus units? This question can be meaningful for two reasons: (i) it is possible that the cost of modular units is high before the technology gets fully mature; and (ii) the analysis of this question can shed light on addressing the scenarios where safety drivers are required for these units in the early stage of deployment. To address this question, in this subsection we conduct additional experiments for the following two cases: (i) $\pi_m = 60$ CHF/hr; and (ii) $\pi_m = 120$ CHF/hr. Note that in the first/second case, four/two combined modular bus units yields the same operating cost as that of the conventional bus. Consequently, the operating cost of the modular bus consisting of the maximum number of modular units is, in both cases, significantly higher than that of the conventional bus. Although these two cases are rather unrealistic, they help to shed light on the robustness of the proposed optimization framework.

Table 2: Effects of the operating unit cost of modular bus units on the cost functions ($p_m = 0.1$).

| | total cost | | | operator cost | | | system delay | | |
|---|---|---|---|---|---|---|---|---|---|
| $\pi_m$ | 30 | 60 | 120 | 30 | 60 | 120 | 30 | 60 | 120 |
| LL | -16.0% | -12.3% | -7.8% | -44.5% | -34.1% | -18.8% | -5.0% | -5.0% | -4.2% |
| LM | -13.3% | -10.9% | -6.9% | -41.0% | -30.9% | -16.8% | -4.2% | -4.3% | -3.6% |
| LH | -11.7% | -9.5% | -5.9% | -36.9% | -27.6% | -14.7% | -3.6% | -3.7% | -3.1% |
| ML | -13.2% | -10.7% | -6.9% | -42.9% | -32.5% | -18.1% | -4.0% | -4.0% | -3.4% |
| MM | -11.6% | -9.3% | -5.9% | -39.0% | -29.2% | -16.1% | -3.3% | -3.3% | -2.8% |
| MH | -9.9% | -8.0% | -5.0% | -34.2% | -25.8% | -13.9% | -2.7% | -2.8% | -2.3% |
| HL | -11.0% | -9.0% | -5.8% | -40.0% | -30.4% | -17.3% | -2.9% | -3.0% | -2.5% |
| HM | -9.5% | -7.7% | -4.8% | -35.6% | -26.7% | -14.8% | -2.3% | -2.4% | -2.1% |
| HH | -7.9% | -6.4% | -3.9% | -31.3% | -23.2% | -12.5% | -1.6% | -1.8% | -1.6% |

The results are shown in Table 2, comparing the performance of the proposed bus dispatching system with the

penetration rate of $p_m = 0.1$ and the operating unit costs of $\pi_m = 30$ CHF/hr, 60 CHF/hr, and 120 CHF/hr. Note that for the last two cases the system provides a very similar bus dispatching policy for all penetration rates under a given operating unit cost of modular bus units (either 60 CHF/hr or 120 CHF/hr). In other words, the results for the two tested cases with higher $\pi_m$ are almost insensitive to the penetration rate of modular bus units. This, in turn, implies that the marginal benefits of additional modular units are negligible. Nonetheless, notice that, even when the operating unit cost of modular bus units is significantly increased, the proposed system still substantially outperforms the existing bus dispatching system with only conventional buses. As expected, with an increase in the operating unit cost of modular bus units, the improvements (in particular those from the operator perspective) reduce. Finally, in terms of the average number of combined modular bus units, the analysis reveals that the system recognizes higher costs for combining more modular units, especially in case of $\pi_m = 120$ CHF/hr. Consequently, the capacity of modular buses is reduced compared to the case of $\pi_m = 30$ CHF/hr with $p_m = 0.1$. That being said, in case of $\pi_m = 60$ CHF/hr, the system dispatches up to two combined modular bus units, whereas in case of $\pi_m = 120$ CHF/hr the system operates mostly with individual modular units. We omit those results here for brevity.

### *4.5. Effects of the size of the bus network and the bus fleet*

Note that, in the previous analyses, we used a rather larger bus network (in terms of the number of bus lines) inspired by that of the City of Zurich, Switzerland, as it exhibits a well-defined empirical 3D-MFD. Although a higher number of bus lines provides more flexibility in adjusting the number of combined modular units and their dispatching frequencies to the passenger demand, it also adds more constraints to the formulated optimization problem. Therefore, another interesting question that can be posed is: How well does the proposed system perform when implemented on a smaller bus network?

To address this question, in this subsection we apply the proposed optimization framework on the bus network operating with two bus lines. For the purpose of computing the operator cost, we use the initial operating unit cost of modular bus units of $\pi_m = 30$ CHF/hr. To analyze how sensitive the results are to the size of the bus fleet, we analyze the following two cases indicating the number of conventional buses within the bus fleet for zero penetration rate of modular units: (i) $F = 8$; and (ii) $F = 16$. For brevity, in both cases, we only investigate the penetration rates of modular unit of $p_m \in \{0.1, 0.2, 0.3, 0.4\}$. Recall that for the scenarios with $p_m > 0$, the number of conventional buses is given as $F_r = F \cdot (1 - p_m)$.

Table 3: Effects of the size of the bus network and the bus fleet on the cost functions.

| | | total cost | | | | operator cost | | | | system delay | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $F$ | $p_m$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.1 | 0.2 | 0.3 | 0.4 | 0.1 | 0.2 | 0.3 | 0.4 |
| 8 | LL | -14.9% | -27.2% | -31.1% | -32.1% | -35.8% | -69.0% | -70.7% | -72.7% | -6.2% | -9.9% | -14.7% | -15.3% |
| | LM | -14.5% | -26.6% | -30.0% | -31.0% | -36.7% | -61.5% | -66.7% | -69.1% | -5.8% | -12.8% | -15.5% | -15.9% |
| | LH | -14.6% | -25.2% | -29.0% | -30.0% | -37.7% | -56.6% | -62.7% | -65.0% | -6.1% | -13.6% | -16.6% | -17.0% |
| | ML | -14.5% | -26.1% | -29.7% | -30.6% | -35.8% | -64.1% | -70.3% | -72.8% | -5.7% | -10.5% | -13.1% | -13.3% |
| | MM | -13.9% | -24.9% | -28.4% | -29.4% | -36.0% | -58.8% | -65.7% | -68.8% | -5.5% | -12.0% | -14.2% | -14.4% |
| | MH | -14.8% | -24.4% | -27.6% | -28.5% | -36.4% | -57.6% | -60.8% | -63.3% | -7.3% | -12.9% | -16.1% | -16.4% |
| | HL | -14.3% | -25.3% | -28.2% | -29.1% | -38.7% | -65.6% | -69.9% | -72.5% | -4.7% | -9.4% | -11.8% | -12.0% |
| | HM | -13.6% | -24.0% | -27.0% | -28.0% | -34.8% | -59.0% | -65.0% | -67.9% | -6.0% | -11.5% | -13.4% | -13.7% |
| | HH | -14.4% | -23.5% | -26.1% | -27.1% | -37.0% | -59.4% | -63.3% | -65.2% | -6.3% | -10.7% | -12.8% | -13.4% |
| 16 | LL | -27.6% | -31.9% | -32.0% | -32.0% | -67.9% | -74.4% | -74.7% | -74.7% | -9.1% | -12.3% | -12.3% | -12.3% |
| | LM | -26.8% | -30.4% | -30.9% | -30.9% | -67.5% | -71.6% | -72.9% | -72.9% | -8.6% | -12.0% | -12.1% | -12.1% |
| | LH | -24.7% | -29.0% | -29.7% | -29.8% | -62.8% | -68.6% | -71.0% | -71.0% | -7.9% | -11.6% | -11.6% | -11.7% |
| | ML | -26.8% | -30.5% | -30.6% | -30.6% | -68.4% | -74.0% | -74.4% | -74.4% | -8.5% | -11.4% | -11.4% | -11.4% |
| | MM | -25.4% | -29.0% | -29.5% | -29.5% | -66.1% | -71.2% | -72.6% | -72.6% | -7.8% | -10.9% | -11.0% | -11.0% |
| | MH | -24.1% | -27.6% | -28.3% | -28.4% | -62.5% | -68.0% | -70.6% | -70.6% | -8.0% | -10.6% | -10.6% | -10.7% |
| | HL | -25.7% | -29.0% | -29.2% | -29.2% | -69.2% | -73.6% | -74.0% | -74.0% | -7.5% | -10.3% | -10.4% | -10.4% |
| | HM | -24.6% | -27.6% | -28.1% | -28.1% | -66.7% | -70.7% | -72.3% | -72.3% | -7.3% | -9.8% | -9.8% | -9.8% |
| | HH | -22.8% | -26.1% | -26.9% | -27.0% | -62.6% | -67.4% | -70.0% | -70.2% | -6.7% | -9.5% | -9.5% | -9.6% |

The results are shown in Table 3, comparing the performance of the proposed flexible bus dispatching system with different penetration rates of modular units to that of the existing dispatching system (i.e. consisting only of conventional buses). Notice that, even in case of smaller bus fleet, the proposed system significantly outperforms

the existing one, especially from the operator perspective. As the size of the bus fleet increases, the improvements increase. Potential reduction in the total and the operator cost function can be up to 32% and 75%, respectively, depending on the level of car and passenger demand, penetration rate of modular bus units, and the size of the bus fleet. Interestingly, there seems to be a critical number of modular bus units, after which the marginal benefits of additional modular units are negligible (observe similar improvements in the total system cost for $p_m \geq 0.4$ and $p_m \geq 0.2$ in case of $F = 8$ and $F = 16$, respectively). This indicates the robustness of the proposed optimization framework.

## 5. Conclusion

In this study, we propose a novel concept, called *flexible bus dispatching system*, which offers new perspectives and enormous flexibility to better manage the dispatching frequencies and the allocation of the vehicle resources, reducing thereby the operating cost. In such a flexible bus dispatching system, the bus fleet consists not only of conventional buses, but also of modular and fully automated bus units that can either operate individually or combined together (forming thereby a single modular unit of a higher passenger capacity). To determine the optimal number of combined modular bus units and the optimal frequency at which the units (both conventional and modular) should be dispatched across different bus lines, while accounting for the traffic dynamics at the network level, we propose an optimization framework based on the recently proposed three-dimensional macroscopic fundamental diagram (3D-MFD). To the best of our knowledge, this is the first application of the 3D-MFD and modular bus units for the frequency setting problem in the domain of bus operations.

To test the performance of the proposed system, we analyze various scenarios, characterizing different levels of car and passenger demand, and penetration rates of modular bus units. Numerical results show that the proposed concept can significantly outperform the existing bus dispatching system with only conventional buses. The improvements in the total system cost are achieved by adapting the number of combined modular bus units and their dispatching frequencies to the evolution of both, the car and the passenger demand. For example, in case of low passenger demand (typically during the off-peak period), the proposed system dispatches compositions of modular bus units that contain only few combined (if not individual) modular units. However, in case of high passenger demand (typically during the peak hour), the dispatching compositions include higher number of combined modular bus units. This way, the proposed system makes a trade-off between the service frequency and the allocation of bus units on the one hand, and the level of service provided to users on the other hand. Moreover, a comparison with the approach that considers only the bus system (neglecting the complex multimodal interactions and congestion propagation dynamics) reveals the value of the proposed modeling framework. Finally, by studying the effect of the operating unit cost of modular bus units, the size of the bus network, and the size of the bus fleet, we shed light on the robustness of the proposed optimization framework.

This work opens several research directions. First, we are interested in integrating the proposed bus dispatching system with other traffic management strategies, e.g. a bi-modal perimeter control (Dakic et al., 2019; He et al., 2019), to further improve the operations of bi-modal traffic systems. Second, we would like to extend the proposed methodological framework to more complex scenarios (e.g. multiple-region networks) and develop a hierarchical control framework (Yang et al., 2017) that converts the macroscopic level bus dispatching decision to account for more detailed operational features (e.g. timetables, en-route combining/splitting of modular units along a bus line). Third, the en-route combining/splitting features can be used to further develop other bus-related strategies (e.g. stop skipping and dynamic routing). Fourth, it would be interesting to incorporate the automated modular bus units into other promising paradigms (e.g. shared mobility). For example, we can design real-time demand responsive bus services with these units to further improve passenger mobility. We may also integrate the proposed system with mobility-on-demand systems, where the public transport operator maintains a larger fleet of automated modular units and some units provide last-mile on-demand services. Fifth, it would be meaningful to evaluate the long-term benefits of the proposed strategy from a planning perspective by accounting for shift in mode choice, route choice, and departure time.

## 6. Acknowledgments

This work was supported by the Swiss National Science Foundation (SNSF) under the project name DIPLOMAT, contract 205121_165644. Kaidi Yang acknowledges funding support by the Swiss National Science Foundation (SNSF) Early Postdoc.Mobility with project number P2EZP2_184244. Monica Menendez acknowledges funding support by the NYUAD Center for Interacting Urban Networks (CITIES), funded by Tamkeen under the NYUAD Research Institute Award CG001 and by the Swiss Re Institute under the Quantum Cities™ initiative.

## References

Ambühl, L., Loder, A., Menendez, M., Axhausen, K.W., 2017. Empirical Macroscopic Fundamental Diagrams: New Insights from Loop Detector and Floating Car Data. Presented at 96th Annual Meeting of the Transportation Research Board .

Ambühl, L., Loder, A., Zheng, N., Axhausen, K.W., Menendez, M., 2019. Approximative network partitioning for mfds from stationary sensor data. Transportation Research Record 2673, 94–103.

Ampountolas, K., Zheng, N., Geroliminis, N., 2017. Macroscopic modelling and robust control of bi-modal multi-region urban road networks. Transportation Research Part B: Methodological 104, 616–637.

Axhausen, K.W., König, A., Abay, G., Bates, J.J., Bierlaire, M., 2006. Swiss value of travel time savings. Arbeitsberichte Verkehrs-und Raumplanung 253.

Badia, H., Estrada, M., Robusté, F., 2016. Bus network structure and mobility pattern: A monocentric analytical approach on a grid street layout. Transportation Research Part B: Methodological 93, 37–56.

Boggs, P.T., Tolle, J.W., 1995. Sequential quadratic programming. Acta numerica 4, 1–51.

Cacchiani, V., Caprara, A., Toth, P., 2010. Solving a real-world train-unit assignment problem. Mathematical Programming 124, 207–231.

Cadarso, L., Marín, Á., Maróti, G., 2013. Recovery of disruptions in rapid transit networks. Transportation Research Part E: Logistics and Transportation Review 53, 15–33.

Cats, O., Haverkamp, J., 2018. Optimal infrastructure capacity of automated on-demand rail-bound transit systems. Transportation Research Part B: Methodological 117, 378–392.

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

Ceder, A., Wilson, N.H., 1986. Bus network design. Transportation Research Part B: Methodological 20, 331–344.

Chen, Z., Li, X., Zhou, X., 2019. Operational design for shuttle systems with modular vehicles under oversaturated traffic: Discrete modeling method. Transportation Research Part B: Methodological 122, 1–19.

Chen, Z., Li, X., Zhou, X., 2020. Operational design for shuttle systems with modular vehicles under oversaturated traffic: Continuous modeling method. Transportation Research Part B: Methodological 132, 76–100.

Chiabaut, N., Küng, M., Menendez, M., Leclercq, L., 2018. Perimeter control as an alternative to dedicated bus lanes: A case study. Transportation Research Record 2672, 110–120.

Cordeau, J.F., Soumis, F., Desrosiers, J., 2001. Simultaneous Assignment of Locomotives and Cars to Passenger Trains. Operations Research 49, 531–548.

Daganzo, C.F., 1994. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. Transportation Research Part B: Methodological 28, 269–287.

Daganzo, C.F., 2007. Urban gridlock: Macroscopic modeling and mitigation approaches. Transportation Research Part B: Methodological 41, 49–62.

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

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

Dakic, I., Ambühl, L., Schümperlin, O., Menendez, M., 2020. On the modeling of passenger mobility for stochastic bi-modal urban corridors. Transportation Research Part C: Emerging Technologies 113, 146–163.

Dakic, I., Menendez, M., 2018. On the use of Lagrangian observations from public transport and probe vehicles to estimate car space-mean speeds in bi-modal urban networks. Transportation Research Part C: Emerging Technologies 91, 317–334.

Dakic, I., Yang, K., Menendez, M., 2019. Evaluating the effects of passenger occupancy dynamics on a bi-modal perimeter control, in: TRB Annual Meeting Online, Transportation Research Board.

De Montjoye, Y.A., Hidalgo, C.A., Verleysen, M., Blondel, V.D., 2013. Unique in the crowd: The privacy bounds of human mobility. Scientific reports 3, 1376.

Desaulniers, G., Hickman, M.D., 2007. Public transit. Handbooks in operations research and management science 14, 69–127.

Eberlein, X.J., Wilson, N.H., Bernstein, D., 2001. The holding problem with real–time information available. Transportation science 35, 1–18.

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. Procedia - Social and Behavioral Sciences 17, 113–135.

Fioole, P.J., Kroon, L., Maróti, G., Schrijver, A., 2006. A rolling stock circulation model for combining and splitting of passenger trains. European Journal of Operational Research 174, 1281–1297.

Furth, P.G., Wilson, N.H.M., 1981. Setting frequencies on bus routes: theory and practice. Transportation Research Record Journal of the Transportation Research Board 818, 1–7.

Ge, Q., Ciuffo, B., Menendez, M., 2015. Combining screening and metamodel-based methods: An efficient sequential approach for the sensitivity analysis of model outputs. Reliability Engineering & System Safety 134, 334–344.

Ge, Q., Menendez, M., 2017. Extending morris method for qualitative global sensitivity analysis of models with dependent inputs. Reliability Engineering & System Safety 162, 28–39.

Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. Transportation Research Part B: Methodological 42, 759–770.
Geroliminis, N., Zheng, N., Ampountolas, K., 2014. A three-dimensional macroscopic fundamental diagram for mixed bi-modal urban networks. Transportation Research Part C: Emerging Technologies 42, 168–181.
Gkiotsalitis, K., Cats, O., 2018. Reliable frequency determination: Incorporating information on service uncertainty when setting dispatching headways. Transportation Research Part C: Emerging Technologies 88, 187–207.
Guo, Q.W., Chow, J.Y.J., Schonfeld, P., 2018. Stochastic dynamic switching in fixed and flexible transit services as market entry-exit real options. Transportation Research Part C: Emerging Technologies 94, 288–306.
Hadas, Y., Shnaiderman, M., 2012. Public-transit frequency setting using minimum-cost approach with stochastic demand and travel time. Transportation Research Part B: Methodological 46, 1068–1084.
Han, A.F., Wilson, N.H., 1982. The allocation of buses in heavily utilized networks with overlapping routes. Transportation Research Part B: Methodological 16, 221–232.
He, H., Yang, K., Liang, H., Menendez, M., Guler, S.I., 2019. Providing public transport priority in the perimeter of urban networks: A bimodal strategy. Transportation Research Part C: Emerging Technologies 107, 171–192.
Hickman, M.D., 2001. An analytic stochastic model for the transit vehicle holding problem. Transportation Science 35, 215–237.
Ibarra-Rojas, O., Delgado, F., Giesen, R., Muñoz, J., 2015. Planning, operation, and control of bus transport systems: A literature review. Transportation Research Part B: Methodological 77, 38–75.
Krugman, P., 1991. Geography and trade MIT Press. Cambridge, MA .
Li, Y., Xu, W., He, S., 2013. Expected value model for optimizing the multiple bus headways. Applied Mathematics and Computation 219, 5849–5861.
Lingaya, N., Cordeau, J.F., Desaulniers, G., Desrosiers, J., Soumis, F., 2002. Operational car assignment at VIA Rail Canada. Transportation Research Part B: Methodological 36, 755–778.
Liu, W., Geroliminis, N., 2017. Doubly dynamics for multi-modal networks with park-and-ride and adaptive pricing. Transportation Research Part B: Methodological 102, 162–179.
Loder, A., Ambühl, L., Menendez, M., Axhausen, K.W., 2017. Empirics of multi-modal traffic networks - Using the 3D macroscopic fundamental diagram. Transportation Research Part C: Emerging Technologies 82, 88–101.
Loder, A., Ambühl, L., Menendez, M., Axhausen, K.W., 2019a. Understanding traffic capacity of urban networks. Scientific reports 9, 1–10.
Loder, A., Dakic, I., Bressan, L., Ambühl, L., Bliemer, M.C., Menendez, M., Axhausen, K.W., 2019b. Capturing network properties with a functional form for the multi-modal macroscopic fundamental diagram. Transportation Research Part B: Methodological 129, 1–19.
Mahmassani, H.S., Saberi, M., Zockaie, A., 2013. Urban network gridlock: Theory, characteristics, and dynamics. Transportation Research Part C: Emerging Technologies 36, 480–497.
Martínez, H., Mauttone, A., Urquhart, M.E., 2014. Discrete Optimization Frequency optimization in public transportation systems: Formulation and metaheuristic approach. European Journal of Operational Research 236, 27–36.
Muñoz, J.C., Cortés, C.E., Giesen, R., Sáez, D., Delgado, F., Valencia, F., Cipriano, A., 2013. Comparison of dynamic control strategies for transit operations. Transportation Research Part C: Emerging Technologies 28, 101–113.
Newell, G.F., 1971. Dispatching Policies for a Transportation Route. Transportation Science 5, 91–105.
Paipuri, M., Leclercq, L., 2020a. Bi-modal macroscopic traffic dynamics in a single region. Transportation research part B: methodological 133, 257–290.
Paipuri, M., Leclercq, L., 2020b. Empirical validation of bimodal mfd models. Frontiers in Future Transportation 1, 1.
Roca-Riu, M., Menendez, M., Dakic, I., Buehler, S., Ortigosa, J., 2020. Urban space consumption of cars and buses: an analytical approach. Transportmetrica B: Transport Dynamics 8, 237–263.
Saeedmanesh, M., Geroliminis, N., 2017. Dynamic clustering and propagation of congestion in heterogeneously congested urban traffic networks. Transportation Research Part B: Methodological 105, 193–211.
Salzborn, F.J.M., 1972. Optimum Bus Scheduling. Transportation Science 6, 137–148.
Salzborn, F.J.M., 1980. Scheduling Bus Systems with Interchanges. Transportation Science 14, 211–231.
Schéele, S., 1980. A supply model for public transit services. Transportation Research Part B: Methodological 14, 133–146.
Sinner, M., 2019. Effects of the Autonomous Bus on the Railway System. Ph.D. thesis. ETH Zurich.
Sirmatel, I.I., Geroliminis, N., 2018. Mixed logical dynamical modeling and hybrid model predictive control of public transport operations. Transportation Research Part B: Methodological 114, 325–345.
Szeto, W.Y., Wu, Y., 2011. A simultaneous bus route design and frequency setting problem for tin shui wai, hong kong. European Journal of Operational Research 209, 141–155.
Verbas, I.Ö., Frei, C., Mahmassani, H.S., Chan, R., 2015. Stretching resources: sensitivity of optimal bus frequency allocation to stop-level demand elasticities. Public Transport 7, 1–20.
Verbas, I.Ö., Mahmassani, H.S., 2013. Optimal Allocation of Service Frequencies over Transit Network Routes and Time Periods. Transportation Research Record: Journal of the Transportation Research Board 2334, 50–59.
Vuchic, V.R., Day, F.B., Dirshimer, G.N., Kikuchi, S., Rudinger, D.J., 1978. Transit operating manual .
Wang, Y., D'Ariano, A., Yin, J., Meng, L., Tang, T., Ning, B., 2018. Passenger demand oriented train scheduling and rolling stock circulation planning for an urban rail transit line. Transportation Research Part B: Methodological 118, 193–227.
Xuan, Y., Argote, J., Daganzo, C.F., 2011. Dynamic bus holding strategies for schedule reliability: Optimal linear control and performance analysis. Transportation Research Part B: Methodological 45, 1831–1845.
Yang, K., Menendez, M., Zheng, N., 2019. Heterogeneity aware urban traffic control in a connected vehicle environment: a joint framework for congestion pricing and perimeter control. Transportation Research Part C: Emerging Technologies 105, 439–455.
Yang, K., Zheng, N., Menendez, M., 2017. Multi-scale perimeter control approach in a connected-vehicle environment. Transportation Research Part C: Emerging Technologies 94, 32–49.
Yue, Y., Han, J., Wang, S., Liu, X., 2017. Integrated Train Timetabling and Rolling Stock Scheduling Model Based on Time-Dependent Demand

for Urban Rail Transit. Computer-Aided Civil and Infrastructure Engineering 32, 856–873.
Zhang, F., Liu, W., 2019. Responsive bus dispatching strategy in a multi-modal and multi-directional transportation system: A doubly dynamical approach. Transportation Research Part C: Emerging Technologies .
Zheng, N., Geroliminis, N., 2013. On the distribution of urban road space for multimodal congested networks. Transportation Research Part B: Methodological 57, 326–341.