Transportation Research Part C 175 (2025) 105119

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Exploiting the flexibility of modular buses in an urban transit system

Carlo Filippi, Gianfranco Guastaroba, Lorenzo Peirano *, M. Grazia Speranza

University of Brescia, Department of Economics and Management, Brescia, Italy

ARTICLE INFO

Keywords:
Modular buses
Flexible-capacity vehicles
Transit systems
Integer linear programming

ABSTRACT

Urban transit systems usually operate according to fixed-route and fixed-schedule schemes by employing fixed-capacity vehicles, despite the mobility demand is unevenly spread out in both space and time. Modular buses are an emerging technology in which modules of relatively small capacity can be dynamically docked together to form greater capacity buses and can, therefore, make the transit system capable of adapting the capacity to the actual mobility demand. A module can be shifted from one line to another at pre-defined intersections and can be relocated when empty, if beneficial. We call these two operations sharing and rebalancing, respectively. Given a transit network comprising multiple bus lines and a mobility demand, we present an integer linear program to determine an optimal assignment of modules to lines, so that the mobility demand is met with a minimum total number of modules. Computational experiments show that, by exploiting the flexibility of modular buses, the total capacity deployed can be reduced by 49% with respect to a conventional transit system, whereas the average occupancy ratio increases from 41.22% to 72.85%.

1. Introduction

As of 2023, roughly 56% of the world population – around 4.4 billion inhabitants – live in urban areas, a percentage that is expected to increase to 70% by 2050.[1] This continuously growing level of urbanization poses several challenges, including how to meet the ever-increasing human mobility demand. As urban areas become more densely populated and more congested, a growing percentage of urban population lean upon transit systems rather than their private vehicles, especially for their regular trips (e.g., for home-work commuting trips). This trend is strengthened by a rising awareness of the environmental impacts that conventional internal combustion engine vehicles have. Furthermore, transit systems support the well-functioning of societies by serving low-income communities, providing them with an affordable alternative to private vehicles.

Most of the modern transit systems – such as buses, trains, subways, and light rail vehicles – operate according to a Fixed-Route and Fixed-Schedule (FRFS) scheme. This approach simplifies the design of the service for transit authorities, and also offers simple to implement schedules that are easy to comprehend for the passengers. However, services running along a FRFS scheme may not be attractive to a considerable percentage of the mobility demand due to the lack of comfort and flexibility. Recently, a number of researchers have analyzed innovative transit systems that are more capable than FRFS systems of adapting the service provided to the passenger demand. Along this line of research, we mention the customized buses (e.g., see Gong et al., 2021) and the demand-responsive transit systems (Lee et al., 2021). The reader interested in an overview of the recent trends emerging in the literature on flexible passengers transit systems is referred to Filippi et al. (2023).

* Corresponding author.
E-mail addresses: carlo.filippi@unibs.it (C. Filippi), gianfranco.guastaroba@unibs.it (G. Guastaroba), lorenzo.peirano@unibs.it (L. Peirano), grazia.speranza@unibs.it (M.G. Speranza).

[1] https://www.worldbank.org/en/topic/urbandevelopment/overview (Last accessed: 26th November 2024)

https://doi.org/10.1016/j.trc.2025.105119
Received 18 July 2024; Received in revised form 28 January 2025; Accepted 29 March 2025
Available online 16 April 2025
0968-090X/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Fig. 1. A modular bus designed by NExT S.r.l..

Besides running along a FRFS scheme, the majority of the current urban transit systems employ fixed-capacity vehicles. This setting is often unable to provide an effective service to a mobility demand that is spread out unevenly in space and time. In fact, the mobility demand is well-known to be concentrated during certain periods of the day, as highlighted in Chapuis et al. (2020). Consequently, current transit systems are usually overcrowded during rush hours and underutilized in off-peak hours, leading to a shortage of vehicles during the former periods and an excessive number of vehicles operating during the latter periods (Xia et al., 2023). To mitigate these shortcomings, the common approach is to modify the frequency of the service throughout the day. However, this produces long headways (i.e., long interarrival times between vehicles) for passengers during off-peak hours. Additionally, because of the small mobility demand arising, isolated urban areas are often not effectively served.

Modular vehicles are an emerging technology that has the potential to mitigate the main issues discussed above concerning fixed-capacity vehicles. Our main interest is in the use of *modular buses* in an urban transit system, such as those designed by the Italian company NExT S.r.l..[2] Each modular bus is composed of multiple modules (also called pods). As of the beginning of 2024, up to 5 modules can be joined together. Fig. 1 shows a modular bus comprising 2 modules. The modules are electric vehicles, and the current version (called NX24) has a driving range of 300 kilometers.[3] Each module can accommodate up to 15 passengers and is capable of docking to (and undocking from) other modules. It is worth noting that these modules are capable of docking and undocking with each other without stopping, so that the delay incurred to carry out these operations is negligible. When joined together, modules create an open bus-like corridor, allowing passengers to move from one module to another. Each module can be driven separately, and in certain conditions (e.g., in suitable private areas) in fully driverless and automated mode. The full potential of these modular buses will be achieved in the future, when fully automated trips will be possible (and lawfully allowed). The possibility of docking and undocking modules makes this technology more capable of dynamically adapting the capacity deployed to the mobility demand. Indeed, modular buses comprising several modules can be employed in certain segments of a bus line during peak hours, where and when the mobility demand is concentrated, thereby providing a service similar to conventional fixed-capacity buses. Modules can also be detached to create small modular buses that serve low-demand areas or run during off-peak hours, while maintaining a regular service frequency. Additionally, when idle or during relocation trips, a module can possibly be used as a ride-sharing vehicle, as proposed, among others, in Zhang et al. (2020), Gong et al. (2021), Liu et al. (2021), Hannoun and Menendez (2022) and Guo et al. (2023). Given all the aforementioned advantages, this emerging technology is attracting increasing attention among academic researchers, as discussed below in Section 2.

In this paper, we consider a given transit network with multiple lines and a given mobility demand. Our main objective is to show that modular buses can be a sound alternative to conventional fixed-capacity buses, taking advantage of their capability to flexibly adapt the capacity to the actual demand, and, therefore, they can offer a better service with a smaller total capacity deployed. To this aim, we develop an Integer Linear Program (ILP) to determine an optimal assignment of modules to lines, so that the mobility demand given is all met by employing a minimum total number of modules. The proposed ILP enables the transit system to adapt the capacity along lines by sharing modules among lines according to the passengers demand. Modules can be docked and undocked only at given stops of the network. Empty modules can be moved between stops to rebalance the capacity. The proposed ILP can support decision-makers (e.g., a transit authority) in the evaluation of the strategic-level decision of employing a fleet of modular,

[2] https://www.next-future-mobility.com/scenarios (Last accessed: 26th November 2024)
[3] https://www.next-future-mobility.com/nextisnow (Last accessed: 26th November 2024)

rather than conventional, buses. Given this strategic-level perspective, in the proposed ILP, several day-to-day issues, which must be incorporated when adopting an operational-level viewpoint, are neglected.

Extensive computational results show that the total capacity deployed with the flexible setting of modular buses is reduced, on average, by 49% with respect to a conventional transit system with fixed-capacity buses.

Structure of the paper. The paper is structured as follows. In Section 2, the literature most closely related to our research is reviewed, and the contributions of this paper are highlighted. Section 3 provides the problem description, along with the main mathematical notation. A mathematical formulation of the problem is detailed in Section 4. In Section 5 we analyze two extreme cases to illustrate, on idealized conditions, the potential reduction in the number of modules deployed. The potential advantages of module flexibility are empirically highlighted in Section 6 through an extensive experimental analysis. Some concluding remarks are outlined in Section 7.

2. Literature review

This section reviews the literature most closely related to our research. As mentioned above, modular buses are an emerging technology. Consequently, the related literature is very recent and limited. The readers interested in more general aspects related to line and transit network design problems are referred to the surveys published by Schöbel (2012) and Guihaire and Hao (2008). We conclude this section by pointing out our contributions.

To the best of our knowledge, the first paper analyzing the potential advantages that can be achieved by adopting modular buses is Gecchelin and Webb (2019). In their research, the authors highlight the potential congestion reduction achievable by decreasing the size and number of vehicles employed at particular periods of time, as well as the prospective cost-effectiveness of this form of transport. Motivated by these potential advantages, an increasing number of researchers studied the use of modular buses as an alternative to, or in combination with, conventional buses. Table 1 provides a summary of the main characteristics of the papers reviewed in the following and that address an optimization problem arising with the use of modular buses. These main characteristics include some details on the type of problem studied, the objective optimized, and the solution method employed. The meaning of the abbreviations used in Table 1 are explained in the following.

Several researchers tackled the problem of jointly optimizing modular bus composition (often also called formation), i.e., the number of modules composing a modular bus, and the headway between bus departures. This class of problems is hereafter referred to as the class of Joint Headway and Composition Optimization (*JHCO*) problems. In contrast with the strategic-level viewpoint adopted in the present research, these problems have an operational-level perspective: Given a fleet of modular vehicles, these problems decide upon the number of them composing each modular bus along with the headway between bus departures. The basic variant for this class is defined on a single-line corridor traversed by buses in only one direction. In the majority of the related papers, the objective function is represented by the minimization of the total cost of the system usually given by two components: the costs of operating the transit system and passengers-related costs. These two cost components are denoted in Table 1 as *Op* and *Pas*, respectively. The following papers tackle such a basic variant or a simple extension. In two related papers, Chen et al. (2019, 2020) investigate a basic JHCO problem under over-saturated traffic conditions. In their research, decisions taken in one time period directly impact the environment under which decisions must be taken in subsequent time periods. In the former article, the authors develop a Mixed Integer Linear Program (MILP) to capture the problem, whereas in the latter paper a continuum approximation model is proposed. Dai et al. (2020) develop an integer nonlinear model for a JHCO problem where vehicle routes can be adapted in order to adjust the dispatch headway and serve the mobility demand by means of a hybrid fleet comprising modular and conventional buses. The authors study a transit network comprising a single line that departs from one origin and travels toward multiple destinations. Chen and Li (2021) introduce the expression of station-wise docking where modules can dock and undock at any stop along their route. To show the benefits that such flexibility can produce, they develop a MILP for a JHCO problem that explicitly incorporates passengers' behaviors in terms of origin and destination of their trips, mode preferences, and time of departure. Shi and Li (2021) develop a mathematical formulation for a JHCO problem that also considers first-in, first-out passenger queuing rules. The experimental analysis highlights the potential improvements of employing modular buses, in terms of both costs and levels of service provided to passengers. Tian et al. (2023) address a basic JHCO problem where the mobility demand is time-dependent. Additionally, the authors explicitly consider the limited availability of the modules at stations, as well as the incurred rebalancing costs of the modules. For this problem, the authors devise a mixed-integer nonlinear program. A data-driven distributionally robust optimization approach is applied in Xia et al. (2023) for a JHCO problem where both travel times and passenger demands are time-dependent.

Liu et al. (2023) address an extension of the basic JHCO problem where the single-line corridor is traversed in both ways, and modules can be exchanged by buses traveling in opposite directions. Shi et al. (2020) develop a MILP for a JHCO problem that comprises a single corridor and is traversed by multiple lines.

The following authors extend the scope of the basic JHCO problem to consider a transit network that comprises multiple lines. Pei et al. (2021) propose a nonlinear optimization model, which is then linearized and solved with a commercial solver. The experimental analysis shows that using modular buses both improves the quality of service for passengers compared to a standard fixed-capacity system, and reduces operational costs compared to a system where all passengers use their private cars. A multiple-lines JHCO problem is studied also in Gao et al. (2023) where operational-level issues, such as the driving range and the charging of modules, are explicitly captured in an optimization model as constraints. Despite the many potential advantages that can be achieved by using modular vehicles, as of today, it is hard to envisage a rapid and complete switch from conventional to modular buses. More likely, hybrid systems based on the simultaneous use of both types of vehicles will probably operate until

Table 1
A summary of the literature on modular buses using an optimization approach.

Reference	Class of problems	Setting	Objective function (Min.)	Solution method
Chen et al. (2019)	JHCO	Single-line corridor	Op & Pas costs	Dynamic programming
Chen et al. (2020)	JHCO	Single-line corridor	Op & Pas costs	Continuum approximation
Dai et al. (2020)	JHCO	Single-line corridor	Op & Pas costs	Dynamic programming
Chen and Li (2021)	JHCO	Single-line corridor	Op & Pas costs	Branch-and-Bound
Shi and Li (2021)	JHCO	Single-line corridor	Op & Pas costs	Dynamic programming
Tian et al. (2023)	JHCO	Single-line corridor	Op & Pas costs	Genetic algorithm
Xia et al. (2023)	JHCO	Single-line corridor	Op & Pas costs	L-Shaped method
Liu et al. (2023)	JHCO	2-ways corridor	Op & Pas costs	Decomposition-based
Shi et al. (2020)	JHCO	Shared corridors	Op & Pas costs	Commercial solver
Pei et al. (2021)	JHCO	Network-wise	Op & Pas costs	Commercial solver
Gao et al. (2023)	JHCO	Network-wise	Op costs	Column generation
Dakic et al. (2021)	JHCO	Network-wise	Op & Pas costs	Sequential quadratic programming
Gong et al. (2021)	TND	Network-wise	Op & Pas & uncovered demand costs	Particle swarm optimization
Cheng et al. (2024)	TND	Network-wise	Op & Pas costs	Heuristic
Liu et al. (2020)	FS	Single-line corridor	# of Modules	Deficit function

a complete transition is accomplished. Based on this observation, Dakic et al. (2021) examine a transit system with a hybrid fleet composed of autonomous modular buses and conventional fixed-capacity buses.

Some authors study the use of modular buses to mitigate the problem of bus bunching. The latter is a phenomenon afflicting traditional bus transit systems. Since travel times are very unpredictable, two (or more) successive buses may arrive at the same stop at times that are much closer than originally planned. As an outcome, the first bus serves most of (sometimes all) the passengers, whereas the following ones travel with low occupancy rates (or even empty). This generates several inconveniences to the passengers, including longer than planned waiting times. Note that, also in this case, the viewpoint is mainly on the operational-level. Furthermore, in contrast with our approach, the majority of the related papers do not apply an optimization method, but rather a simulation approach to evaluate given operating policies. Along this line, Khan et al. (2023) suggest mitigating bus bunching by exploiting the flexibility of modular buses. Instead of stopping the whole modular bus, one or more modules are detached from the latter while in motion, and rerouted to a stop. Passengers alighting at that stop leave the detached modules. Passengers boarding at that stop get on the detached modules once they arrive. After boarding, the modules can either wait for the next bus or depart immediately and try to catch up to the bus from which they were originally detached. Khan and Menéndez (2023) extend the previous research combining the previous strategy with a bus holding strategy. The latter is a simple strategy where the modules detached stay on hold at a bus stop as long as needed. The previous two strategies are also considered in Liu et al. (2024). After proposing an infinite-horizon stochastic optimization model, the latter authors develop a reinforcement learning algorithm to learn the optimal strategy for each bus. Finally, Lin et al. (2024) explore the role of boarding dynamics upon boarding times and bus bunching. Particularly, they compare a fixed-capacity boarding system against a modular and flexible boarding one.

As previously mentioned, one of the key features of the modules designed by NExT S.r.l. is that they can couple and decouple while in motion. The following authors study some related challenges by applying simulation approaches. Wu et al. (2021) investigate when coupling and decoupling operations should be carried out. Khan and Menéndez (2024) study how autonomous modular vehicles can be employed to design a seamless transit network, where the majority of modules operates without stopping, and where passengers can travel from their origin to their destination without any external transfers.

The papers reviewed above address operational-level problems. The study of tactical- and strategic-level problems is still relatively unexplored. Among the latter, Gong et al. (2021) and Cheng et al. (2024) address Transit Network Design (*TND*) problems for transit systems that employ modular buses. In the former paper, the authors devise a mixed-integer nonlinear program aiming at minimizing a weighted sum of three cost components: operational costs, passenger travel costs, and external costs related to the unserved passengers. Cheng et al. (2024) study a stylized design model for a transit system using autonomous modular buses on a grid network, in a range of demand density scenarios with both homogeneous and heterogeneous spatial distributions.

To the best of our knowledge, Liu et al. (2020) is the only article that tackles the problem of determining an optimal Fleet Size (*FS*) consisting of autonomous modular buses. In their study, the authors analyze the problem for a single-line transit system and capture the problem by adapting ideas from deficit function theory. Finally, the authors illustrate a case study of a single-line system in Singapore. Compared to Liu et al. (2020), in the present research we consider a more realistic transit system comprising multiple lines. This setting allows taking full advantage of the flexibility of modular buses. Indeed, it enables the transit system to adapt the capacity by sharing modules among lines according to the passengers demand, as well as rebalancing the capacity by moving empty modules between stops.

Contributions of the paper. Given a transit network with multiple lines and given a mobility demand, we can summarize the contributions of this paper as follows:

- ✓ we introduce a novel optimization problem, called the *Modular Buses Assignment* (MBA) problem, to support the strategic-level decision of employing a fleet of modular, rather than conventional, buses;

Table 2
A summary of the main notation.

Notation	Description	Notation	Description
L	Set of lines	o_k	Origin of request k
V	Set of all nodes	d_k	Destination of request k
S	Set of ordinary stops	P_k	Path followed by request k
J	Set of junctions	p_k	Average number of passengers associated with request k in the time unit
T	Set of terminals	$u_{ij\ell}$	Average number of modules departing from i along line arc (i,j,ℓ) in the time unit
A	Set of line arcs	v_{ij}	Average number of modules departing from i along rebalancing arc (i,j) in the time unit
R	Set of rebalancing arcs	$t_{ij\ell}$	Average travel time of line arc (i,j,ℓ)
K	Set of mobility requests	τ_{ij}	Average travel time of rebalancing arc (i,j)

- ✓ we present an ILP to determine an optimal assignment of modules to lines capable of capturing the flexibility of modular buses, that is sharing and rebalancing of modules;
- ✓ we illustrate the results of extensive computational experiments, showing that the total capacity deployed with modular buses is reduced, on average, by 49.44% compared to conventional buses, with a minimum reduction of 29.32% and a maximum of 70.77%. Additionally, we show that the average occupancy ratio increases from 41.22% to 72.85%.

3. Problem description

The MBA problem can be described as follows. We assume that a transport network with multiple bus lines is given, as well as the average mobility demand expressed with respect to a given unit of time (e.g., an hour or a few hours). The length of the time unit is chosen so that the demand can be assumed constant over time. We consider a decision-maker who has to determine the minimum number of modules that must be deployed to meet the mobility demand. To thoroughly describe the MBA problem, we present three components: (a) the structure of the transport network; (b) the nature of the mobility demand; and (c) how modules can be used to meet the mobility demand. Each of the three components is detailed in the following. The main notation introduced below is summarized in Table 2.

3.1. Transport network structure

The network structure is represented by a directed multi-graph denoted as $G = (V, A \cup R)$, where the node set V is the set of all *stops*, and the arc set is given by the union of two disjoint sets, A and R. We clarify their meaning below. We assume that passengers can board to and alight from buses at any stop in set V.

A *line* is defined as a closed path connecting a subset of nodes in V. Let L be a set of lines. Each line visits either one (in the case of a circularly shaped line) or two distinct stops called *terminals*, representing the ends of the line. Let $T \subset V$ denote the set of terminals. A nonempty subset $J \subset V$ of *junctions* is given. A junction is a stop where two or more lines intersect. Modules can change line only at a junction. Hereafter, this operation is called *module sharing*. Hence, J is the set of nodes where module sharing is allowed. Note that a node can be both a terminal and a junction, when it is an end of a given line and other lines intersect. Finally, let $S = V \setminus (T \cup J)$ be the set of *ordinary stops*. The role of ordinary stops, terminals, and junctions will be explained in detail later.

A *line arc* is denoted (i,j,ℓ), where i and j are nodes visited consecutively by line ℓ. Set A comprises all line arcs, i.e.:

$$A = \{(i,j,\ell) : i,j \in V,\ i \neq j,\ \ell \in L\}.$$

Two stops i and j may be connected by multiple arcs in A, i.e., multiple lines travel from stop i directly to stop j. Hence, G is a multi-graph.

We assume that modules can be relocated (empty) between junctions and/or terminals without traveling on line arcs, but following a more direct connection between the stops considered. To this end, let R be the set of all arcs, called hereafter *rebalancing* arcs, connecting two nodes in $T \cup J$ along which modules can be rebalanced. More formally:

$$R = \{(i,j) : i,j \in T \cup J,\ i \neq j\}.$$

In Fig. 2, we show a small example of a network G comprising a blue and a red line. For the sake of clarity, we depict in the left panel the arcs in set A, whereas the right panel shows the arcs in set R.

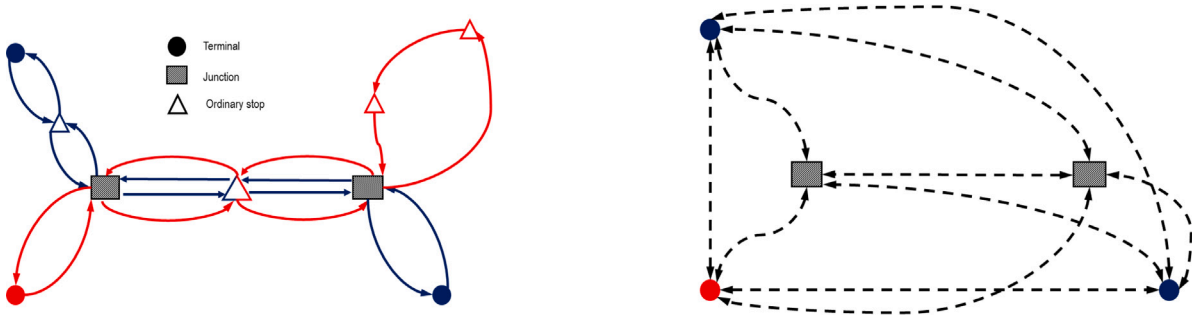


Fig. 2. An example of network G: the line arcs in set A (left panel), and the rebalancing arcs in set R (right panel). To simplify the picture on the right panel, every dashed link represents a pair of arcs with opposite direction.

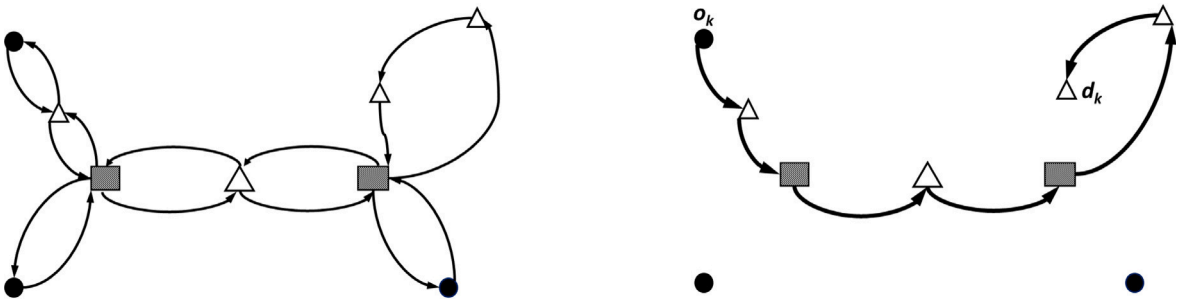


Fig. 3. The left panel shows graph $\bar{G}$ derived from the network in Fig. 2, whereas the right panel depicts an example of a path P_k on graph $\bar{G}$.

3.2. Mobility demand

The mobility demand is described as a set K of *mobility requests*, where each request $k \in K$ is associated with a set of customers and is defined as the tuple (o_k, d_k, P_k, p_k), where $o_k \in V$ denotes the origin of request k; $d_k \in V$ its destination; P_k a path from o_k to d_k; and p_k the rate of departures, that is, the average number of passengers, per time unit, requesting a transit from o_k to d_k. To formally define path P_k, it is convenient to introduce the following notation. Let $\bar{A}$ be defined as:

$$\bar{A} = \{(i,j) : (i,j,\ell) \in A \text{ for at least one } \ell \in L\},$$

and let $\bar{G} = (V, \bar{A})$ be the (simple) graph, induced by transport network G, comprising all arcs served by at least one line. Path P_k is the path on graph $\bar{G}$ that passengers associated with request k follow. Fig. 3 depicts graph $\bar{G}$ for the example in Fig. 2 and a possible path P_k. The determination of path P_k for each request is detailed below in Section 6.1.

Each node of the network (ordinary stop, terminal, or junction) can be simultaneously origin and destination of multiple requests. Further, different passengers may have common origin and destination, but follow different paths — i.e., passengers departing from the same origin may follow different paths to reach the same destination. In the mathematical model presented below, they can be treated separately as different mobility requests, say k and k' following paths P_k and $P_{k'}$ with $P_k \neq P_{k'}$.

We are assuming all passengers associated with a given mobility request k follow the same path P_k, which is defined on the simple graph $\bar{G}$. More precisely, passengers choose the path P_k to follow, but not the lines. The motivations of this assumption are as follows. On the one hand, we assume that passengers have no preference on the line to take. Indeed, the choice among multiple lines might occur only for those arcs from stop i to stop j that are traversed by more than one line. As previously mentioned, we assume that all such lines travel from i directly to j — i.e., they travel along the same road section with comparable travel times. Furthermore, to avoid unnecessary line changes, the latter are penalized in the objective function (see below for further details). As a consequence, from a passenger's perspective the choice of the line is irrelevant. On the other hand, we assume that the path followed by each mobility request is given, since our goal is to gauge the advantages of a transit system employing modular buses in a setting where passengers follow their favorite path, and the system must adapt to their behaviors. This is the typical situation that occurs when passengers are familiar with the structure of the transit network.

3.3. Modules

Lines are served by modular buses. Each bus is composed of one or more homogeneous modules, where every module has a fixed capacity Q expressed in terms of number of passengers. We define the *composition* of a modular bus as the number of modules it is composed of. A modular bus with a composition of two is shown in Fig. 1. Akin to the *station-wise* docking considered in Chen et al.

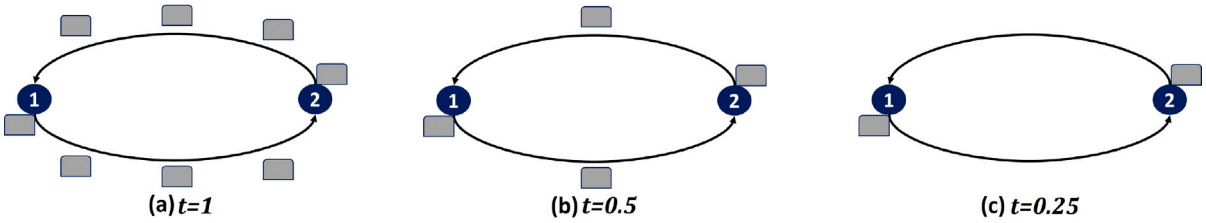


Fig. 4. Number of modules in the system with different travel times (1, 0.5, 0.25), but the same rate of modules.

(2020) and Shi and Li (2021), we assume that the composition of a bus can only change at junctions and terminals, but cannot at ordinary stops. The rationale for such an assumption is that junctions and terminals are often the main stops of a line, and provide enough room for docking and undocking operations, as well as room for modules to be stationary while waiting to be coupled to a modular bus.

The characterization of module flows is key in our approach and deserves a detailed analysis. We first introduce the concept of the rate of modules traveling along an arc, then we explore the sharing of modules among lines at junctions, and finally we discuss the rebalancing of modules among junctions and terminals.

Module rates

We describe the flow of modules by specifying, for each line arc, the module rate. More precisely, for each $(i, j, \ell) \in A$ let $u_{ij\ell}$ be the average number of modules departing in the time unit from i along arc (i, j, ℓ). Note that $u_{ij\ell}$ is also the average number of modules observed in a time unit by an observer positioned at any point along arc (i, j, ℓ). Similarly, for each rebalancing arc $(i, j) \in R$, we denote as v_{ij} the average number of modules departing in the time unit from i along arc (i, j).

A key observation is that the rates coupled with travel times determine the number of modules flowing in the system at any time. To clarify this point, consider the simple network comprising one line (denoted as b) composed of two line arcs, $(1, 2, b)$ and $(2, 1, b)$. Assume that the travel time along both arcs is the same, and it is denoted as t. Let $u_{12b} = u_{21b} = 4$, and let us assume that a module departs from each node every 0.25 time units. Consider now the following three cases. If $t = 1$, then every time a module departs from node 1, we can observe other three modules along arc $(1, 2, b)$, those departed 0.25, 0.50, and 0.75 time units before. At the same time, we expect a module to depart from node 2, and three other modules traveling along arc $(2, 1, b)$. Hence, at any moment, $1 \cdot u_{12b} + 1 \cdot u_{21b} = 8$ modules are traveling in the system. Fig. 4(a) depicts this first case. If $t = 0.5$, every time a module departs from node 1, we can observe only one module along arc $(1, 2, b)$, that departed 0.25 time units before. The module that departed 0.5 time units before from 1 has just reached node 2, and it is departing along arc $(2, 1, b)$, whereas the module that departed from node 1 0.75 time units before is already halfway on arc $(2, 1, b)$. Hence, at any moment, $0.5 \cdot u_{12b} + 0.5 \cdot u_{21b} = 4$ modules are traveling in the system. Fig. 4(b) shows this second case. Finally, if $t = 0.25$, every time a module departs from node 1, the module that departed 0.25 time units before is departing from node 2 toward node 1. The total number of modules traveling in the system is $0.25 \cdot u_{12b} + 0.25 \cdot u_{21b} = 2$. Fig. 4(c) depicts this third case. From a passengers' perspective, the rate of modules traveling along an arc represents a better indicator of the level of service provided than the number of modules assigned to that arc. In fact, the number of modules flowing is very different across the three networks depicted in Fig. 4. Nevertheless, the average number of modules observed in a time unit by a passenger positioned at any point along each arc, i.e., the rate, is exactly the same. The mathematical formulation presented in Section 4 uses rates of modules as main decision variables.

To generalize the above observations, let $t_{ij\ell}$ be the average travel time of line arc $(i, j, \ell) \in A$, and let τ_{ij} be the average travel time of rebalancing arc $(i, j) \in R$. Hence, for any line arc $(i, j, \ell) \in A$, if $u_{ij\ell}$ modules per time unit are departing from i along arc (i, j, ℓ), on average $t_{ij\ell} \cdot u_{ij\ell}$ modules are traveling along that arc. A similar argument applies to any rebalancing arc $(i, j) \in R$. Thus, the total number of modules in the system at any time is:

$$\sum_{(i,j,\ell)\in A} t_{ij\ell} \cdot u_{ij\ell} + \sum_{(i,j)\in R} \tau_{ij} \cdot v_{ij}. \tag{1}$$

Another key observation is that the rate of modules does not specify how modules are grouped. Consider again the example depicted in Fig. 4(a), with $t = 1$. In that case, the rate $u_{12b} = 4$ can be achieved using different grouping strategies: single modules departing every 0.25 time units (as in Fig. 4(a)), two modules coupled together departing every 0.50 time units, four modules attached together departing every time unit. Grouping strategies are crucial at an operational level to ensure an effective real-time management of modular buses, but are not relevant in the optimization problem considered in this paper that, as already mentioned, supports decisions at a strategic level. For the latter reason, we neglect grouping in our analysis.

Finally, notice that the observations above can be easily extended from single arcs to any subpath of a line. Consider, as an example, the red line (denoted as r) depicted in the right panel of Fig. 5. Consider, initially, the two arcs $(4, 5, r)$ and $(5, 2, r)$ separately, and assume that the travel time of each arc is $t = 0.5$. Let $u_{45r} = u_{52r} = 4$ and let a module depart from each node 4 and 5 every 0.25 time units. As $t = 0.5$, every time a module departs from node 4 (and, similarly, from node 5) one can observe only one module along arc $(4, 5, r)$ (arc $(5, 2, r)$), the one that departed 0.25 time units before. Hence, the number of modules traveling along arc $(4, 5, r)$ is $0.5 \cdot u_{45r} = 2$. The same value is obtained for arc $(5, 2, r)$. Consider, now, the subpath $(4, 5, 2)$ whose travel time is $0.5 + 0.5 = 1$. In this case, as the module rate of the two arcs is the same, we can conclude that if a module departs from node 4 every 0.25 time

Table 3
The mobility requests for the example in Fig. 5.

k	1	2	3	4	5	6	7	8
o_k	1	1	4	5	3	7	7	7
d_k	3	7	7	7	1	1	4	6
p_k	Q	Q	Q	Q	Q	Q	Q	Q

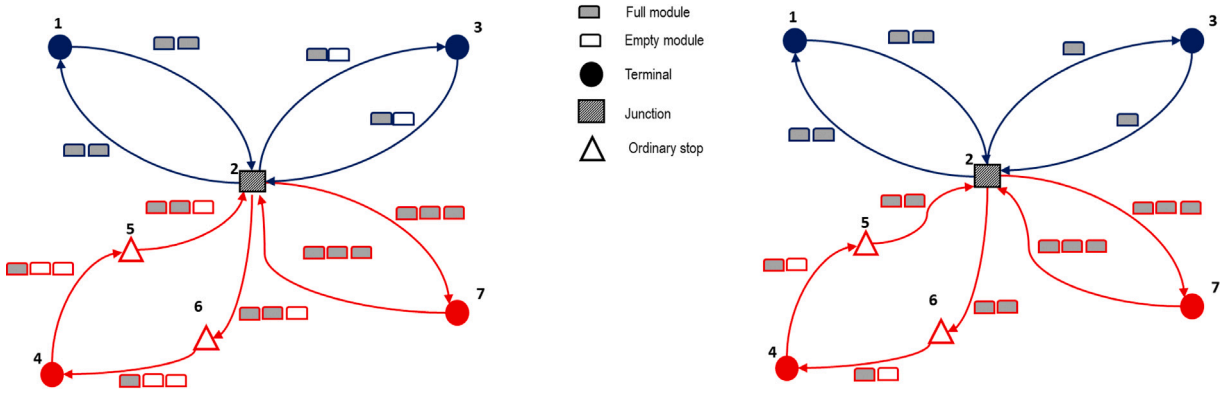


Fig. 5. Two feasible solutions without (left panel) and with (right panel) sharing of modules.

units, at its departure one can observe three other modules traveling along subpath $(4, 5, 2)$. Hence, the number of modules traveling along this subpath is $1 \cdot u_{45r} = 4$. We take advantage of the latter observation to reduce the number of integer variables in the ILP developed in Section 4.

Module sharing

We introduce the concept of *module sharing* with the following example. Consider the transport network comprising the two lines depicted in Fig. 5. The blue line (denoted as b) corresponds to the closed path $(1, 2, 3, 2, 1)$, where nodes 1 and 3 are terminals, whereas node 2 is a junction. The red line (denoted as r) corresponds to the closed path $(4, 5, 2, 7, 2, 6, 4)$, where nodes 4 and 7 are terminals, whereas nodes 5 and 6 are ordinary stops. Node 2 is the only junction, being the only node where the two lines intersect. All arcs have a unit travel time, i.e., $t_{ij\ell} = 1$ for all line arcs (i, j, ℓ). The latter assumption simplifies the following exposition, as for each arc the rate equals the number of modules. Consider the mobility requests reported in Table 3, and assume that each request k travels according to the shortest path from its origin to the corresponding destination (that is, path P_k). One can observe that: (i) the passenger rate is Q on arcs $(2, 3, b)$, $(3, 2, b)$, $(4, 5, r)$, and $(6, 4, r)$; (ii) the passenger rate is $2Q$ on arcs $(1, 2, b)$, $(2, 1, b)$, $(2, 6, r)$, and $(5, 2, r)$; (iii) the passenger rate is $3Q$ on arcs $(2, 7, r)$ and $(7, 2, r)$.

Consider the following two policies. In the first policy, we assume that the composition of a modular bus cannot be changed once allocated to a line. This implies that along each arc of the line, the rate of modules is the same, and, to satisfy the mobility demand, this rate matches the largest passenger rate across all arcs traversed by the line. In the example provided, a module rate of 2 is required on all arcs of line b, whereas a module rate of 3 is required on all arcs of line r. Since every arc has a unit travel time, a total of 26 modules are required in the system. However, this policy implies that in every time unit, one empty module travels along arcs $(2, 3, b)$, $(3, 2, b)$, $(2, 6, r)$, and $(5, 2, r)$, and two empty modules travel along arcs $(4, 5, r)$ and $(6, 4, r)$ (see the left panel in Fig. 5). As a result, at any time, 8 modules out of 26 are traveling empty in the system.

In the second policy, we assume that in junction 2 modules can change the line they are assigned to. In other words, modular bus compositions can change (only) at node 2. In any time unit, let two modules depart from 1 ($u_{12b} = 2$) and from 5 ($u_{52r} = 2$), one module departs from 3 ($u_{32b} = 1$), and three modules depart from 7 ($u_{72r} = 3$). Every two modules of line b arriving to node 2 from node 1, one can be assigned to line r, coupling it to the 2 modules arriving from 5. In this way, three modules depart from 2 headed to 7 ($u_{27r} = 3$). Similarly, every three modules of line r arriving to node 2 from 7, one can be assigned to line b, coupling it to the module arriving from 3. In this way, two modules depart from 2 headed to 1 ($u_{21b} = 2$). The rate of modules assigned to arcs $(4, 5, r)$ and $(6, 4, r)$ is forced to be 2 (see the right panel in Fig. 5), since no composition change is allowed at ordinary stops. This policy, based on module sharing among lines at junctions, enables the reduction of the number of modules in the system from 26 to 20. Additionally, the number of modules traveling empty is reduced from 8 to 2.

Module rebalancing

We introduce the concept of *module rebalancing* with the following example. Consider the network shown in Fig. 6 and having a single line b given by the closed path $(1, 2, 3, 4, 5, 6, 3, 7, 1)$. Let $t_{ijb} = 1$ for all line arcs (i, j, b) of the network. We consider two mobility requests: from node 1 to node 5 with passenger rate $2Q$, and from node 5 to node 1 with passenger rate Q.

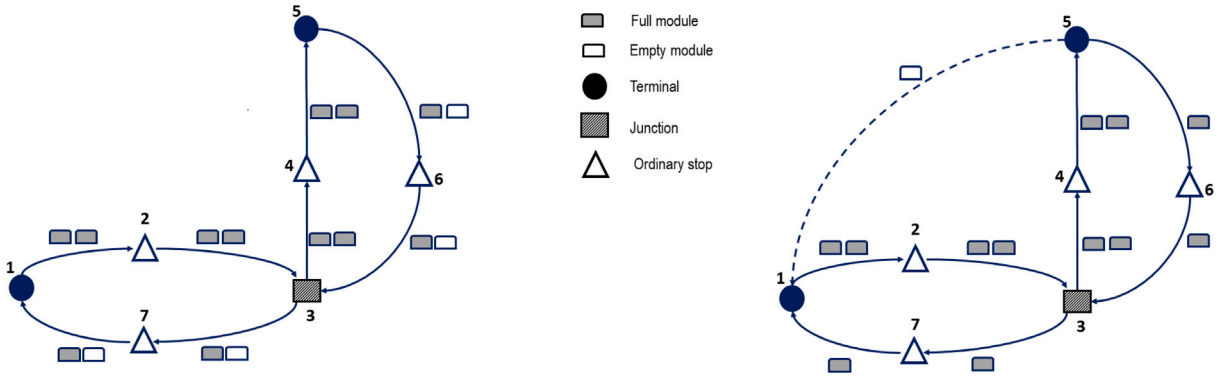


Fig. 6. Two feasible solutions without (left panel) and with (right panel) the rebalancing of modules.

As in this simple single-line network sharing modules with other lines is not possible, to meet the mobility demand a rate of two modules per time unit must be assigned to each arc. As a consequence, 16 modules are needed to operate the system, but half of the modules traveling along line arcs $(5,6,b)$, $(6,3,b)$, $(3,7,b)$, $(7,1,b)$ are empty (see the left panel in Fig. 6). This inefficiency is caused by the imbalanced distribution of the mobility demand.

To mitigate this inefficiency, for every two modules arriving at node 5, one empty module can be relocated directly to terminal 1 along the rebalancing arc $(5,1) \in R$ (see the right panel in Fig. 6). As with this movement the module capacity allocated is rebalanced, we call this operation *module rebalancing*. Note that, in the example, 4 is the total travel time on the subpath (5,6,3,7,1). As long as $\tau_{51} < 4$, rebalancing allows reducing the total number of modules. For example, if $\tau_{51} = 2$, two less modules can be used compared to the solution where rebalancing is not allowed. In the right panel of Fig. 6, it is assumed that $t_{51} = 1$, and, hence, 3 modules can be saved.

The simplified example above illustrates how module rebalancing operates on a single line. In the problem at hand, we consider a more general setting where module rebalancing can also be carried out among different lines.

4. The mathematical formulation

This section introduces the MBA model, an ILP to determine an optimal assignment of modules to lines, so that all mobility requests are met, employing a minimum total number of modules.

In the optimization model developed below, two strictly intertwined levels of decisions are optimized simultaneously. Recall that path P_k followed by the passengers of request $k \in K$ is given as an input of the problem. Further, recall that such a path is defined on the simple graph $\bar{G}$. In the first level of decisions, the model must map each path P_k into a path on the multi-graph G. In other words, for each pair of nodes, say i and j, that are connected by multiple arcs, the model must determine which line serves each request whose path P_k travels from i to j. This mapping must satisfy the following two restrictions. Splitting a given request k among different lines is not allowed, and requests can change line only at junctions. To avoid unnecessary transfers for requests, the number of line changes is penalized in the objective function. To this aim, let $\alpha \geq 0$ be a penalty factor for each line change. The first level of decisions determines the total number of passengers, per time unit, traveling along each line arc of multi-graph G, and thereby, implies a lower bound for the transport capacity allocated to such an arc. In the second level of decisions, the model must determine how to assign module rates to each arc of network G (including the rebalancing arcs in R) so that the above mentioned lower bounds are satisfied, as well as a set of flow conservation constraints. The goal is to employ a minimum number of modules in the system, while reducing the number of times requests must change lines.

In Section 3.3, we denoted as $u_{ij\ell}$ the rate of modules traversing line arc (i,j,ℓ), for all $(i,j,\ell) \in A$. In expression (1) we used these quantities to express the total number of modules in the system. A natural idea is to use $u_{ij\ell}$ as decision variables in an optimization model. However, since modular bus compositions can only change at junctions or terminals, using variables $u_{ij\ell}$ is inefficient, since a large number of them must be forced to take identical values (i.e., all those such that at least one between i and j is neither a junction nor a terminal). To illustrate this issue, consider the subpath $(4,5,2)$ of the red line in Fig. 5. Node 5 is an ordinary stop, i.e., $5 \in S$. Since module sharing is allowed only at junctions, and module rebalancing can be carried out only between a pair of nodes if they are terminals and/or junctions, neither sharing nor rebalancing can take place at 5, and hence we should impose in the model $u_{45r} = u_{52r}$. This observation is exploited to reduce the number of integer variables in the mathematical model, and can be generalized as follows.

By construction, any line $\ell \in L$ visits at least two nodes in $T \cup J$. Let j_0^ℓ be an arbitrarily chosen node in $T \cup J$ visited by line $\ell \in L$ (for example, j_0^ℓ may be a terminal of ℓ). Let $j_0^\ell, j_1^\ell, \ldots, j_{n_\ell}^\ell$ be all the nodes in $T \cup J$ encountered by traveling along line ℓ and starting from j_0^ℓ. Since lines are assumed to be closed paths, $j_{n_\ell}^\ell = j_0^\ell$. As an example, consider again the red line r depicted in Fig. 5. Assume that $j_0^r = 4$, then the nodes mentioned above are $4,2,7,2,4$. Consequently, we can partition a general line ℓ into a collection of sections (subpaths) $P_1^\ell, P_2^\ell, \ldots, P_{n_\ell}^\ell \subset A$, where P_h^ℓ is the section of line ℓ going from terminal or junction j_{h-1}^ℓ to terminal or

junction j_h^ℓ (with $h = 1, \dots, n_\ell$). Given the above example from Fig. 5, P_1^r is the subpath composed by the two line arcs $(4, 5, r)$ and $(5, 2, r)$. For convenience, let $N_\ell = \{1, \dots, n_\ell\}$ be the set of all indices denoting the sections composing line ℓ. Since it is assumed that sharing of modules can be done at junction stations only, and that rebalancing is between junction and terminal stations only, for any $h \in N_\ell$, all module rates $u_{ij\ell}$ with $(i, j, \ell) \in P_h^\ell$ must take the same value (e.g., in the example above $u_{45r} = u_{52r}$). Hence, we can replace all the latter rates with a single non-negative integer variable $w_{\ell h}$. Let $t'_{\ell h}$ be the average time needed to travel along section P_h^ℓ, i.e., the sum of the travel times of the arcs composing subpath P_h^ℓ. It is computed as follows:

$$t'_{\ell h} = \sum_{(i,j,\ell) \in P_h^\ell} t_{ij\ell} \qquad \text{for all } \ell \in L \text{ and } h \in N_\ell.$$

Recall that, for each $(i, j) \in R$, τ_{ij} denotes the average time needed by an empty module to travel along rebalancing arc (i, j). From expression (1), we derive the following expression for the total number of modules in the system:

$$\sum_{\ell \in L} \sum_{h \in N_\ell} t'_{\ell h} \cdot w_{\ell h} + \sum_{(i,j) \in R} \tau_{ij} \cdot v_{ij}. \tag{2}$$

The mathematical formulation developed below uses the following decision variables:

$w_{\ell h} \in \mathbb{Z}_0^+$, with $\ell \in L$ and $h \in N_\ell$, is a non-negative integer variable representing the rate of modules traveling along subpath P_h^ℓ;
$v_{ij} \in \mathbb{Z}_0^+$, with $(i, j) \in R$, is a non-negative integer variable representing the rate of modules relocated along rebalancing arc (i, j);
$x_{kij\ell} \in \{0, 1\}$, with $k \in K$ and $(i, j, \ell) \in A$, is a binary variable that takes value 1 if and only if mobility request k travels from node i to node j by using line ℓ;
$z_{kj} \in \{0, 1\}$, with $k \in K$ and $j \in J$, is a binary variable that takes value 1 if and only if mobility request k changes line at junction j.

To formulate the model, we need to introduce some additional notation. For each $\ell \in L$ and $k \in K$, and given path P_k associated with the latter request, let $B_{\ell k}$ be the set of triples of nodes (i, j, h) that are on path P_k, are visited consecutively by line ℓ, and are such that the inner node j is neither the origin nor the destination of mobility request k. More formally:

$$B_{\ell k} = \{(i, j, h) : (i, j, \ell), (j, h, \ell) \in A; (i, j), (j, h) \in P_k; o_k \neq j \neq d_k\}.$$

For each $j \in T \cup J$, let $\Delta^+(j)$ be the set of subpaths P_h^ℓ departing from j across all lines, and let $\Delta^-(j)$ be the set of subpaths P_h^ℓ across all lines arriving to j.

The MBA can be stated as the following ILP:

[MBA model]

$$\min \quad \sum_{\ell \in L} \sum_{h \in N_\ell} t'_{\ell h} \cdot w_{\ell h} + \sum_{(i,j) \in R} \tau_{ij} \cdot v_{ij} + \alpha \sum_{k \in K} \sum_{j \in J} p_k \cdot z_{kj} \tag{3}$$

$$\text{s.t.} \quad \sum_{\ell \in L} x_{kij\ell} = 1 \qquad k \in K, (i, j) \in P_k \tag{4}$$

$$x_{kij\ell} = x_{kjh\ell} \qquad k \in K, \ell \in L, (i, j, h) \in B_{\ell k} \mid j \in S \tag{5}$$

$$-z_{kj} \leq x_{kij\ell} - x_{kjh\ell} \leq z_{kj} \qquad k \in K, \ell \in L, (i, j, h) \in B_{\ell k} \mid j \in J \tag{6}$$

$$\sum_{k \in K} p_k \cdot x_{kij\ell} \leq Q \cdot w_{\ell h} \qquad \ell \in L, h \in N_\ell, (i, j, \ell) \in P_h^\ell \tag{7}$$

$$\sum_{P_h^\ell \in \Delta^-(j)} w_{\ell h} + \sum_{(i,j) \in R} v_{ij} = \sum_{P_h^\ell \in \Delta^+(j)} w_{\ell h} + \sum_{(j,i) \in R} v_{ji} \qquad j \in T \cup J \tag{8}$$

$$w_{\ell h} \in \mathbb{Z}_0^+ \qquad \ell \in L, h \in N_\ell \tag{9}$$

$$v_{ij} \in \mathbb{Z}_0^+ \qquad (i, j) \in R \tag{10}$$

$$x_{kij\ell} \in \{0, 1\} \qquad k \in K, (i, j, \ell) \in A \tag{11}$$

$$z_{kj} \in \{0, 1\} \qquad k \in K, j \in J. \tag{12}$$

Objective function (3) has three terms. According to (2), the first two terms represent the total number of modules flowing in the system. The third term penalizes line changes. The penalty is weighted by the average number p_k of passengers per time unit associated with request k. In a situation where the model must choose between two requests which one to move to another line, including p_k in the penalty term will promote the movement of the requests with the smallest number of passengers associated. In other terms, to provide a good service, the inconvenience of changing the line should involve the smallest possible number of

passengers. For each request $k \in K$ and for any arc $(i,j) \in \bar{A}$ that is part of path P_k, constraints (4) ensure that such a request is assigned to one line $\ell \in L$. Furthermore, constraints (4) along with binary restriction (11) ensure that any request $k \in K$ is not split among different lines. For each request $k \in K$ and each triple of nodes $(i,j,h) \in B_{\ell k}$, constraints (5) impose that the request does not change line at any ordinary stop $j \in S$. As a consequence, requests can change line only at junctions. For each request $k \in K$, constraints (6) guarantee that if the request changes line at junction $j \in J$ (i.e., $|x_{kij\ell} - x_{kjh\ell}| = 1$, where $|\cdot|$ is the absolute value of a quantity), variable z_{kj} takes value 1, and, thereby, the corresponding penalty is accounted for in the objective function. For each subpath P_h^ℓ of each line $\ell \in L$, constraints (7) impose that the number of modules assigned to such a subpath is sufficient to satisfy the total mobility demand assigned to every arc of the same subpath. For each terminal or junction $j \in T \cup J$, constraints (8) ensure the flow conservation: the rate of modules entering into node j must be equal to the rate of modules leaving the same node. Both rates entering and leaving a node consider the modules assigned to all lines plus the empty modules used for rebalancing the capacity. Finally, constraints (9)–(12) define the domain of the decision variables.

5. The value of sharing and rebalancing

In this section, we analyze the value of module sharing among lines and, separately, the value of empty module rebalancing by analyzing two extreme cases. The following results are based on idealized assumptions, leading to extreme situations. Such situations prove that there is no theoretical bound to the advantages that sharing and rebalancing might produce, but they are not representative of a concrete case arising from a real-world application.

Recall that sharing represents the possibility of exchanging modules among different lines at junctions. In Section 3.2, the mobility demand is defined as a set K of mobility requests, where each request $k \in K$ is identified by the tuple (o_k, d_k, P_k, p_k). For convenience, in this section we use the following simplified notation. Given a network $G = (V, A \cup R)$, we let $M = [m_{ij}]$ be a $|V| \times |V|$ origin–destination matrix (where $|\cdot|$ denotes the cardinality of a set), where each entry m_{ij} represents the total rate of passengers requesting a transport service from i to j in the time unit. Notice that $m_{ij} = \sum_{k:o_k=i,d_k=j} p_k$, where an empty summation takes value zero. In the following, the *value of sharing*, denoted as $\mathrm{VOS}(G,M)$, is the relative reduction in the number of modules required by the system when sharing is allowed compared to the situation where sharing is not permitted. We recall that rebalancing refers to the possibility of moving empty modules among terminals and junctions. The *value of rebalancing*, denoted as $\mathrm{VOR}(G,M)$, is the relative reduction in the number of modules required by the system when rebalancing is allowed with respect to the situation where rebalancing is not.

The higher the value of the VOS (resp., VOR), the higher the potential advantage of adopting module sharing (resp., module rebalancing). For example, if for a certain pair G and M we have $\mathrm{VOS}(G,M) = 0.25$, then sharing modules among lines reduces the number of required modules by 25%. By construction, for every G and compatible M, the values of the VOS and VOR are such that:

$$0 \le \mathrm{VOS}(G,M) < 1 \quad \text{and} \quad 0 \le \mathrm{VOR}(G,M) < 1.$$

In the following, we show that both VOS and VOR can be made arbitrarily close to 1.

Let us define:

$$\overline{\mathrm{VOS}} = \sup_{G,M}\{\mathrm{VOS}(G,M)\},$$

$$\overline{\mathrm{VOR}} = \sup_{G,M}\{\mathrm{VOR}(G,M)\}.$$

The following results hold.

Theorem 1. $\overline{VOS} = 1$.

Theorem 2. $\overline{VOR} = 1$.

The proofs of Theorems 1 and 2 can be found in Appendix. They are based on the analysis of two simple and idealized networks.

We highlight that it is complicated to empirically measure the values of the VOS and VOR when module sharing and rebalancing are simultaneously allowed. In fact, it is hard to discern the impact on the reduction in the number of modules of the sharing and, separately, of the rebalancing. Nevertheless, it is simple to measure the value of the VOS only by hindering module rebalancing. This can be achieved by removing all rebalancing arcs from an instance of the MBA model (or, equivalently, set to zero all v_{ij} variables). Conversely, measuring the value of the VOR only remains complicated, but can be gauged indirectly by comparing the value of the VOS, computed as just explained, and the value of a statistic measuring the combined effect of both module sharing and rebalancing. More details on the values taken by the VOS and VOR in our experimental analysis are provided in the following section.

6. Computational experiments

This section is devoted to presentation and discussion of the computational experiments. They were conducted on a Workstation HP Intel(R)-Xeon(R) at 3.5 GHz with 64 GB RAM (Win 10 Pro, 64 bits). The processor is equipped with 6 physical cores, and all threads were used while solving each instance. The MBA model was implemented in Julia, compiled within Visual Studio Code, and solved by means of CPLEX 22.1.1. All CPLEX parameters were set at their default values.

The section is organized as follows. Section 6.1 describes the generation of the instances we used in our experiments. Section 6.2 describes the transit systems we considered to assess the flexible use of modular buses, and presents the statistics computed to evaluate the performance of each system. Finally, in Section 6.3 detailed computational results are provided.

6.1. Instance generation

From a public repository, widely used in the literature on transportation research,[4] we selected a set of networks with the following properties: the structure of the network allows the generation of multiple and heterogeneous instances; each network has a number of nodes smaller than 1000; the coordinates of the nodes are provided; and, for each node, there are at least two arcs leaving and, separately, two arcs entering. The following five networks satisfy these properties: (*a*) Berlin-Friedrichshain, hereafter denoted shortly as `FRI`; (*b*) Berlin-Mitte-Center (`BMC`); (*c*) Berlin-Mitte-Prenzlauerberg-Friedrichshain-Center (`MPF`); (*d*) Berlin-Tiergarten (`TIE`); and (*e*) Chicago-Sketch (`CHI`).

Each of the above networks specifies the underlying street network, but does not contain any data regarding the transit system available. We generated the latter as follows. For each of the five networks above, we generated three heterogeneous transit networks, denoted as `A`, `B`, and `C`. Each transit network is generated so as to satisfy the following features:

- ✓ the transit network has a minimum of 6 and a maximum of 8 lines (set L). Lines are generated so that the resulting transit network covers most areas of the original street network, while avoiding an excessive overlapping of lines;
- ✓ terminals (set T) are located in the outer sections of the street network;
- ✓ all nodes that are traversed by more than one line are junctions (set J);
- ✓ each remaining node, which is neither a terminal nor a junction, is an ordinary stop (set S);
- ✓ for each line (resp., rebalancing) arc of the network, the corresponding average travel time $t_{ij\ell}$ (resp., τ_{ij}) is proportional to the Euclidean distance between the two associated nodes.

The mobility demand is generated according to the following two assumptions. Firstly, all passengers from a given origin, say o_k, to a given destination, say d_k, follow the same path, say P_k. In other words, there is no pair of mobility requests, say k and k', such that $o_k = o_{k'}, d_k = d_{k'}$, and $P_k \neq P_{k'}$. Secondly, each request k follows the path P_k with the minimum travel time from o_k to d_k. As described in Section 3.2, this path is computed on graph $\bar{G}$. Based on such assumptions, the set of requests K is generated through the following two steps. Given a network, in a first step we generated set K_{all} that contains all possible pairs of nodes $i, j \in V$ such that $i < j$. Subsequently, a given percentage θ of the pairs of nodes in K_{all} is randomly selected. We generated the mobility requests according to five values of θ: 10%, 20%, 50%, 80%, and 100%. For each pair of nodes i, j randomly selected, we generated two mobility requests: (*a*) request k where $o_k = i$ and $d_k = j$, and (*b*) request k' where $o_{k'} = j$ and $d_{k'} = i$. The average number of passengers associated is the same, i.e., $p_k = p_{k'}$, where p_k is an integer random number drawn from the interval [1, 50]. As a result, the mobility demand is symmetrical for each pair of nodes selected. As previously mentioned, for each request k path P_k is determined by solving a shortest (i.e., a minimum travel time) path problem from o_k to d_k in $\bar{G}$. Finally, each instance is solved by setting $\alpha = 0.001$. We chose this value since our primary objective is the minimization of the number of modules in the system, i.e., the first two terms in objective function (3). The minimization of the total penalty for line changes can be seen as a secondary objective. Hence, the choice of a small value of α. It is worth highlighting that in preliminary experiments we tested several other values of α, and the differences compared to the results reported below were negligible.

Summarizing, for each of the 5 original street networks (i.e., `FRI`, `BMC`, `MPF`, `TIE`, and `CHI`), we generated 3 transit networks (i.e., `A`, `B`, and `C`). For each of the resulting combinations (e.g., `FRI` combined with `A`), we considered 5 different settings for the mobility demand (i.e., θ equal to 10%, 20%, 50%, 80%, and 100%). By considering all possible combinations, we generated 75 instances. These data sets are publicly available at http://or-brescia.unibs.it/instances. Each instance is labeled according to the corresponding combination. As an example, the instance generated from the network called Berlin-Tiegarten, which considers the second transit network with $\theta = 80\%$ is labeled as `TIE.B.80`.

A summary of the instances in our test bed is reported in Table 4. For each transit network (indicated in column *Net*), column $|L|$ shows the number of lines. Column $|V|$ reports the total number of nodes (i.e., bus stops) composing the network, whereas column $|J|$ shows the number of junctions. The following column ($\frac{|J|}{|V|}$%) indicates the percentage of junctions compared to the total number of bus stops in the transport network. Finally, column $2|K_{all}|$ shows the total number of mobility requests when $\theta = 100\%$.

As illustrative examples, Fig. 7 displays the transport networks denoted as `FRI.C` (left panel) and as `BMC.B` (right panel). To simplify the readability, for each line (characterized by different colors), we only show the path from one terminal to the other. The path on the way back is very similar. The dots represent bus stops. Terminals are the stops at the ends of each line, whereas stops traversed by two or more lines are junctions.

6.2. Transit systems and performance evaluation

To assess the advantages achievable by the use of modular buses in a flexible setting, we analyzed the performance of the following three transit systems:

`m-flex:` this is the transit system described in this paper. It uses modular buses in a flexible setting, i.e., where module sharing and rebalancing are both allowed;
`m-rigid:` this system uses modular buses in a rigid setting, where module sharing and rebalancing are both neglected;
`b-rigid:` this transit system uses conventional buses with a fixed capacity Q.

[4] *Transportation Networks for Research.* https://github.com/bstabler/TransportationNetworks (Last accessed: 26th November 2024)

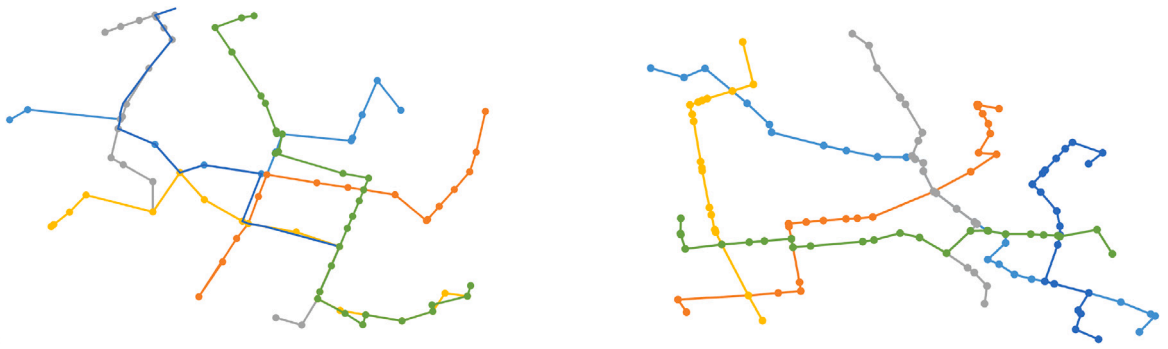


Fig. 7. Two examples of the transit networks generated: the FRI.C (left panel) and the BMC.B (right panel) networks.

Table 4
A summary of the structure of the instances tested.

Net	$\|L\|$	$\|V\|$	$\|J\|$	$\frac{\|J\|}{\|V\|}$%	$2\|K_{all}\|$	Net	$\|L\|$	$\|V\|$	$\|J\|$	$\frac{\|J\|}{\|V\|}$%	$2\|K_{all}\|$
FRI.A	6	114	19	16.67%	12,882	MPF.C	7	210	15	7.14%	43,890
FRI.B	6	78	15	19.23%	6,006	TIE.A	6	108	15	13.89%	11,556
FRI.C	6	89	20	22.47%	7,832	TIE.B	6	112	10	8.93%	12,432
BMC.A	6	113	18	15.93%	12,656	TIE.C	8	143	29	20.28%	20,306
BMC.B	6	125	16	12.80%	15,500	CHI.A	6	134	7	5.22%	17,822
BMC.C	7	149	16	10.74%	22,052	CHI.B	6	147	9	6.12%	21,462
MPF.A	6	235	13	5.53%	54,990	CHI.C	6	158	9	5.70%	24,806
MPF.B	8	251	18	7.17%	62,750						

The performance of the m-flex system is evaluated by solving the 75 instances generated with the MBA model. Each instance is solved considering each of the following values for the capacity Q: 10, 20, and 40. The first two values were chosen based on the current capacity of commercial modular vehicles, as those mentioned in the introduction, whereas the last value was taken from the literature (e.g., see Xia et al., 2023). Hence, altogether 225 optimizations were carried out for the m-flex system.

To measure the performance of the b-rigid and m-rigid systems, we proceeded as follows. Both systems are based on a fixed-capacity setting, where for each line the rate of vehicles (conventional or modular buses) assigned is constant along the whole path. To determine a minimum number of vehicles used in each of the two rigid systems, we solved the ILP detailed below, which uses the following additional notation. Let $w_\ell \in \mathbb{Z}_0^+$ be a non-negative integer variable representing the rate of vehicles traveling along line $\ell \in L$. The average travel time for traversing the whole closed path defining line ℓ is denoted as $\bar{t}_\ell$, which is computed as $\bar{t}_\ell = \sum_{(i,j,\ell)\in A} t_{ij\ell}$. The remaining notation has the same meaning as defined above.

[Rigid model]

$$\min \quad \sum_{\ell\in L} \bar{t}_\ell \cdot w_\ell + \alpha \sum_{k\in K}\sum_{j\in J} p_k \cdot z_{kj} \tag{13}$$

s.t. (4)–(6), (11), (12)

$$\sum_{k\in K} p_k \cdot x_{kij\ell} \le Q \cdot w_\ell \qquad \ell \in L, (i,j,\ell) \in A \tag{14}$$

$$w_\ell \in \mathbb{Z}_0^+ \qquad \ell \in L. \tag{15}$$

The first term in objective function (13) minimizes the total number of vehicles (conventional buses or modules) in the systems, whereas the second term penalizes line changes (as in the MBA model). Given line $\ell \in L$, for each of its line arcs $(i,j,\ell) \in A$, constraints (14) ensure that the number of vehicles assigned to such a line is sufficient to satisfy the total mobility demand traversing that arc. Recall that, for each pair of nodes i and j connected by multiple lines, the total mobility demand assigned to each line ℓ traveling from i to j is not known upfront. This observation motivates the need of an optimization model, and, thereby, of the ILP above, where the total mobility demand assigned to each line arc (i,j,ℓ) is a result of the optimization.

For the b-rigid and m-rigid systems we solved each of the 75 instances with the Rigid model. For the b-rigid system, each instance is solved considering each of the following three values of Q: 50, 80, and 100. For the m-rigid system, we tested the following values of Q: 10, 20, and 40. Hence, to evaluate each of the two rigid systems, 225 optimizations were carried out.

To gain insights into the benefits of employing modular buses, we devised the two following comparisons:

(A) b-rigid vs m-rigid: We compare the performance of the two rigid fixed-capacity systems. This comparison aims at measuring the benefits of replacing a fleet of conventional buses with a fleet of modular buses that do not exploit the flexibility enabled by module sharing and rebalancing;

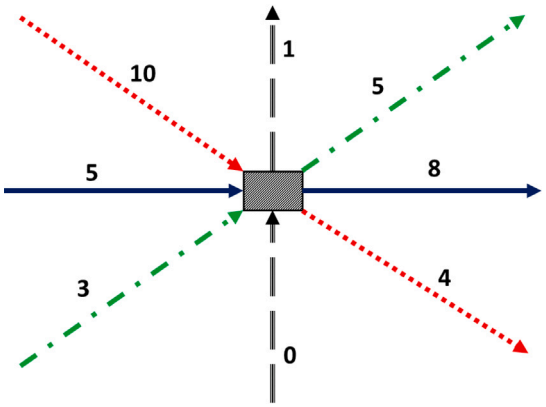


Fig. 8. An illustrative example of how rate Ψ_j is computed.

(B) `b-rigid` vs `m-flex`: We compare the conventional fixed-capacity buses with the transit system analyzed in the present paper, which employs modular buses in a flexible setting. This comparison aims at measuring the benefits of replacing a fleet of conventional buses with a fleet of modular buses where module sharing and rebalancing are both allowed.

For each instance and each of the three transit systems described above, we computed the following statistics:

`TCAP`: Total capacity deployed, expressed in terms of the total number of passengers that can be transported in a given system. For the `m-flex` system, it is computed as $\texttt{TCAP}_{\texttt{f}} = Q(\sum_{\ell\in L}\sum_{h\in N_\ell} t'_{\ell h}\cdot w_{\ell h} + \sum_{(i,j)\in R}\tau_{ij}\cdot v_{ij})$. For the rigid systems, it is computed as $\texttt{TCAP}_{\texttt{r}} = Q(\sum_{\ell\in L}\bar{l}_\ell \cdot w_\ell)$, where `r` is equal to `rb` for the `b-rigid` system, and it is equal to `rm` for the `m-rigid` system.

`VOF`: Value of flexibility. This statistic measures the total capacity that a decision-maker can save by adopting the `m-flex` system, while meeting the same mobility demand. It is computed as $\texttt{VOF} = \left(\texttt{TCAP}_{\texttt{r}} - \texttt{TCAP}_{\texttt{f}}\right)/\texttt{TCAP}_{\texttt{r}}$, for all possible combinations of capacities for the two systems considered. For instance, for each value of the capacity for the `b-rigid` system, say $Q = 100$, the `VOF` is computed for each of the three values of Q considered for the `m-flex` system.

Ω: Average occupancy ratio. This statistic gives a measure of how efficiently the capacity is exploited in each system. Small values of Ω indicate that vehicles, on average, travel with few passengers. On the other hand, very high values indicate situations where vehicles travel at almost full capacity, likely providing a lack of comfort for the passengers. Recall that different lines may have different total travel times. Similarly, different line arcs traversed by the same line have different travel times. These issues must be accounted for while computing Ω. Hence, Ω is computed as:

$$\Omega = \frac{\sum_{\ell\in L}\sum_{(i,j,\ell)\in A} t_{ij\ell}\sum_{k\in K} p_k x_{kij\ell}}{\texttt{TCAP}},$$

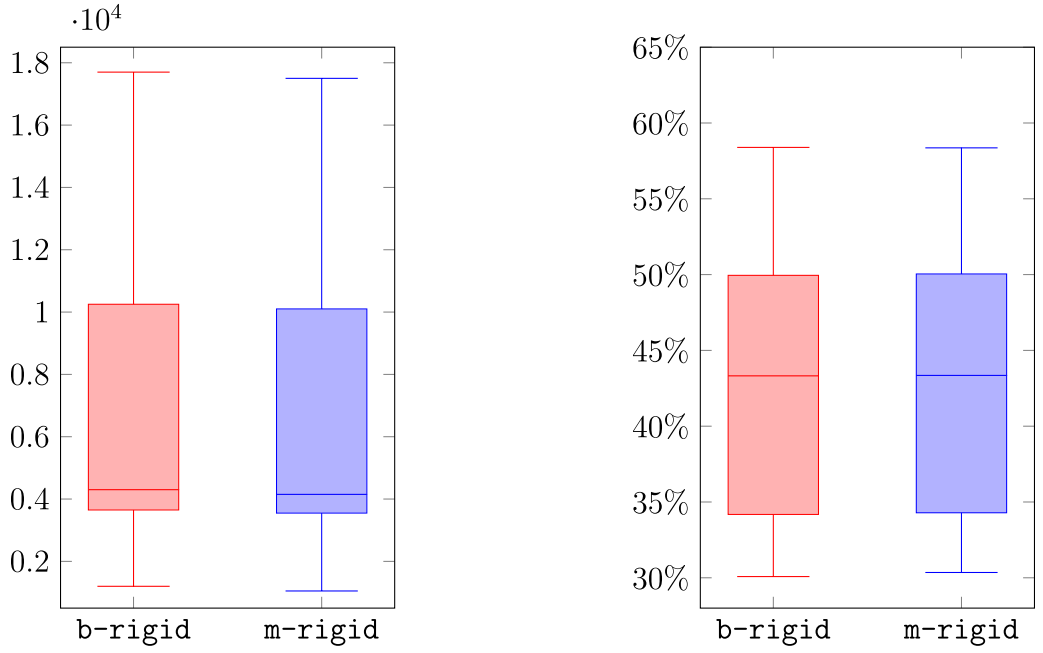
where `TCAP` is computed as defined above. Therefore, we refer to $\Omega_{\texttt{f}}$ when the denominator is $\texttt{TCAP}_{\texttt{f}}$, to $\Omega_{\texttt{rb}}$ when we divide by $\texttt{TCAP}_{\texttt{rb}}$, and to $\Omega_{\texttt{rm}}$ when the denominator is $\texttt{TCAP}_{\texttt{rm}}$.

Ψ: Rate of modules shared or rebalanced. This statistic is computed only for the `m-flex` system, and provides a measure of the rate of modules shared among lines or rebalanced at junctions and terminals. To compute Ψ we proceeded as follows. For each node $j \in T\cup J$ and each line traversing it, we computed the absolute value of the difference between the rate of modules entering j and the rate of modules leaving it. A similar absolute value is computed for the total rate of modules entering and leaving j along rebalancing arcs. The sum of all the above absolute values divided by 2 gives the rate Ψ_j of modules that are shared or rebalanced at node j. The computation of Ψ_j is exemplified in Fig. 8. This junction is traversed by three lines (red, blue, and green). For the sake of clarity, all rebalancing arcs entering (or leaving) are depicted as a single entering (leaving) gray arc. The absolute values mentioned above are equal to 6 modules for the red line, 3 for the blue line, 2 for the green line, and 1 for the rebalancing arcs. The sum of such absolute values is 12. From Fig. 8 one can observe that the red line shares 6 modules, which are distributed among the remaining lines and the rebalancing arcs. Hence, the actual rate of modules shared or rebalanced is $\Psi_j = 12/2 = 6$. The sum of Ψ_j over all junctions or terminals is then divided by the total rate of modules in the system to obtain statistic Ψ:

$$\Psi = \frac{\sum_{j\in T\cup J}\Psi_j}{\sum_{\ell\in L}\sum_{h\in N_\ell} w_{\ell h} + \sum_{(i,j)\in R} v_{ij}}.$$

6.3. Computational results

This section is devoted to the computational results. Firstly, we compare the two rigid fixed-capacity systems, i.e., the `b-rigid` against the `m-rigid` system. Subsequently, we compare the `b-rigid` system against the flexible setting of a transit system



(a) Total capacity (TCAP) (b) Average occupancy ratio (Ω)

Fig. 9. b-rigid vs. m-rigid: Box-and-whisker plots showing the distributions of statistic TCAP (outliers are not shown) and statistic Ω.

employing modular buses — i.e., the m-flex system. It is worth highlighting that computing times for solving the MBA model are negligible (on average, roughly 10 s, and always less than 1 min) and, therefore, are not reported below.

6.3.1. Comparison (A): b-rigid vs m-rigid

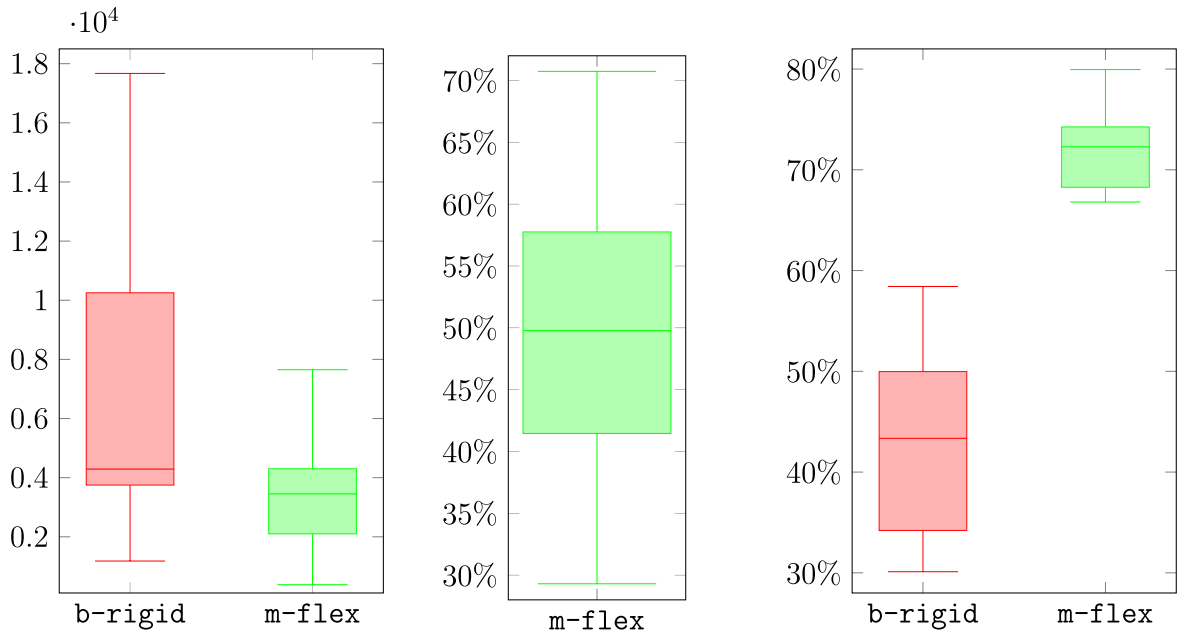
To gauge the potential advantages of adopting an m-rigid system compared to a more conventional b-rigid system, we use statistics TCAP and Ω. The distribution of the values of statistic TCAP is depicted in the left panel of Fig. 9.

Summarizing the insights above in terms of the total capacity deployed, the advantages achievable by adopting an m-rigid system are rather limited, with a small reduction in the average value of TCAP (on average, the total capacity reduces by 50 units). The limited benefits are confirmed by the very similar values of statistic Ω, reported in the right panel of Fig. 9. On average, the occupancy ratio Ω in the m-rigid system increases by 0.3% compared to the b-rigid system. The limited impact of using modules as in the m-rigid system is due to the fact that the number of modules on a line ℓ remains constant across all arcs. Though the number of modules may be different across different lines, such number must be sufficient to meet the largest mobility demand assigned to any of its line arcs (i, j, ℓ). This is true also when the mobility demand is concentrated in a few arcs, whereas a relatively small demand is assigned to the remaining arcs of the same line. The same logic applies to the b-rigid system.

6.3.2. Comparison (B): m-rigid vs. m-flex

We now compare the performance of the m-rigid against the m-flex system. The left panel of Fig. 10 depicts the distribution of the values of statistic TCAP for both systems. These values must be read in combination with those of statistic VOF, whose distribution (in percentage) is shown in the central panel of the same figure. We can see that the total capacity deployed to meet a given mobility demand can be substantially reduced adopting an m-flex system with respect to the b-rigid system. For the b-rigid system, statistic TCAP ranges from a minimum value of 1180 to a maximum non-outlier value (upper whisker) of 17,670. The range of variability of the same statistic for the m-flex system is considerably smaller, ranging from 380 to a maximum non-outlier value of 7650. The total capacity that can be saved by adopting the m-flex system, expressed by statistic VOF, is remarkable: on average, the reduction is equal to 49.44%, whereas the minimum and maximum reductions are equal to 29.32% and 70.77%, respectively.

The saving that can be achieved mostly depends on the structure of the transit network. One relevant factor is the number of junctions in the system: the more the junctions, the more the possibilities for module sharing and rebalancing, the more efficient the usage of modules. In addition, the number of lines traversing a given junction, as well as the number of times that two (or more) lines traverse the same set of junctions, impact the values of both statistics TCAP and VOF as well. By construction, the mobility demand is perfectly symmetric for each pair of nodes selected (see Section 6.1). As a result, the number of modules that are rebalanced is equal to zero in all the instances tested, with some marginal exceptions. Hence, the savings obtained are the results



(a) Total capacity (TCAP) (b) Value of flexibility (VOF) (c) Average occupancy (Ω)

Fig. 10. m-rigid vs. m-flex: Box-and-whisker plots showing the distributions of statistics TCAP, VOF, and Ω.

of module sharing only. In other words, for these instances, the value of the VOR is close to zero and the values of the VOS and VOF are almost equal. We investigated further this issue, and the results are presented at the end of the present section.

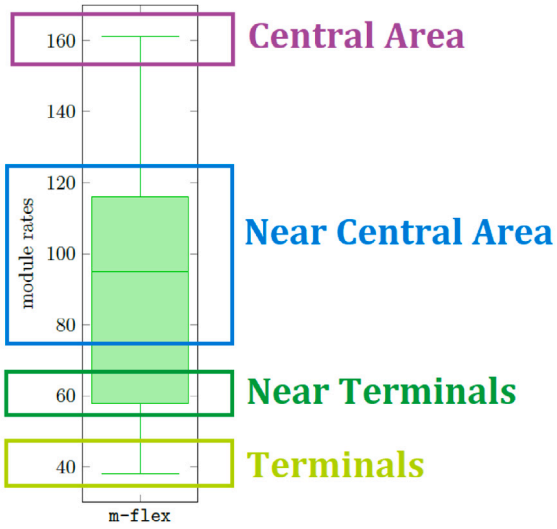
The analysis of the average occupancy ratio Ω confirms that the capacity deployed with modular buses is used much more effectively than if conventional buses are used. The distributions of the values of statistic Ω for both the b-rigid and m-flex systems are shown in the right panel of Fig. 10. The main insights we can gain are:

- ✓ the average occupancy ratio Ω increases remarkably for the m-flex in contrast to the b-rigid system: the average value of Ω_{f} is equal to 72.85%, against an average value equal to 41.22% for Ω_{rb};
- ✓ the range of variability of statistic Ω_{f} is considerably smaller than the one observed for Ω_{rb}. The former ranges from 66.82% to 79.97%, while the latter ranges from 30.07% to 58.91%.

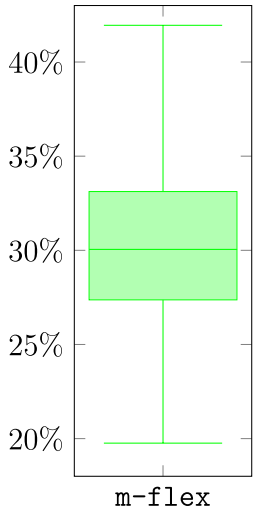
These results indicate that a very efficient usage of the capacity can only be achieved by fully exploiting the flexibility enabled by modular vehicles.

It is interesting to analyze how the rate of modules changes along a line. Fig. 11 shows the distribution of the rate of modules assigned to a specific line in an optimal solution to instance MPF.B.10. For each value of the rate of modules (a number on the left of the box-and-whisker plot), the figure displays the corresponding area of the original street network. As expected, the largest values are observed near and in the central areas, where the mobility demand is high. Conversely, the rate of modules decreases toward and in the outer sections of the network, where terminals are located. The rate of modules assigned to the central section of the line is equal to 161. In the section near the central area, the rate ranges between 92 and 116. It becomes 58 in the section approaching the terminals, while it is equal to 38 in the sections entering and leaving a terminal. To gain insights on the impact that using modular buses in a flexible setting may have on the operational complexity of the transit system, we analyzed the rate of modules that are shared among lines or rebalanced at junctions and terminals, as measured by statistic Ψ. The distribution of the values taken by statistic Ψ is depicted in the right panel of Fig. 11. This figure indicates that, on average, the rate of modules shared or rebalanced at junctions and terminals is approximately equal to 30%. The values of Ψ range from 19.76% to 41.95%. These values indicate that to achieve the flexibility enabled by the m-flex system, the operations involve, on average, approximately 1 out of 3 modules. Interestingly, the experimental results highlight that there is not a strong correlation between the values of statistics Ψ and VOF. While one could presume that a large reduction in the total capacity deployed (i.e., VOF) can only be achieved at the cost of an equally large number of operations of sharing and rebalancing (i.e., Ψ), we observed that there are several instances characterized by similar values of VOF (e.g., around 53%) with significantly different values of Ψ.

To complete the presentation of the computational results, aggregated statistics are reported in Table 5. Each line of the table, excluding the headers and the last line showing averages over all instances, shows averages computed over 15 instances. The average reduction of the total capacity deployed (i.e., statistic VOF) across all the instances is almost 50%. The effectiveness of the system



(a) Rate of modules along a single line.



(b) Rate of modules shared or rebalanced (Ψ).

Fig. 11. `m-flex`: Box-and-whisker plots showing the distributions of the rate of modules along a single line and statistic Ψ.

Table 5
`m-rigid` vs. `m-flex`: Average statistics values.

Net	VOF (%)	Ω_r (%)	Ω_f (%)	Ψ (%)
FRI.A	61.10%	31.22%	74.82%	42.27%
FRI.B	43.41%	49.78%	73.91%	27.39%
FRI.C	60.94%	32.99%	78.55%	27.28%
BMC.A	45.72%	43.31%	73.89%	23.73%
BMC.B	56.76%	34.50%	72.69%	27.74%
BMC.C	47.05%	45.19%	74.33%	28.12%
MPF.A	55.24%	33.37%	68.28%	29.27%
MPF.B	53.97%	34.79%	68.84%	35.91%
MPF.C	56.22%	33.69%	68.38%	29.49%
TIE.A	59.11%	33.93%	70.66%	39.47%
TIE.B	37.54%	45.57%	68.79%	32.31%
TIE.C	51.84%	37.86%	77.69%	31.23%
CHI.A	34.18%	53.39%	73.61%	25.88%
CHI.B	36.80%	57.57%	74.56%	19.96%
CHI.C	41.60%	51.27%	73.73%	26.34%
Average	49.43%	41.23%	72.85%	29.76%

with modular buses is also noticeable when looking at the difference between the average occupancy ratios Ω_r and Ω_f. In particular, the values of the latter statistic are much larger than the former, more than twice as large in many instances. In general, providing high levels of service requires that a non negligible number of modules are shared at junctions or are rebalanced (see statistic Ψ), with almost one third of the modules, on average, involved in sharing or rebalancing operations.

To gain further insights on the empirical values of the VOS and VOR when the mobility demand is not symmetric, we carried out the following analysis. We selected the transit network denoted as `TIE.B` and generated asymmetric mobility requests: for each pair of nodes i, j randomly selected, we generated only one mobility request k such that $o_k = i$ and $d_k = j$. The mobility requests were generated according to the five values θ presented in Section 6.1. We analyzed the performance of the `m-rigid` and `m-flex` transit systems, and of a third system (denoted as `m-flex(NR)`) where module rebalancing is forbidden. For each of the three systems, the module capacity Q is equal to 10.

The left panel of Fig. 12 shows the number of modules deployed in each of the three transit systems, and for each value of θ. The main insights we can gain are as follows. Enabling the possibility of module sharing only allows a remarkable reduction in the number of modules employed (compare the `m-rigid` against the `m-flex(NR)` system). An additional reduction, despite having a smaller magnitude, can be achieved enabling also module rebalancing (compare the two former systems against the `m-flex` one). As expected, the larger the percentage of mobility requests θ, the larger the impact of module sharing only, as well as the combined

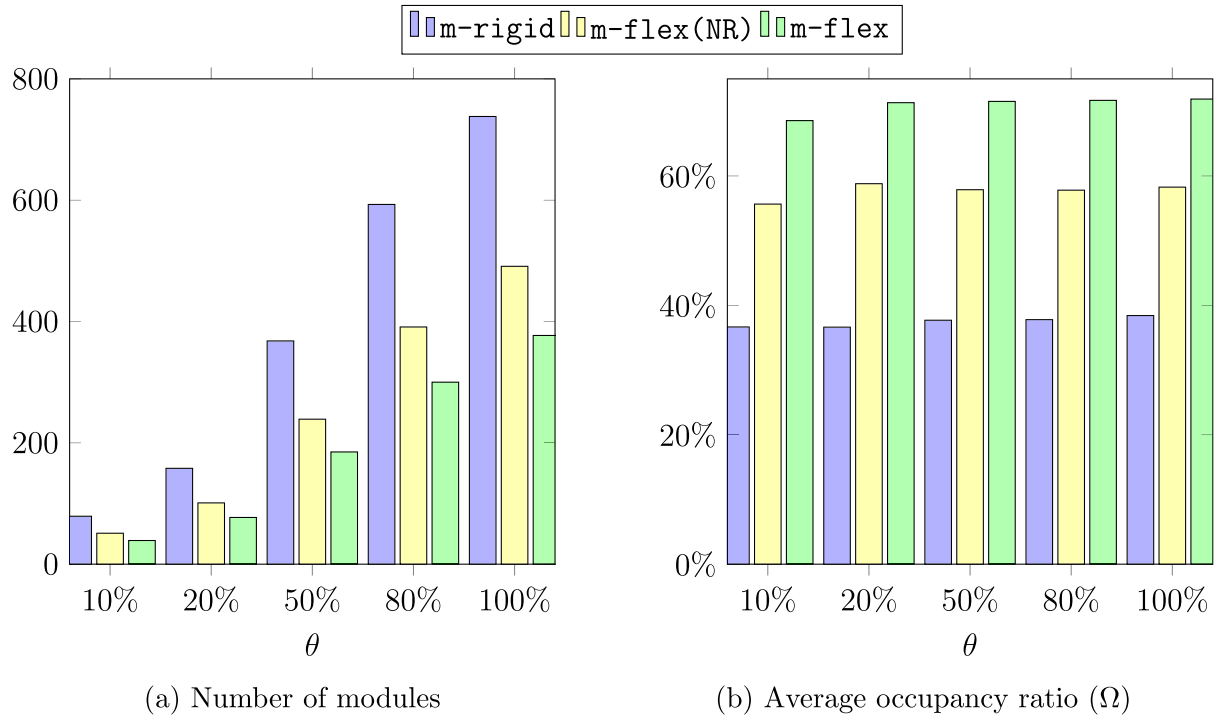


Fig. 12. The impact of allowing module sharing, and both module sharing and rebalancing.

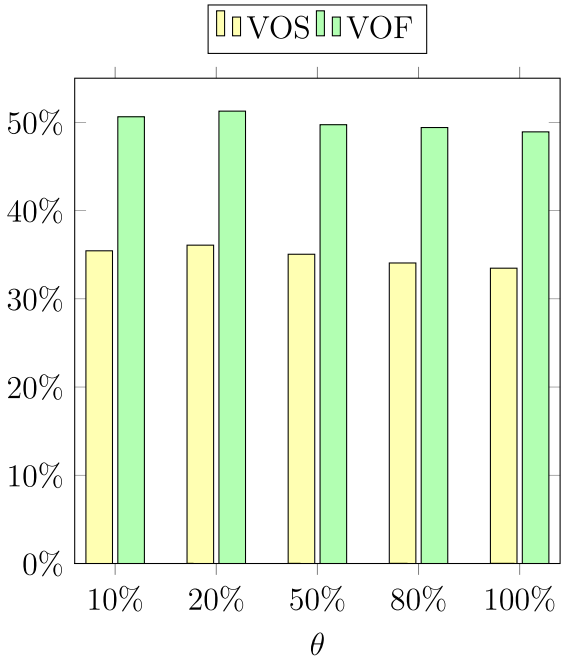


Fig. 13. The impact of allowing only module sharing (VOS) and both module sharing and rebalancing (VOF).

impact of both module sharing and rebalancing. The natural consequence is a more effective usage of the capacity deployed across the three systems, as confirmed by the analysis of the average occupancy ratio Ω (see the right panel of Fig. 12).

Fig. 13 illustrates the percentage reduction in the number of modules required in a transit system if only module sharing is allowed (i.e., the VOS), and the reduction achievable if both module sharing and rebalancing are simultaneously possible (i.e., the

VOF). The values depicted support the insights drawn above. Further, they show that the percentage reductions do not remarkably depend on the percentage of mobility requests θ. The value of the VOS is always roughly 35%, whereas the value of the VOF is larger and always approximately equal to 50%.

7. Conclusions and future research

In this paper, we studied a transit system that uses modular buses and compared its performance to a conventional transit system. We presented an integer linear program to determine an optimal assignment of modules to the sections of each line. The mathematical formulation exploits the flexibility enabled by modular buses, allowing sharing and rebalancing of modules. Extensive empirical experiments investigate the benefits of modular buses. On average, over all the tested instances the total capacity deployed is reduced by 49.44%. Additionally, the average occupancy ratio increases from 41.22% to 72.85%.

The technology of modular buses has the potential to adapt the capacity of a section of a line to the actual mobility demand of that section. Obviously, a complete assessment of this new technology will have to consider the costs of operating such a system. This aspect is out of the scope of the present paper, and it is left for a future research.

Future developments may concern the analysis of more tactical or operational issues. In this paper, we assumed that the demand is constant over time. Nevertheless, it would be interesting to incorporate into the analysis the variability over time of the mobility demand. It would also be interesting to further explore the relations between the layout of the network and the benefits provided by employing modular buses. Instead of assuming that the paths followed by the passengers are given, another interesting development would be to let the optimization model determine the best path assigned to each mobility request. In this case, it would be interesting to incorporate into the mathematical formulation some considerations regarding the inconvenience, as measured by a given metric, possibly generated to the passengers. For instance, a fairness measure could be introduced in the model to stimulate an even distribution, among all passengers, of the additional travel distance (or time) compared to their shortest path. A further extension of this work would be to study the combined optimization of headways (i.e., waiting times at stops) and number of modules assigned to each bus departure in a transit system with multiple lines. Such an extension, being more operational-oriented compared to the current paper, could also incorporate technical constraints modeling the main real-world complexities, such as the time needed to assemble/disassemble modules, battery capacities, charging breaks, energy consumption based on the modules composition and their load, road inclinations. Finally, while this paper assumed that the transit lines are given, it would also be interesting to study the design of the transit line network for the modular buses.

CRediT authorship contribution statement

Carlo Filippi: Writing – review & editing, Writing – original draft, Supervision, Methodology, Formal analysis, Conceptualization. **Gianfranco Guastaroba:** Writing – review & editing, Visualization, Validation, Supervision, Data curation, Conceptualization. **Lorenzo Peirano:** Writing – original draft, Visualization, Software, Data curation, Conceptualization. **M. Grazia Speranza:** Writing – review & editing, Visualization, Supervision, Methodology, Formal analysis, Conceptualization.

Acknowledgments

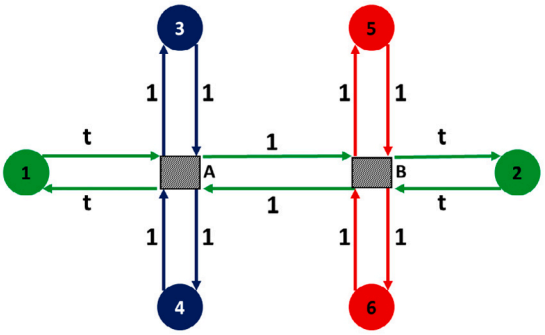
- This research has been supported and funded by the European Union (EU) and Italian Ministry for Universities and Research (MUR), National Recovery and Resilience Plan (NRRP), within the project “Sustainable Mobility Center (MOST)”,2022–2026, CUP D83C22000690001, Spoke N° 7, “CCAM, Connected networks and Smart Infrastructures”.
- This research has been supported and funded by the European Union (EU) with Project PRIN 2022: “Time-dependent optimization for sustainable transportation”, Code: 20223MHHA8, CUP: D53D23005590006, Funded by the European Union - Next Generation EU

Appendix

Proofs of Theorems 1 and 2

In the following, we refer to the notation introduced in Section 5.

Proof of Theorem 1. Consider the network $G(t)$, depending on parameter t, and the corresponding matrix $M(p)$, depending on parameter p, depicted in Fig. 14. For convenience, rebalancing arcs are not shown. In network $G(t)$, there are two junctions (nodes A and B), 6 terminals (nodes 1 through 6), and three lines: the horizontal green line H is the closed path $(1, A, B, 2, B, A, 1)$; the left blue vertical line LV is the closed path $(3, A, 4, A, 3)$; the right red vertical line RV is the closed path $(5, B, 6, B, 5)$. The figure associated with each arc in $G(t)$ represents the average travel time. Assume that the mobility demand is the one represented in matrix $M(p)$, and that the unit mobility demand in matrix $M(p)$ corresponds to the capacity of a module. According to the network structure and matrix $M(p)$, the passenger rates must be those reported in Fig. 15. If lines cannot share modules, the minimum rates of modules along all arcs of lines H, LV, and RV are $p + 2$, p, and p, respectively. Recall that without module sharing, the rate of modules assigned to a line remains constant over all line arcs of that line. Hence, it must be sufficient to transport the highest



	1	2	3	4	5	6	A	B
1	-	1	0	0	0	0	0	0
2	1	-	0	0	0	0	0	0
3	0	0	-	0	0	p	0	0
4	0	0	0	-	1	0	0	0
5	0	0	0	1	-	0	0	0
6	0	0	p	0	0	-	0	0
A	0	0	0	0	0	0	-	0
B	0	0	0	0	0	0	0	-

Fig. 14. The left panel shows network $G(t)$ along with the travel times, whereas matrix $M(p)$ is reported in the right panel.

passenger rate across all arcs traversed by a line. Thus, the minimum number of modules that must be in the system to satisfy that mobility demand is:

$$(4t + 2)(p + 2) + 8p.$$

If lines can share modules, the number of modules assigned to each arc can be adapted to the mobility demand on that specific link. In fact, by sharing modules at junctions A and B as shown in Fig. 16, it is possible to assign to each arc a rate of modules equal to the passenger flows shown in Fig. 15. Particularly, considering the blue line starting with p modules from the stop at the upper-left corner, $p - 1$ modules are shared at junction A with the green line (see the dotted line arrow) traveling toward the rightmost stop. Subsequently, at junction B, $p-1$ modules are shared between the green and the red line traveling toward the stop at the bottom-right corner. A similar sharing is carried out at junction B between the red line, starting from the stop at the bottom-right corner, and the green line, and then at junction A from the latter and the blue line traveling toward the stop at the upper-left corner. Furthermore, two modules assigned to the green line travel only between the two junctions A and B. At the latter stops, the two modules change their traveling direction, as indicated by the curved arrows. Technically speaking, there is no sharing between different lines, but two modules are shared between different modular buses of the same line that travel in opposite directions. Hence, if module sharing is allowed, the total required number of modules is:

$$4t + 4p + 2(p + 2) + 4.$$

Since:

$$\text{VOS}(G(t), M(p)) = \frac{[(4t + 2)(p + 2) + 8p] - [4t + 4p + 2(p + 2) + 4]}{(4t + 2)(p + 2) + 8p},$$

with a little algebra, one obtains:

$$\text{VOS}(G(t), M(p)) = \frac{2tp + 2t + 2p - 2}{2tp + 4t + 5p + 2}.$$

From this expression, one can determine:

$$\lim_{p\to\infty} \lim_{t\to\infty} \text{VOS}(G(t), M(p)) = 1$$

This proves the theorem. □

Proof of Theorem 2. Consider the network $G(t)$ and the corresponding matrix M depicted in Fig. 17. In network $G(t)$, all nodes are terminals without any junction visited by both lines. For convenience, only the rebalancing arcs relevant to this discussion are drawn. In this transport network, there are two lines: the upper red line is the closed path $(1, 2, 1)$; the lower blue line is the closed path $(3, 4, 3)$. The figure close to each arc indicates the corresponding average travel time. Assume that the unit mobility demand in matrix M corresponds to the capacity of a module. According to the network structure and matrix M, we have one unit of demand along arc $(1, 2)$, one unit of demand along arc $(3, 4)$, and zero demand along arcs $(2, 1)$ and $(4, 3)$.

If modules cannot be rebalanced, the two lines are independent, and the minimum rate of modules assigned to each arc is 1. Thus, the minimum number of modules that must be in the system to satisfy the mobility demand is:

$$2(t + 1).$$

If modules can be rebalanced using the dashed arcs in the left panel of Fig. 17, the number of modules assigned can be adapted to the mobility demand on that specific arc. This implies that we can avoid assigning modules to the arcs of duration t, since the

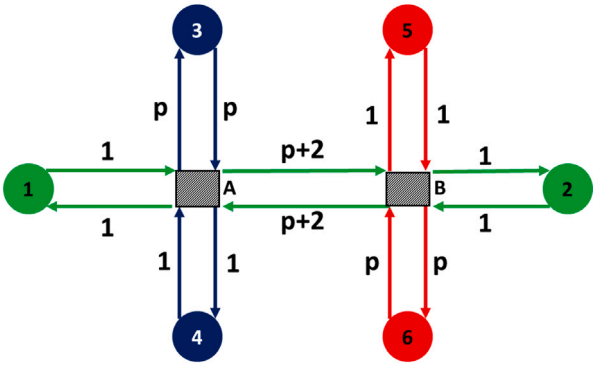


Fig. 15. Passenger rates for the example in Fig. 14.

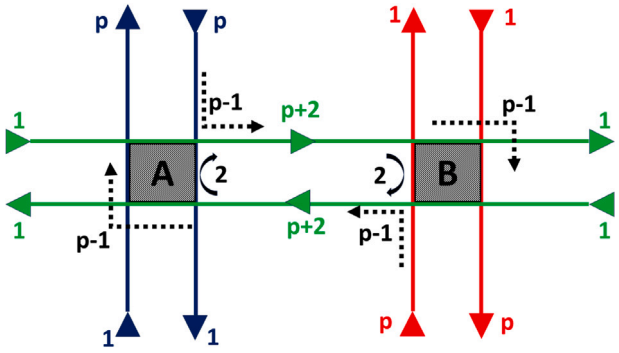
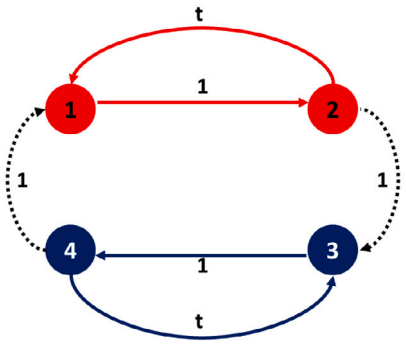


Fig. 16. The sharing of modules at junctions for the example in Fig. 14.



	1	2	3	4
1	-	1	0	0
2	0	-	0	0
3	0	0	-	1
4	0	0	0	-

Fig. 17. The left panel depicts network $G(t)$, whereas matrix M is reported in the right panel.

associated demand is zero, and one module travels along each rebalancing arcs $(2, 3)$ and $(4, 1)$. Thus, if rebalancing is allowed, the total required number of modules is 4. Hence:

$$\mathrm{VOR}(G(t), M) = \frac{2(t+1)-4}{2(t+1)} = 1 - \frac{2}{t+1}.$$

Since:

$$\lim_{t\to\infty} \mathrm{VOR}(G(t), M) = 1,$$

the claim follows. □

References

Chapuis, R., Tadjeddine, K., Chinn, D., Holmes, R., Knol, A., Speksnijder, L., Wolfs, K., Lotz, C., Stern, S., 2020. Restoring Public Transit Amid COVID-19: What

European Cities Can Learn from One Another. McKinsey & Company.
Chen, Z., Li, X., 2021. Designing corridor systems with modular autonomous vehicles enabling station-wise docking: Discrete modeling method. Transp. Res. Part E: Logist. Transp. Rev. 152, 102388.
Chen, Z., Li, X., Zhou, X., 2019. Operational design for shuttle systems with modular vehicles under oversaturated traffic: Discrete modeling method. Transp. Res. Part B: Methodol. 122, 1–19.
Chen, Z., Li, X., Zhou, X., 2020. Operational design for shuttle systems with modular vehicles under oversaturated traffic: Continuous modeling method. Transp. Res. Part B: Methodol. 132, 76–100.
Cheng, X., Nie, Y.M., Lin, J., 2024. An autonomous modular public transit service. Transp. Res. Part C: Emerg. Technol. 168, 104746.
Dai, Z., Liu, X.C., Chen, X., Ma, X., 2020. Joint optimization of scheduling and capacity for mixed traffic with autonomous and human-driven buses: A dynamic programming approach. Transp. Res. Part C: Emerg. Technol. 114, 598–619.
Dakic, I., Yang, K., Menendez, M., Chow, J.Y., 2021. On the design of an optimal flexible bus dispatching system with modular bus units: Using the three-dimensional macroscopic fundamental diagram. Transp. Res. Part B: Methodol. 148, 38–59.
Filippi, C., Guastaroba, G., Peirano, L., Speranza, M.G., 2023. Trends in passenger transport optimisation. Int. Trans. Oper. Res. 30 (6), 3057–3086.
Gao, H., Liu, K., Wang, J., Guo, F., 2023. Modular bus unit scheduling for an autonomous transit system under range and charging constraints. Appl. Sci. 13 (13), 7661.
Gecchelin, T., Webb, J., 2019. Modular dynamic ride-sharing transport systems. Econ. Anal. Policy 61, 111–117.
Gong, M., Hu, Y., Chen, Z., Li, X., 2021. Transfer-based customized modular bus system design with passenger-route assignment optimization. Transp. Res. Part E: Logist. Transp. Rev. 153, 102422.
Guihaire, V., Hao, J.-K., 2008. Transit network design and scheduling: A global review. Transp. Res. Part A: Policy Pr. 42 (10), 1251–1273.
Guo, R., Guan, W., Vallati, M., Zhang, W., 2023. Modular autonomous electric vehicle scheduling for customized on-demand bus services. IEEE Trans. Intell. Transp. Syst. 24 (9), 10055–10066.
Hannoun, G.J., Menendez, M., 2022. Modular vehicle technology for emergency medical services. Transp. Res. Part C: Emerg. Technol. 140, 103694.
Khan, Z.S., He, W., Menéndez, M., 2023. Application of modular vehicle technology to mitigate bus bunching. Transp. Res. Part C: Emerg. Technol. 146, 103953.
Khan, Z.S., Menéndez, M., 2023. Bus splitting bus holding: A new strategy using autonomous modular buses for preventing bus bunching. Transp. Res. Part A: Policy Pr. 177, 103825.
Khan, Z.S., Menéndez, M., 2024. A seamless bus network without external transfers using autonomous modular vehicles. Transp. Res. Part C: Emerg. Technol. 104822.
Lee, E., Cen, X., Lo, H.K., Ng, K.F., 2021. Designing zonal-based flexible bus services under stochastic demand. Transp. Sci. 55 (6), 1280–1299.
Lin, X., Chen, Z., Li, M., He, F., 2024. Bunching-proof capabilities of modular buses: An analytical assessment. Transp. Sci. 58 (5), 925–946.
Liu, T., Ceder, A., Rau, A., 2020. Using deficit function to determine the minimum fleet size of an autonomous modular public transit system. Transp. Res. Rec. 2674 (11), 532–541.
Liu, Y., Chen, Z., Wang, X., 2024. Alleviating bus bunching via modular vehicles. Transp. Res. Part B: Methodol. 189, 103051.
Liu, Z., de Almeida Correia, G.H., Ma, Z., Li, S., Ma, X., 2023. Integrated optimization of timetable, bus formation, and vehicle scheduling in autonomous modular public transport systems. Transp. Res. Part C: Emerg. Technol. 155, 104306.
Liu, X., Qu, X., Ma, X., 2021. Improving flex-route transit services with modular autonomous vehicles. Transp. Res. Part E: Logist. Transp. Rev. 149, 102331.
Pei, M., Lin, P., Du, J., Li, X., Chen, Z., 2021. Vehicle dispatching in modular transit networks: A mixed-integer nonlinear programming model. Transp. Res. Part E: Logist. Transp. Rev. 147, 102240.
Schöbel, A., 2012. Line planning in public transportation: Models and methods. OR Spectrum 34 (3), 491–510.
Shi, X., Chen, Z., Pei, M., Li, X., 2020. Variable-capacity operations with modular transits for shared-use corridors. Transp. Res. Rec. 2674 (9), 230–244.
Shi, X., Li, X., 2021. Operations design of modular vehicles on an oversaturated corridor with first-in, first-out passenger queueing. Transp. Sci. 55 (5), 1187–1205.
Tian, Q., Lin, Y.H., Wang, D.Z., 2023. Joint scheduling and formation design for modular-vehicle transit service with time-dependent demand. Transp. Res. Part C: Emerg. Technol. 147, 103986.
Wu, J., Kulcsár, B., Selpi, Qu, X., 2021. A modular, adaptive, and autonomous transit system (MAATS): An in-motion transfer strategy and performance evaluation in urban grid transit networks. Transp. Res. Part A: Policy Pr. 151, 81–98.
Xia, D., Ma, J., Azadeh, S.S., Zhang, W., 2023. Data-driven distributionally robust timetabling and dynamic-capacity allocation for automated bus systems with modular vehicles. Transp. Res. Part C: Emerg. Technol. 155, 104314.
Zhang, Z., Tafreshian, A., Masoud, N., 2020. Modular transit: Using autonomy and modularity to improve performance in public transportation. Transp. Res. Part E: Logist. Transp. Rev. 141, 102033.