Integrated Bus Fleet Electrification Planning Through Accelerated Logic-Based Benders Decomposition and Restriction Heuristics

Robin Legault[*]   Filipe Cabral[†]   Xu Andy Sun[‡]

Abstract

To meet sustainability goals and regulatory requirements, transit agencies worldwide are planning partial and complete transitions to electric bus fleets. This paper presents the first comprehensive and computationally efficient multi-period optimization framework integrating the key planning decisions necessary to support such electrification initiatives. Our model, formulated as a two-stage integer program with integer subproblems, jointly optimizes yearly fleet and charging infrastructure investments as well as hourly vehicle scheduling and charging operations. To solve instances of practical relevance to proven optimality, we develop a logic-based Benders decomposition method enhanced by several techniques, including preprocessing, partial decomposition, and a range of classical and monotone Benders cuts derived from relaxations of the operational subproblems. These accelerations yield speedups of up to three orders of magnitude and lead to practical and theoretical insights into Benders cut selection. We also propose a heuristic tailored for long-term, citywide electrification planning. This approach, which imposes and progressively relaxes additional scheduling constraints, consistently delivers high-quality solutions with optimality gaps below 1% for instances an order of magnitude larger than those considered in prior studies. We illustrate our model using data from the Chicago public bus system, providing managerial insights into optimal investment and operational policies.

Keywords: Integrated planning, Bus fleet electrification, Logic-based Benders decomposition

1 Introduction

Several major metropolitan transit agencies have established ambitious electrification targets for their bus fleets. In 2017, 35 cities across six continents pledged to procure only zero-emission buses starting from 2025 (C40 2023). In the United States, cities such as Boston, Chicago, and New York have committed to operating fully electrified fleets by 2040 (MBTA 2022, CTA 2022, MTA 2024b), and the California Air Resources Board's Innovative Clean Transit regulation mandates that all public transit agencies in the state replace conventional buses with zero-emission models by this time (CARB 2018). Although full electrification is the stated long-term goal, practical considerations such as funding availability, operational complexity, and infrastructure readiness also lead agencies to initiate shorter-term, partial electrification projects targeting selected routes or depots (e.g., CTA 2022, MARTA 2023, MTA 2024a). Despite these initiatives, public data from the Federal Transit Administration indicate that electric buses represented only 3% of the nation's 60,995 transit buses and accounted for just 1.4% of miles driven in 2022 (Davidson 2023).

Designing an effective electrification plan, whether partial or complete, requires jointly optimizing several interdependent decisions. While many studies have addressed individual aspects of electric bus planning and

---

[*]Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA. Email legault@mit.edu

[†]Georgia Institute of Technology, Atlanta, GA. Email fcabral1290@gmail.com

[‡]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA. Email sunx@mit.edu

scheduling, the review article by Perumal et al. (2022) highlights that integrated planning, where long-term strategic and short-term operational decisions are jointly considered, has received very little attention in the operations research literature. In this context, the goal of this work is to propose the first comprehensive and computationally tractable framework for integrated bus fleet electrification planning. We summarize our contributions as follows.

Contributions

1. **Modeling:** We introduce a new model for integrated bus fleet electrification planning (Section 3). At the strategic level, our model optimizes the procurement of battery electric buses (BEBs), retirement of conventional buses, and placement of chargers over multiple investment periods. At the operational level, we propose a flexible flow-based formulation to model the hourly operations of the fleet and track the state of charge at the individual vehicle level. The operations are optimized for a representative day of each period, ensuring that service-level requirements on each bus line are satisfied and that the same operations can be repeated daily. The resulting problem can be formulated as a two-stage program with integer subproblems. Using historical service schedules and geospatial data from the public transit agencies of eight major US cities, we construct a set of realistic benchmark instances for the electrification of partial and complete bus networks.

2. **Algorithmic Design and Computation:**

   • To solve small to medium-scale instances representative of real-life partial electrification projects, we develop an accelerated logic-based Benders decomposition algorithm leveraging preprocessing, partial decomposition, and custom cuts obtained from different relaxations of the operational model (Sections 4.1-4.2). We review Benders cut selection techniques, which leads us to establish the equivalence of two methods from the literature (Section 4.2.3). We evaluate the impact of each proposed acceleration through an ablation study, showing that our method achieves speedups of three orders of magnitude compared to a decomposition framework recently proposed for the same class of problems (Section 5.1). Our exact method significantly outperforms Gurobi in both computing time and optimality gaps (Section 5.2).

   • To tackle long-term citywide electrification planning, we propose an easily implementable heuristic that solves a sequence of problems in which restrictions on the schedule of BEBs are imposed and progressively removed (Section 4.3). Our results show that our approach consistently achieves optimality gaps around 1% within reasonable computation times for challenging 10-year-long electrification planning instances defined on complete networks with more than 100 potential charging locations, 100 routes, and 1000 buses (Section 5.3).

3. **Case Study:** We present a case study of the Chicago bus network (Section 5.4). We analyze the optimal sequencing of investments, the spatial deployment of charging infrastructure by technology, and trends in bus utilization throughout the electrification process. We provide sufficient conditions under which the optimal investment sequence exhibits a three-phase structure: (i) high-return early investments, (ii) a period of inaction, and (iii) deferred, low-return investments required for regulatory compliance. Our results suggest prioritizing early electrification of high-usage routes by deploying BEBs supported by fast chargers, while completing electrification in less dense areas with BEBs relying on overnight depot charging.

The remainder of the paper consists of a literature review (Section 2) and a conclusion (Section 6).

2 Integrated bus fleet electrification planning

The literature on BEB systems planning can be divided into three streams. First, in the electric vehicle scheduling problem (E-VSP), the goal is to construct schedules that respect the charging dynamics of a given fleet and satisfy predefined service requirements (e.g., Parmentier et al. 2023, de Vos et al. 2024). The second stream disregards vehicle routing and scheduling, and focuses on strategic choices such as fleet composition and charger placement (e.g., Xylia et al. 2017, Pelletier et al. 2019). Finally, some studies integrate strategic and operational decisions to design electrification plans that balance capital expenditures with operational costs. Our work contributes to this last stream. In Table 1, we review selected publications on integrated bus fleet electrification planning based on the decisions they model and the scale of the instances considered. For general reviews on electric bus planning and scheduling, see Perumal et al. (2022) and Zhou et al. (2024).

Table 1: Selected publications on integrated bus fleet electrification planning

	Strategic planning			Tactical/operational planning				Largest instance		
	Bus/battery selection	Chargers placement	Multi-period investments	Depot charging	On-route charging	Charging scheduling	Vehicle scheduling	Routes	Buses	Charging locations
Rogge et al. (2018)	✓			✓		✓	✓	3	14	1
Liu et al. (2021)		✓		✓	✓	✓	✓	17	252	15
Dirks et al. (2022)	✓	✓	✓	✓	✓	✓		60	357	182
Hu et al. (2022)	✓	✓			✓	✓		3	16	111
Wang et al. (2022)		✓		✓	✓	✓	✓	4	28	31
He et al. (2023a)	✓	✓		✓	✓	✓	✓	3	6	3
He et al. (2023b)	✓	✓	✓	✓	✓	✓		36	170	29
Gairola and Nezamuddin (2023)		✓		✓	✓	✓		18	285	21
This work	✓	✓	✓	✓	✓	✓	✓	140	1817	200

The first model to jointly optimize BEB fleet composition, charging infrastructure, and vehicle scheduling was proposed by Rogge et al. (2018). Their formulation constructs a heterogeneous fleet, determines the number of chargers to install at a single depot, and designs schedules to cover a set of timetabled trips. Building on this approach, Wang et al. (2022) and He et al. (2023a) also optimize the deployment of fast on-route chargers, which can be used between service trips during dwelling periods at terminal stations. In addition to the location and configuration of chargers Liu et al. (2021), models the hourly flow of vehicles between each route and charging station to construct bus schedules without relying on predefined timetabled trips. This first flow-based approach relies on simplifying assumptions, such as aggregating the state of charge for all buses assigned to the same route, but scales to larger instances than other models that optimize vehicle scheduling.

Hu et al. (2022) and Gairola and Nezamuddin (2023) focus principally on chargers placement. Both studies assume that the service trips of the BEBs are given and design charging infrastructure and charging schedules to meet the fleet's needs, allowing for heterogeneous batteries in the former study. The models of He et al. (2023b) and Dirks et al. (2022) are the only ones to address strategic planning as a multi-period process, thereby recognizing that fleet electrification is typically phased over time rather than accomplished in a single step. For computational tractability, these studies assign each new BEB to a predefined service schedule currently performed by a conventional bus, substituting a complex vehicle scheduling problem with a simple assignment decision for each vehicle. While this modeling choice greatly reduces problem complexity,

it limits operational decisions to inserting charging events into existing schedules, which comes at the cost of ignoring opportunities to optimize vehicle schedules in light of BEB range limitations and charging dynamics.

Most existing literature on integrated bus fleet electrification planning prioritizes new modeling approaches and case-study insights, with algorithmic development often treated as an afterthought. As a result, these studies either omit vehicle scheduling or are restricted to small bus systems. Notably, none of the works that incorporate vehicle scheduling solve instances to proven optimality. Rogge et al. (2018) and He et al. (2023a) use heuristic genetic algorithms, Liu et al. (2021) solve a surrogate relaxation without providing global optimality bounds, and Wang et al. (2022) solve the mixed-integer linear programming formulation of their model using Gurobi, which achieves a 4% optimality gap after 25 hours of computation on a single-period instance comprising four routes.

To address the limitations of existing models, we introduce a multi-period framework that jointly optimizes fleet management and charger placement while explicitly integrating vehicle scheduling. Our operational model employs a flow-based formulation to construct bus schedules without relying on timetabled trips and to track the state of charge at the individual vehicle level. With novel algorithms, we can achieve proven optimality for realistically sized instances and deliver near-optimal solutions on systems an order of magnitude larger than those considered in previous studies.

3 Problem formulation

This section presents our modeling approach for the bus fleet electrification problem (BFEP). The model integrates a strategic problem and an operational problem, which are respectively introduced in Section 3.1 and Section 3.2. Important properties of the problem are presented in Section 3.3.

3.1 Strategic problem

We consider the problem of minimizing the investment and operational costs incurred by a central planner over a multi-period transition horizon in which a bus fleet and its associated charging infrastructure are progressively updated to achieve electrification targets. In each investment period $p \in \mathcal{P} = \{1, \ldots, P\}$, the state $x_p = (\chi_p, \{\eta_{pr}\}_{r \in \mathcal{R}}) \in \mathbb{Z}_+^n$ of the system is described by charger-related variables $\chi_p = (\bar{\chi}^p, \widetilde{\chi}^p)$, and vehicle assignment variables $\eta_{pr} = (\bar{\eta}_r^p, \widetilde{\eta}_r^p, \widehat{\eta}_r^p)$ for each bus line $r \in \mathcal{R}$.

It is assumed that the BEBs available for purchase can be partitioned into depot BEBs that only use plug-in chargers located in depots and on-route BEBs that rely on fast chargers such as DC plug-in chargers or pantographs installed at bus terminals. In each period $p \in \mathcal{P}$, each vehicle operates on a unique bus route. The number of depot BEBs of each type $b \in \mathcal{B}$, of on-route BEBs, and of conventional buses assigned to route $r \in \mathcal{R}$ are respectively controlled by the decision variables $\bar{\eta}_{rb}^p$, $\widetilde{\eta}_r^p$, and $\widehat{\eta}_r^p$. The number of chargers of type $k \in \mathcal{K}$ installed at depot $i \in \mathcal{I}$ is given by $\bar{\chi}_{ik}^p$, and $\widetilde{\chi}_j^p$ denotes the number of on-route chargers at terminal $j \in \mathcal{J}$. The BFEP is formulated as follows:

$$\min_{x_p, \, \forall p \in \mathcal{P}} \quad \sum_{p \in \mathcal{P}} \gamma^{p-1} \left(I_p(x_p - x_{p-1}) + O_p(x_p) \right) \tag{1a}$$

$$\text{s.t.} \quad I_p(x_p - x_{p-1}) \leq I_p^{\mathrm{UB}}, \qquad\qquad\qquad \forall p \in \mathcal{P}, \tag{1b}$$

$$\bar{\chi}_{ik}^p \geq \bar{\chi}_{ik}^{p-1}, \quad \widetilde{\chi}_j^p \geq \widetilde{\chi}_j^{p-1}, \qquad\qquad \forall p \in \mathcal{P}, \, i \in \mathcal{I}, \, k \in \mathcal{K}, \, j \in \mathcal{J}, \tag{1c}$$

$$\sum_{r \in \mathcal{R}} \bar{\eta}_{rb}^p \geq \sum_{r \in \mathcal{R}} \bar{\eta}_{rb}^{p-1}, \quad \sum_{r \in \mathcal{R}} \widetilde{\eta}_r^p \geq \sum_{r \in \mathcal{R}} \widetilde{\eta}_r^{p-1}, \qquad \forall p \in \mathcal{P}, \, b \in \mathcal{B}, \tag{1d}$$

$$\sum_{r\in\mathcal{R}}\widehat{\eta}_r^p \leq \sum_{r\in\mathcal{R}}\widehat{\eta}_r^{p-1}, \qquad\qquad \forall p\in\mathcal{P},\ b\in\mathcal{B}, \qquad\qquad (1\text{e})$$

$$\sum_{r\in\mathcal{R}}\widehat{\eta}_r^p \leq \widehat{\eta}_p^{\mathrm{UB}}, \qquad\qquad \forall p\in\mathcal{P}, \qquad\qquad (1\text{f})$$

$$\sum_{k\in\mathcal{K}}\bar{\chi}_{ik}^p \leq \bar{\chi}_i^{\mathrm{UB}}, \quad \widetilde{\chi}_j^p \leq \widetilde{\chi}_j^{\mathrm{UB}}, \qquad\qquad \forall p\in\mathcal{P},\ i\in\mathcal{I},\ j\in\mathcal{J}, \qquad\qquad (1\text{g})$$

$$\bar{\chi}^p\in\mathbb{Z}_+^{\mathcal{I}\times\mathcal{K}},\ \widetilde{\chi}^p\in\mathbb{Z}_+^{\mathcal{J}}, \qquad\qquad \forall p\in\mathcal{P}, \qquad\qquad (1\text{h})$$

$$\bar{\eta}_r^p\in\mathbb{Z}_+^{\mathcal{B}},\ \widetilde{\eta}_r^p\in\mathbb{Z}_+,\ \widehat{\eta}_r^p\in\mathbb{Z}_+, \qquad\qquad \forall p\in\mathcal{P},\ r\in\mathcal{R}. \qquad\qquad (1\text{i})$$

The discount factor $\gamma > 0$ and the initial state x_0 are given. The function $O_p(x_p) := H_p(x_p) + \mathcal{Q}_p(x_p)$ denotes the cost of maintaining and operating the system in state x_p during period p. It is composed of a nondecreasing linear function $H_p(x_p)$ modeling fixed maintenance costs and the optimal value $\mathcal{Q}_p(x_p)$ of a fleet scheduling problem. The investment costs $I_p(x_p - x_{p-1})$ are assumed to depend linearly on the number of each type of assets acquired and sold in period p. Constraints (1b) impose an investment budget I_p^{UB} for each period $p\in\mathcal{P}$. Constraints (1c)–(1d) ensure that the chargers and BEBs acquired at any period remain in the system during the following periods, whereas (1e) indicates that the size of the conventional bus fleet cannot increase. Constraints (1f) impose retirement targets for the conventional buses, with $\widehat{\eta}_p^{\mathrm{UB}}$ being the maximal number of conventional buses allowed to remain in use in period $p\in\mathcal{P}$. Constraints (1g) specify the charger hosting capacity of each location. Constraints (1h)–(1i) give the domain of the strategic variables. We denote by $\mathcal{X}$ the set of feasible solutions to problem (1).

3.2 Operational problem

For each investment period $p\in\mathcal{P}$, the operational problem is to construct a minimum-cost schedule that satisfies minimum service level requirements and can be performed using the assets allocated at the strategic level. A day is divided into a set of time intervals $\mathcal{T} = \{0,1,\ldots,T-1\}$, viewed as the residue classes modulo T, to encode the daily cyclicity of the fleet's operations.

For each route $r\in\mathcal{R}$, the number of conventional buses in service during interval $t\in\mathcal{T}$ is denoted by $\widehat{w}_r^{pt}$. We consider that on-route BEB in operation must be assigned to a terminal $j\in\mathcal{J}(r)\subseteq\mathcal{J}$ to charge during layover times, with $\mathcal{R}(j) = \{r\in\mathcal{R}: j\in\mathcal{J}(r)\}$ denoting the set of routes connected to terminal j. We capture the fast chargers' power rating by a parameter ρ, defined as the maximum number of operating BEBs each charger can accommodate in a given service interval. The number of on-route BEBs of route r that rely on terminal j during a service interval t is denoted by $\widetilde{w}_{rj}^{pt}$. The operations of conventional buses and on-route BEBs are thus modeled as simple task assignments.

Depot BEBs differ in that their schedules are represented as circulations on cyclic time-expanded graphs of state of charge (see Appendix A.3 for an illustrative example). Specifically, for each route $r\in\mathcal{R}$ and each type of bus $b\in\mathcal{B}$, we maintain a graph with nodes $(t,s)\in\mathcal{T}\times\{0,\ldots,s_b\}$, where $s = 0$ and $s = s_b$ respectively represent the fully depleted and fully charged states. In each interval, a depot BEB can be in service, idle, or initiate a charging trip to one of the depots $i\in\mathcal{I}$ equipped with plug-in chargers of any type $k\in\mathcal{K}$. The flow variables w_{rbs}^{pt}, v_{rbs}^{pt}, and $\{z_{rbiks}^{pt}\}_{(i,k)\in\mathcal{I}\times\mathcal{K}}$ respectively represent the number of vehicles initiating each possible service, idling, and charging operations from node (t,s). Working for one interval reduces the state of charge by one unit, idling does not affect the battery level, and κ_{rbiks} intervals are needed to perform a round-trip from route r to depot i and fully recharge from state s using a type k charger. We denote by $\mathcal{S}_b^w = \{1,2,\ldots,s_b\}$, $\mathcal{S}_b = \{0,1,\ldots,s_b\}$ and $\mathcal{S}_b^z = \{0,1,\ldots,s_b-1\}$ the charging states from which service,

5

charging and idling operations can respectively be initiated. Denoting by $y_{pr} = (w_r^p, v_r^p, z_r^p, \widetilde{w}_r^p, \widehat{w}_r^p) \in \mathbb{Z}_+^{m_{pr}}$ the decisions variables that pertain to each route $r \in \mathcal{R}$, the problem can be expressed as:

$$\mathcal{Q}_p(x_p) := \min_{y_{pr}, \, \forall r \in \mathcal{R}} \quad \sum_{r \in \mathcal{R}} c_{pr}^{y \top} y_{pr} \tag{2a}$$

$$\text{s.t.} \quad \sum_{b \in \mathcal{B}} \sum_{r \in \mathcal{R}} \sum_{s \in \mathcal{S}_b^z} \sum_{l=0}^{\kappa_{rbiks}-1} z_{rbiks}^{p(t-l)} \leq \bar{\chi}_{ik}^p, \qquad \forall \, i \in \mathcal{I}, \; k \in \mathcal{K}, \; t \in \mathcal{T}, \tag{2b}$$

$$\sum_{r \in \mathcal{R}(j)} \widetilde{w}_{rj}^{pt} \leq \rho \widetilde{\chi}_j^p, \qquad \forall j \in \mathcal{J}, \; t \in \mathcal{T}, \tag{2c}$$

$$\sum_{b \in \mathcal{B}} \sum_{s \in \mathcal{S}_b^w} w_{rbs}^{pt} + \sum_{j \in \mathcal{J}(r)} \widetilde{w}_{rj}^{pt} + \widehat{w}_r^{pt} \geq d_r^{pt}, \qquad \forall r \in \mathcal{R}, \; t \in \mathcal{T}, \tag{2d}$$

$$w_{rbs}^{pt} + v_{rbs}^{pt} = \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} \sum_{s' \in \mathcal{S}_b^z} z_{rbiks'}^{p(t-\kappa_{rbiks'})} + v_{rbs}^{p(t-1)}, \qquad \forall r \in \mathcal{R}, \; b \in \mathcal{B}, \; t \in \mathcal{T}, \; s = s_b, \tag{2e}$$

$$w_{rbs}^{pt} + v_{rbs}^{pt} + \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} z_{rbiks}^{pt} = w_{rb(s+1)}^{p(t-1)} + v_{rbs}^{p(t-1)}, \qquad \forall r \in \mathcal{R}, \; b \in \mathcal{B}, \; t \in \mathcal{T} \; s \in \mathcal{S}_b \setminus \{0,1\}, \tag{2f}$$

$$\sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} z_{rbiks}^{pt} + v_{rbs}^{pt} = w_{rb(s+1)}^{p(t-1)} + v_{rbs}^{p(t-1)}, \qquad \forall r \in \mathcal{R}, \; b \in \mathcal{B}, \; t \in \mathcal{T}, \; s = 0, \tag{2g}$$

$$\sum_{s \in \mathcal{S}_b^w} w_{rbs}^{p0} + \sum_{s \in \mathcal{S}_b} v_{rbs}^{p0} + \sum_{i \in I} \sum_{k \in K} \sum_{s \in \mathcal{S}_b^z} \sum_{l=0}^{\kappa_{rbiks}-1} z_{rbiks}^{p(-l)} \leq \bar{\eta}_{rb}^p, \qquad \forall r \in \mathcal{R}, \; b \in \mathcal{B}, \tag{2h}$$

$$\sum_{j \in \mathcal{J}(r)} \widetilde{w}_{rj}^{pt} \leq \widetilde{\eta}_r^p, \quad \widehat{w}_r^{pt} \leq \widehat{\eta}_r^p, \qquad \forall r \in \mathcal{R}, \; t \in \mathcal{T}, \tag{2i}$$

$$w_r^p \in \mathbb{Z}_+^{\mathcal{T} \times \mathcal{B} \times \mathcal{S}_b^w}, \; v_r^p \in \mathbb{Z}_+^{\mathcal{T} \times \mathcal{B} \times \mathcal{S}_b}, \; z_r^p \in \mathbb{Z}_+^{\mathcal{T} \times \mathcal{B} \times \mathcal{I} \times \mathcal{K} \times \mathcal{S}_b^z}, \qquad \forall r \in \mathcal{R}, \tag{2j}$$

$$\widetilde{w}_r^p \in \mathbb{Z}_+^{\mathcal{T} \times \mathcal{J}(r)}, \; \widehat{w}_r^p \in \mathbb{Z}_+^{\mathcal{T}}, \qquad \forall r \in \mathcal{R}. \tag{2k}$$

The nonnegative objective coefficients c_{pr}^y include the energy costs, driver wages, and variable maintenance costs associated with each operational decision. For each depot $i \in \mathcal{I}$ and each charger type $k \in \mathcal{K}$, constraints (2b) ensure that the capacity $\bar{\chi}_{ik}^p$ is never exceeded by the number of BEBs simultaneously in charge. The charger usage at time $t \in \mathcal{T}$ is given by the number of BEBs of any type $b \in \mathcal{B}$ that initiated a charging operation toward a charger of type k at depot i from any route $r \in \mathcal{R}$ and any state of charge $s \in \mathcal{S}_b^z$ in the last κ_{rbiks} intervals. Constraints (2c) ensure that on-route charger usage respects the installed capacity at each terminal location $j \in \mathcal{J}$. Constraints (2d) indicate that the number of buses in service on each route $r \in \mathcal{R}$ and interval $t \in \mathcal{T}$ must be at least d_r^{pt}. Constraints (2e)–(2g) model the depot BEBs charging dynamics by enforcing flow conservation on the cyclic time-expanded graphs of state of charge for each route $r \in \mathcal{R}$ and bus type $b \in \mathcal{B}$. In each flow balance equation, the left- and right-hand sides respectively correspond to the outflow and inflow at a node $(t, s) \in \mathcal{T} \times \mathcal{S}_b$. For the fully charged state $s = s_b$, the outflow is given by the number of vehicles in service or idling in state s at time t, whereas the inflow corresponds to the number of vehicles that were idling in state s in interval $t - 1$, or initiated, $\kappa_{rbiks'}$ intervals earlier, a charging trip to any location $i \in \mathcal{I}$ and charger type $k \in \mathcal{K}$ from state $s' \in \mathcal{S}_b^z$, and thus become available and fully charged at time t. A similar logic applies to the partly depleted states $s \in \mathcal{S}_b \setminus \{0, 1\}$ and the fully depleted state $s = 0$. Constraints (2h) and (2i) ensure that the allocated fleet size is respected. Constraints (2j)–(2k) give the domain of the operational variables.

Throughout the paper, we will use the following compact form of problem (2):

$$\mathcal{Q}_p(x_p) := \min_{y_{pr} \in \mathbb{Z}_+^{m_{pr}}, \; \forall r \in \mathcal{R}} \quad \sum_{r \in \mathcal{R}} c_{pr}^{y\top} y_{pr} \tag{3a}$$

$$\text{s.t.} \quad \sum_{r \in \mathcal{R}} A_r y_{pr} \leq B\chi_p, \tag{3b}$$

$$D_r y_{pr} \leq e_{pr} + E\eta_{pr}, \quad \forall r \in \mathcal{R}. \tag{3c}$$

Constraints (3b) represent the charger availability constraints (2b)–(2c), and the route-separable constraints (3c) include the service level, battery dynamics, and fleet size constraints (2d)–(2i). If solution x_p leads to an infeasible operational problem, then $\mathcal{Q}_p(x_p) = +\infty$.

3.3 Properties of the problem

Although our depot BEB scheduling model closely resembles a minimum-cost circulation problem due to the flow conservation constraints (2e)–(2g), the chargers capacity constraints (2b) and service level requirements (2d) act as side constraints that break the total unimodularity property (see Schrijver 2003). It follows that instances with nonzero integrality gaps can easily be constructed.

Remark 1. *The operational problem* (2) *does not have the integrality property.*

As a consequence of Remark (1), classical Benders cuts do not suffice to obtain an exact decomposition algorithm for the BFEP. However, the strategic variables only appear as positive terms in the operational constraints, i.e., matrices B and E are nonnegative. This implies the result of Remark 2, which will be leveraged in our exact logic-based Benders decomposition approach.

Remark 2. *The optimal value* $\mathcal{Q}_p(x_p)$ *of problem* (2) *is a nonincreasing function of* x_p.

We conclude this section by highlighting that, even under simplifying assumptions, the BFEP is strongly NP-hard. The proofs of Propositions 1 and 2 are by reduction from the set covering problem (Hartmanis 1982). They are presented in Appendix A.2.

Proposition 1. *The BFEP with one investment period, one time interval, and only on-route BEBs is strongly NP-hard.*

Proposition 2. *The BFEP with one investment period and only depot BEBs is strongly NP-hard.*

4 Solution methodology

The BFEP can be expressed in extensive form by replacing in problem (1) the value function $\mathcal{Q}_p(x_p)$ by its integer programming formulation (3). However, the resulting formulation rapidly becomes intractable. Leveraging the structure of the problem is thus essential to achieve efficient solution methods. Section 4.1 presents preprocessing steps to reduce the size of the problem and strengthen its LP relaxation. In Section 4.2, we present our logic-based Benders decomposition method along with several acceleration techniques. Section 4.3 introduces our restriction heuristic algorithm.

4.1 Preprocessing

This section presents valid inequalities (i.e., inequalities satisfied by any feasible solution) and dual reductions (i.e., inequalities that can remove feasible solutions while guaranteeing that at least one optimal solution remains) for the BFEP. In Section 4.1.1, we exploit the set covering structure of the on-route BEB charging dynamics to devise dominance relations between terminals and eliminate unnecessary locations. Section 4.1.2 presents valid inequalities on the size of the depot BEB fleet.

4.1.1 Dominance relation between terminal locations

Since we assume the installation cost of on-route chargers to be linear in the number of chargers and to be the same at each terminal, the choice of the best location for fast chargers is driven by route covering. Definition 1 establishes a dominance relation between terminals $j \in \mathcal{J}$ based on their connectivity to bus lines and their hosting capacity $\widetilde{\chi}_j^{\text{UB}}$. We break equalities based on an arbitrary indexing $\mathcal{J} = \{j_1, j_2, \ldots, j_m\}$.

Definition 1. *Terminal j_b is dominated by terminal j_a if one of the following conditions holds:*

1. $\mathcal{R}(j_b) \subset \mathcal{R}(j_a)$,

2. $\mathcal{R}(j_b) = \mathcal{R}(j_a)$ and $\widetilde{\chi}_{j_b}^{UB} < \widetilde{\chi}_{j_a}^{UB}$,

3. $\mathcal{R}(j_b) = \mathcal{R}(j_a)$ and $\widetilde{\chi}_{j_b}^{UB} = \widetilde{\chi}_{j_a}^{UB}$ and $b < a$.

Proposition 3 provides valid dual reductions on the number of on-route chargers installed at dominated terminals. The proof is given in Appendix A.2.

Proposition 3. *Let $\mathcal{J}(j)$ be the set of terminals dominated by $j \in \mathcal{J}$. Imposing the following inequalities for each $j \in \mathcal{J}$, $j' \in \mathcal{J}(j)$, $p \in \mathcal{P}$ is a valid dual reduction for the BFEP :*

$$\widetilde{\chi}_{j'}^p \leq \max \left\{ 0, \left\lceil \frac{1}{\rho} \max_{p' \in \mathcal{P}, \ t \in \mathcal{T}} \sum_{r \in \mathcal{R}(j)} d_r^{p't} \right\rceil - \widetilde{\chi}_j^{UB} \right\}. \tag{4}$$

The upper bounds of Proposition (3) can be computed outside the optimization process. In addition to strengthening the upper bounds $\widetilde{\chi}_j^{\text{UB}}$, this inexpensive preprocessing step generally eliminates several terminals, hence reducing the size of both the strategic and operational models.

4.1.2 Valid inequalities on fleet size

The service level requirements (2d) drive investments in new vehicles. Indeed, they imply that the number of buses assigned to a route $r \in \mathcal{R}$ in period $p \in \mathcal{P}$ cannot be less than the peak demand $\max_{t \in \mathcal{T}} d_r^{pt}$. For conventional buses, which can be in service without interruption, this trivial bound on fleet size is tight, and the same holds for on-route BEBs given that enough on-route chargers can be installed. However, the LP relaxation of the operational model does not provide a tight lower bound on the number of depot BEBs needed to satisfy the service level constraints. We thus propose to precompute lower bounds on fleet size for each route.

Given that at most m conventional buses and on-route BEBs are assigned to route $r \in \mathcal{R}$ in period $p \in \mathcal{P}$, the minimum number of depot BEBs needed to satisfy the residual service level requirements can be computed by solving the following restricted problem:

$$\bar{\eta}_{pr}^{\text{LB}}(m) := \min_{\tilde{\eta}_r^p, \hat{\eta}_r^p \in \mathbb{Z}_+, \ \bar{\eta}_r^p \in \mathbb{Z}_+^{\mathcal{B}}, \ y_{pr} \in \mathbb{Z}_+^{m_{pr}}} \sum_{b \in \mathcal{B}} \bar{\eta}_{rb}^p \quad \text{s.t.} \quad D_r y_{pr} \le e_{pr} + E\eta_{pr}, \quad \hat{\eta}_r^p + \bar{\eta}_r^p \le m. \tag{5}$$

Problem (5) is constructed by projecting the operational problem (3) onto y_{pr}, relaxing the charging capacity constraints (3b), and taking the fleet assignment parameters η_{pr} as decision variables. It follows that $\bar{\eta}_{pr}^{\text{LB}}(m)$ is a valid lower bound on the number of depot BEBs that must be assigned to route r in period p to complement any fleet of $\hat{\eta}_r^p + \tilde{\eta}_r^p \in [0, m]$ conventional buses and on-route BEBs. These lower bounds can be added directly to the strategic problem (1) as Big-M constraints. However, doing so requires introducing up to $\max_{t \in \mathcal{T}} d_r^{pt}$ binary variables in the model for each route r and period p, i.e., one for each integer value of m for which problem (5) can provide a nontrivial lower bound. Instead, we propose to approximate these constraints using the piecewise-linear lower envelope of the set of points $\mathcal{M}_{pr} := \{(m, \bar{\eta}_{pr}^{\text{LB}}(m))\}_{m=0}^{\max_{t \in \mathcal{T}} d_r^{pt}}$. This lower envelope corresponds to the convex hull of the epigraph of the function $\bar{\eta}_{pr}^{\text{LB}}(m)$ evaluated for a number of buses m ranging from 0 to the peak demand on route r in period p, and can be expressed as a set of linear inequalities on the original strategic variables. The following constraints can thus be added to model (1) without introducing additional variables:

$$\left(\tilde{\eta}_r^p + \hat{\eta}_r^p, \sum_{b \in \mathcal{B}} \bar{\eta}_{rb}^p \right) \in \text{conv} \left(\text{epi} \left(\mathcal{M}_{pr} \right) \right). \tag{6}$$

4.2 Accelerated logic-based Benders decomposition algorithm

For any clause C, we define its indicator function by

$$1[C] := \begin{cases} 1, & \text{if } C \text{ holds}, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Let $\mathcal{X}_p$ be the projection of $\mathcal{X}$ onto x_p. By Remark 2, the value function $\mathcal{Q}_p : \mathcal{X}_p \to [0, +\infty]$ is nonincreasing. Therefore, for any solution $x_p' \in \mathcal{X}_p$, the function $\mathcal{D}_p^{\text{IP}} : \mathcal{X}_p \to [0, +\infty]$, defined as:

$$\mathcal{D}_p^{\text{IP}}(x_p; x_p') := \mathcal{Q}_p(x_p') 1[x_p \le x_p'], \tag{8}$$

where the inequality is taken componentwise, is a lower approximation of $\mathcal{Q}_p$ (i.e., $\mathcal{D}_p^{\text{IP}}(x_p; x_p') \le \mathcal{Q}_p(x_p')$ $\forall x_p \in \mathcal{X}_p$), and is strong at x_p' (i.e., $\mathcal{D}_p^{\text{IP}}(x_p'; x_p') = \mathcal{Q}_p(x_p')$). It follows that bounding the value function $\mathcal{Q}_p$ of each period $p \in \mathcal{P}$ by the family of functions $\{\mathcal{D}_p^{\text{IP}}(\ \cdot\ ; x_p') : x_p' \in \mathcal{X}_p\}$ allows us to reformulate the BFEP as the following master problem:

$$\min_{\substack{x \in \mathcal{X} \\ \Theta \in \mathbb{R}_+^{\mathcal{P}}}} \sum_{p \in \mathcal{P}} f_p^\top x_p + \sum_{p \in \mathcal{P}} \gamma^{p-1} \Theta_p \tag{9a}$$

$$\text{s.t.} \ \Theta_p \ge \mathcal{D}_p^{\text{IP}}(x_p; x_p'), \quad \forall p \in \mathcal{P}, \ \forall x_p' \in \text{S}(\mathcal{Q}_p), \tag{9b}$$

$$1[x_p \le x_p'] = 0, \quad \forall p \in \mathcal{P}, \ \forall x_p' \in \mathcal{X}_p \backslash \text{S}(\mathcal{Q}_p), \tag{9c}$$

where the support $\text{S}(\mathcal{Q}_p) := \{x_p \in \mathcal{X}_p \,|\, \mathcal{Q}_p(x_p) < +\infty\}$ of the extended real-valued function $\mathcal{Q}_p$ corresponds to the subset of $\mathcal{X}_p$ for which problem (3) is feasible, and the vector of coefficients $f_p \in \mathbb{R}^n$ synthesizes the

discounted investment costs and fixed maintenance costs of model (1). In this formulation, the monotone cuts (9b) and (9c) respectively act as optimality and feasibility cuts.

Problem (9) can be solved by constraint generation using a multi-cut approach. Constraints (9b) and (9c) are initially relaxed. At each iteration $l \in \{1, 2, \dots\}$, the current relaxed master problem is solved to optimality, and its optimal solution $(x^l, \Theta^l) \in \mathcal{X} \times \mathbb{R}_+^{\mathcal{P}}$ provides a lower bound $\sum_{p \in \mathcal{P}} f_p^\top x_p^l + \sum_{p \in \mathcal{P}} \gamma^{p-1} \Theta_p^l$ on the optimal value of the master problem. For each period $p \in \mathcal{P}$, the operational problem (3) is then solved for $x_p = x_p^l$. The feasibility cut $1[x_p \leq x_p^l] = 0$ is added to the relaxed master problem if $\mathcal{Q}(x_p^l) = +\infty$, and the optimality cut $\Theta_p \geq \mathcal{D}_p^{\mathrm{IP}}(x_p; x_p^l)$ is otherwise generated. Replacing each lower bound Θ^l by $\mathcal{Q}(x_p^l)$ in the objective function of the relaxed master problem then yields the upper bound $\sum_{p \in \mathcal{P}} f_p^\top x_p^l + \sum_{p \in \mathcal{P}} \gamma^{p-1} \mathcal{Q}_p(x_p^l)$. This process is repeated until the optimality gap reaches the desired tolerance. Given that $\mathcal{X}$ has finite cardinality (which holds for the BFEP assuming finite budget constraints), this logic-based Benders decomposition algorithm is guaranteed to converge in a finite number of iterations (Hooker and Ottosson 2003).

Unfortunately, this simple algorithm performs poorly in practice. It is also what Liu et al. (2024) recently observed in the context of local multi-energy system's optimal design. Their problem is similar to the BFEP in that it is formulated as a multi-period planning problem with MILP operational problems with nondecreasing value functions of the investment variables. For this class of problems, minimizing the investment and operational costs are antagonistic objectives. Since the operational model is completely ignored in the initial relaxed master problem, the first iterations of the algorithm visit solutions characterized by significant underinvestment. Consequently, as the monotone cuts (9b) and (9c) are only active over the lower orthant associated with the visited points $x_p^l \in \mathcal{X}_p$, the cuts generated in early iterations only provide a nontrivial approximation of the value function $\mathcal{Q}_p$ over a very limited subset of $\mathcal{X}_p$, and are therefore very weak. To alleviate this drawback, we develop a range of techniques to achieve a tight and computationally tractable approximation of the value functions $\mathcal{Q}_p$ as early as possible in the execution of the algorithm.

Sections 4.2.1 and 4.2.2 introduce relaxations of model (3) that are respectively used to disaggregate the operational costs by route and to strengthen the relaxed master problem formulation. Section 4.2.3 reviews Benders cut selection techniques and their application to the LP relaxations of our operational problems. Section 4.2.4 presents problem-specific feasibility cuts and an improved encoding technique for monotone cuts. The outline of our accelerated logic-based Benders decomposition algorithm is presented in Section 4.2.5.

4.2.1 Disaggregation of the operational problem

For a fixed strategic solution $x \in \mathcal{X}$, the BFEP decomposes into a set of P independent operational problems. This is exploited in the master problem (9), where a variable Θ_p bounds the scheduling costs for each period $p \in \mathcal{P}$. Here, we propose to further decompose each operational problem by route. Since the operational variables $\{y_{pr}\}_{r \in \mathcal{R}}$ of each period are readily partitioned by route, it suffices to apply the linking constraints (3b) individually to each route to obtain a separable relaxation of problem (3). The single-route operational problem of route $r \in \mathcal{R}$ in period $p \in \mathcal{P}$ is given by:

$$\widetilde{\mathcal{Q}}_{pr}(x_p) := \min_{y_{pr} \in \mathbb{Z}_+^{m_{pr}}} \quad c_{pr}^{y\top} y_{pr} \tag{10a}$$

$$\text{s.t. } A_r y_{pr} \leq B\chi_p, \tag{10b}$$

$$D_r y_{pr} \leq e_{pr} + E\eta_{pr}. \tag{10c}$$

10

Since the variables $\{y_{pr}\}_{r \in \mathcal{R}}$ and the matrices $\{A_r\}_{r \in \mathcal{R}}$ are nonnegative, constraints (10b) are implied by (3b) for each route $r \in \mathcal{R}$. For any feasible solution $\{\bar{y}_{pr}\}_{r \in \mathcal{R}}$ to the operational problem (3), $\bar{y}_{pr}$ is thus feasible for (10) for each route $r \in \mathcal{R}$. The objective coefficients of each variable being identical in both formulations, solving the single-route subproblems separately and summing their objectives yields a lower bound on the original operational problem, i.e. $\sum_{r \in \mathcal{R}} \widetilde{\mathcal{Q}}_{pr}(x_p) \leq \mathcal{Q}_p(x_p)$. To take advantage of the single-route subproblems, we thus define an auxiliary variable θ_{pr} for each period $p \in \mathcal{P}$ and each route $r \in \mathcal{R}$, and add the following constraints to the master problem:

$$\Theta_p = \sum_{r \in \mathcal{R}} \theta_{pr}, \quad \forall p \in \mathcal{P}. \tag{11}$$

The LP relaxation of the single-route subproblems (10) is used to generate classical Benders optimality cuts on the θ_{pr} variables, as well as feasibility cuts. Although they apply to a relaxation of the operational problem, the single-route cuts have the advantage of being sparse, as each of them only involves the subset of strategic decisions that impact the operations on a specific route. Furthermore, they can be generated in large numbers at each iteration of the algorithm, which allows to obtain a good approximation of the value functions $\mathcal{Q}_p$ in fewer iterations.

4.2.2 Partial decomposition

A common drawback of Benders decomposition is that the initial relaxed master usually provides a weak relaxation of the problem. This limitation can be mitigated by using an alternative master formulation that includes explicit information from the subproblems. In the context of two-stage stochastic programming, the partial Benders decomposition approach, which consists of retaining a set of second-stage scenarios in the master problem, can significantly improve the overall performance of the algorithm (Crainic et al. 2021). Alternatively, for two-stage problems with integer subproblems, the initial master problem can be strengthened by retaining all the second-stage variables and relaxing their integrality (Gendron et al. 2016). However, the resulting master formulation, sometimes described as *semi-relaxed* (Liu et al. 2024), may include a prohibitively large number of variables and become impractical to solve.

We propose an alternative master problem formulation inspired by the partial Benders decomposition and semi-relaxation approaches. The idea is to include in the master problem a compact LP relaxation of the operational problem. To this end, we define for each period $p \in \mathcal{P}$ a collection $\omega_p = \{\omega_{pr}\}_{r \in \mathcal{R}}$ of nonnegative continuous variables representing the average service level provided by each type of bus on each route. Their feasible set $\Omega_p(x_p)$ is constructed from surrogate relaxations and valid linear inequalities devised from model (2), and lower bounds on the operating costs of each type of vehicle are taken to define their objective coefficients $\{c_{pr}^{\omega}\}_{r \in \mathcal{R}}$. The operational costs on each route are then bounded by the following constraints:

$$\theta_{pr} \geq c_{pr}^{\omega} \omega_{pr}, \quad \forall p \in \mathcal{P}, \forall r \in \mathcal{R}. \tag{12}$$

A detailed formulation of the approximate model is provided in Appendix A.4.

4.2.3 Benders cut selection

The monotone cuts (9b) and (9c) suffice to construct strong approximation of the value functions $\mathcal{Q}_p$, $p \in \mathcal{P}$. However, relying exclusively on these non-convex constraints is a bad strategy for three reasons. First, their generation requires solving integer subproblems, which can be computationally expensive. Second, monotone

cuts alone are generally ineffective at increasing the optimal value of the relaxed master problem, as they are active on a small region of the master problem domain. Third, generating monotone cuts requires adding binary variables to the master problem. These limitations can be mitigated by generating Benders cuts from LP relaxations of the subproblems before resorting to monotone cuts.

In this section, we present a unified review of the main Benders cut selection methods from the literature, and study their application to our problem. Although we use the notation of model (3), the discussion is general and not limited to the present application. This will lead us to the main result of this section, which we state in Proposition 4. A standalone proof is provided in Appendix A.2.

Proposition 4. *The closest cuts of Seo et al. (2022) and the Conforti–Wolsey deepest cuts of Hosseini and Turner (2024) are equivalent.*

Standard cuts. Let λ_p and $\{\mu_{pr}\}_{r\in\mathcal{R}}$ be the dual vectors of constraints (3b) and (3c), respectively, and let $\pi = (\lambda_p, \{\mu_{pr}\}_{r\in\mathcal{R}})$ denote their concatenation. The dual of the LP relaxation of the operational problem (3) of period $p \in \mathcal{P}$ can be expressed as:

$$\max_{\pi \in \Pi_p} \; \mathcal{D}_p^{\mathrm{LP}}(x_p; \pi), \tag{13}$$

where the objective is defined by $\mathcal{D}_p^{\mathrm{LP}}(x_p; \pi) := \lambda_p^\top B \chi_p + \sum_{r\in\mathcal{R}} \mu_{pr}^\top (e_{pr} + E\eta_{pr})$ and the dual feasible set is $\Pi_p := \{\pi = (\lambda_p, \{\mu_{pr}\}_{r\in\mathcal{R}}) \leq 0 : A_r^\top \lambda_p + D_r^\top \mu_{pr} \leq c_{pr}^y \; \forall r \in \mathcal{R}\}$. Note that the problem is always dual-feasible given that the vectors of coefficients $\{c_{pr}^y\}_{r\in\mathcal{R}}$ are nonnegative. Consequently, whenever the primal subproblem associated with a master solution $x_p' = (\chi_p', \eta_p') \in \mathcal{X}_p$ is infeasible, the dual admits at least one unbounded direction, i.e., a solution $\bar{\pi}_p = (\bar{\lambda}_p, \{\bar{\mu}_{pr}\}_{r\in\mathcal{R}}) \leq 0$ respecting $A_r^\top \bar{\lambda}_p + D_r^\top \bar{\mu}_{pr} \leq 0 \; \forall r \in \mathcal{R}$ and $\mathcal{D}_p^{\mathrm{LP}}(x_p'; \bar{\pi}_p) > 0$. In this case, the standard cut generation method consists of selecting such an unbounded direction $\bar{\pi}_p$ arbitrarily, which yields the feasibility cut:

$$\mathcal{D}_p^{\mathrm{LP}}(x_p; \bar{\pi}_p) \leq 0. \tag{14}$$

If the dual subproblem is bounded and admits an optimal solution $\bar{\pi}_p \in \Pi_p$, then $\mathcal{D}_p^{\mathrm{LP}}(x_p; \bar{\pi}_p)$ is a lower approximation of the value function $\mathcal{Q}_p^{\mathrm{LP}}(x_p)$ of the LP relaxation of model (3), and the optimality cut (15) can be added to the master:

$$\Theta_p \geq \mathcal{D}_p^{\mathrm{LP}}(x_p; \bar{\pi}_p). \tag{15}$$

The basic Benders cut selection approach consists of solving the primal subproblem and retrieving an optimal dual solution from the LP solver if the primal is feasible, and an unbounded extreme ray otherwise. In both cases, the cut is selected arbitrarily and can often be weak. This motivated the development in the literature of systematic cut selection techniques.

Magnanti-Wong (MW) cuts. A first strategy is to make use of a fixed master solution to guide the selection of Benders cuts. This idea was first exploited in the seminal work of Magnanti and Wong (1981) on the selection of strong optimality cuts for subproblems with complete recourse.

Let $\pi^1, \pi^2 \in \Pi_p$ be two feasible solutions of the dual problem (13). The optimality cut associated with π^1 is said to be dominated by the one provided by π^2 if $\mathcal{D}_p^{\mathrm{LP}}(x_p; \pi^2) \geq \mathcal{D}_p^{\mathrm{LP}}(x_p; \pi^1) \; \forall x_p \in \mathcal{X}_p$ and there is at least one solution in $\mathcal{X}_p$ for which the inequality is strict. Magnanti and Wong (1981) showed that one can leverage any core point, that is, a point in the relative interior $\mathrm{ri}(\mathcal{X}_p^c)$ of the convex hull $\mathcal{X}_p^c$ of $\mathcal{X}_p$, to identify a nondominated (also called Pareto-optimal) optimality cut. More precisely, an optimal solution $\bar{\pi}_p \in \Pi_p$ to

problem (13) that provides the best dual bounding function attainable at a core point $x_p^o \in \mathrm{ri}(\mathcal{X}_p^c)$, i.e., that satisfies $\mathcal{D}_p^{\mathrm{LP}}(x_p^o; \bar{\pi}) \geq \mathcal{D}_p^{\mathrm{LP}}(x_p^o; \pi) \ \forall \pi \in \Pi_p$, is Pareto-optimal. For a solution $x_p \in \mathcal{X}_p$, this yields the MW separation problem:

$$\max_{\pi \in \Pi_p} \ \mathcal{D}_p^{\mathrm{LP}}(x_p^o; \pi) \quad \text{s.t.} \quad \mathcal{D}_p^{\mathrm{LP}}(x_p; \pi) = \mathcal{Q}_p^{\mathrm{LP}}(x_p). \tag{16}$$

It was shown by Papadakos (2008) that, for some classes of problems, milder conditions than being a core point suffice to ensure that a guiding point x_p^o leads to Pareto-optimal cuts. However, the main drawback of the MW cut selection technique is that it cannot be applied when the dual subproblem is unbounded, in which case one has to resort to standard feasibility cuts.

Minimal infeasible subsystem (MIS) cuts. It was observed by Fischetti et al. (2010) that identifying a violated Benders cut can be formulated as a pure separation problem. Given a solution (x', Θ') to the master problem (9), a violated cut exists for the LP relaxation of the operational problem (3) of period $p \in \mathcal{P}$ if and only if the extended primal feasibility subproblem (17) is infeasible.

$$\min_{y_{pr} \geq 0, \ \forall r \in \mathcal{R}} \ 0 \quad \text{s.t.} \quad \sum_{r \in \mathcal{R}} c_{pr}^{y\top} y_{pr} \leq \Theta_p', \quad \sum_{r \in \mathcal{R}} A_r y_{pr} \leq B\chi_p', \quad D_r y_{pr} \leq e_{pr} + E\eta_{pr}', \quad \forall r \in \mathcal{R}. \tag{17}$$

Denoting by $\Pi_p^0 := \{(\pi, \pi_0) = (\lambda_p, \{\mu_{pr}\}_{r \in \mathcal{R}}, \pi_0) \leq 0 : \pi_0 c_{pr}^y + A_r^\top \lambda_p + D_r^\top \mu_{pr} \leq 0 \ \forall r \in \mathcal{R}\}$ the feasible domain of the dual of problem (17), whose objective is to maximize $\pi_0 \Theta_p' + \mathcal{D}_p^{\mathrm{LP}}(x_p'; \pi)$, the infeasibility of (17) equivalently means that the following normalized dual is feasible:

$$\min_{(\pi, \pi_0) \in \Pi_p^0} \ 0 \quad \text{s.t.} \quad \pi_0 \Theta_p' + \mathcal{D}_p^{\mathrm{LP}}(x_p'; \pi) = 1. \tag{18}$$

A solution $(\bar{\pi}, \bar{\pi}_0)$ to problem (18) yields a Benders cut of the form:

$$\bar{\pi}_0 \Theta_p + \mathcal{D}_p^{\mathrm{LP}}(x_p; \bar{\pi}) \leq 0. \tag{19}$$

If $\bar{\pi}_0 = 0$, the unified cut (19) simplifies to the feasibility cut (14) associated with solution $\bar{\pi}$. Otherwise, it corresponds to the optimality cut (15) associated with solution $-\bar{\pi}/\bar{\pi}_0$. When the master solution (x_p', Θ_p') violates both feasibility and optimality cuts, problem (18) yields an arbitrary violated cut, whereas the standard cut selection method always produces a feasibility cut.

The key idea of Fischetti et al. (2010) is to replace the trivial objective function of problem (18) by a linear function $h : \Pi_p^0 \to \mathbb{R}_+$. Since the objective and the left-hand side of the normalization constraint are both positive homogeneous in the dual variables (π, π_0), their role can be inverted (Cornuéjols and Lemaréchal 2006), which yields the unified Benders cut selection problem (20).

$$\max_{(\pi, \pi_0) \in \Pi_p^0} \ \pi_0 \Theta_p' + \mathcal{D}_p^{\mathrm{LP}}(x_p'; \pi) \quad \text{s.t.} \quad h(\pi, \pi_0) = 1. \tag{20}$$

Fischetti et al. (2010) propose to define h as the magnitude of the coefficients (π, π_0) multiplying nontrivial functions of the master problem's variables (x_p, Θ_p) in the unified cut (19). This choice seeks to identify a minimal infeasible system (Gleeson and Ryan 1990) in the constraints of the extended primal (17). Denoting by $S(M)$ the row support of a matrix M, the MIS cut selection problem is obtained by using the following

normalization function:

$$h_{MIS}(\pi, \pi_0) := -\pi_0 - \sum_{i \in S(B)} \lambda_{pi} - \sum_{r \in \mathcal{R}} \sum_{i \in S(E)} \mu_{pri}. \tag{21}$$

Closest cuts. A solution $(x_p^o, \Theta_p^o) \in \mathcal{X}_p^c \times \mathbb{R}_+$ that satisfies $\Theta_p^o \geq \mathcal{Q}_p^{\mathrm{LP}}(x_p^o)$ does not violate any Benders cuts. Consequently, a unified cut (19) is violated by a point (x_p', Θ_p') if and only if it defines a hyperplane that separates (x_p', Θ_p') from the guiding point (x_p^o, Θ_p^o). From this observation, Seo et al. (2022) proposed to select a Benders cut based on its distance from (x_p^o, Θ_p^o) at its intersection point with the half-line $\{(x_p^o, \Theta_p^o) + \beta(x_p' - x_p^o, \Theta_p' - \Theta_p^o) | \beta \geq 0\}$ joining (x_p^o, Θ_p^o) and (x_p', Θ_p'). Assuming that the unified cut (19) associated with solution (π, π_0) is violated by (x_p', Θ_p'), setting it to equality for $(x_p, \Theta_p) = (x_p^o, \Theta_p^o) + \beta(x_p' - x_p^o, \Theta_p' - \Theta_p^o)$ and solving for β gives:

$$\beta = 1 - \frac{\pi_0 \Theta_p' + \mathcal{D}_p^{\mathrm{LP}}(x_p'; \pi)}{\pi_0(\Theta_p' - \Theta_p^o) + \mathcal{D}_p^{\mathrm{LP}}(x_p'; \pi) - \mathcal{D}_p^{\mathrm{LP}}(x_p^o; \pi)}. \tag{22}$$

The closest cut from (x_p^o, Θ_p^o) that is violated by (x_p', Θ_p') can be identified by solving problem (18) with β as the objective function. By imposing the constraint of problem (18) to the numerator of (22), the objective can then be reformulated as minimizing the expression in the denominator. From there, we notice that the closest cut selection problem corresponds to the unified Benders cut selection problem (20), with normalization function $h_{CC}(\pi, \pi_0) := \pi_0(\Theta_p' - \Theta_p^o) + \mathcal{D}_p^{\mathrm{LP}}(x_p'; \pi) - \mathcal{D}_p^{\mathrm{LP}}(x_p^o; \pi)$.

Deepest cuts. For a solution (x_p', Θ_p'), the deepest cut selection framework of Hosseini and Turner (2024) identifies the violated cut (19) defining the hyperplane whose distance from (x_p', Θ_p') is the largest with respect to some pseudonorm. This approach can be seen as a natural extension of the work of Fischetti et al. (2010), where $h : \Pi_p^0 \to \mathbb{R}$ is taken as a general positive homogeneous normalization function in the unified Benders cut selection problem (20). Indeed, Hosseini and Turner (2024) show that taking $h_{\ell_q}(\pi, \pi_0) := ||(\pi_0, \lambda_p^\top B, \{\mu_{pr}^\top E\}_{r \in \mathcal{R}})||_q$ yields the violated cut whose ℓ_q distance from the current solution is maximized.

Moreover, they show that the technique proposed by Conforti and Wolsey (2019) to identify a facet-defining cut separating a point from a polyhedron can be exploited within the deepest cut framework. In the context of Benders cut selection, the considered polyhedron is the epigraph $\mathcal{E} := \{(x_p, \Theta_p) \in \mathcal{X}_p \times \mathbb{R}_+ : \Theta_p \geq \mathcal{Q}_p^{\mathrm{LP}}(x_p)\}$ of the Benders's subproblem value function. The separation problem of Conforti and Wolsey (2019) is defined based on a point (x_p^o, Θ_p^o) in the relative interior of $\mathcal{E}$, i.e., a core point x_p^o and a value $\Theta_p^o > \mathcal{Q}_p^{\mathrm{LP}}(x_p^o)$. It identifies a facet of $\mathcal{E}$ that is traversed by the line joining (x_p', Θ_p') and (x_p^o, Θ_p^o). Hosseini and Turner (2024) show that this facet corresponds to the cut (19) returned by the unified Benders cut selection problem (20) associated with the so-called Conforti–Wolsey pseudonorm $h_{CW}(\pi_0, \pi) := \pi_0(\Theta_p' - \Theta_p^o) + \lambda_p^\top B(\chi_p' - \chi_p^o) + \sum_{r \in \mathcal{R}} \mu_{pr}^\top E(\eta_{pr}' - \eta_{pr}^o)$. By expanding the definition of $\mathcal{D}_p^{\mathrm{LP}}(x_p; \pi)$ in the closest cut normalization function $h_{CC}(\pi, \pi_0)$, we observe that $h_{CC}(\pi, \pi_0) = h_{CW}(\pi_0, \pi)$, and Proposition 4 directly follows. To the best of our knowledge, this equivalence has not been highlighted in the literature. In particular, although Seo et al. (2022) discuss an earlier version of Hosseini and Turner (2024) that already presented the Conforti–Wolsey deepest cuts, neither of the two papers seems to have observed the equivalence between the closest cuts and the Conforti-Wolsey deepest cuts.

Implementation. In our algorithm, we generate Benders cuts from the operational problems (3) and their single-route relaxations (10). The single-route cuts of period $p \in \mathcal{P}$ for route $r \in \mathcal{R}$ are obtained by replacing

everywhere in the above section the set of routes $\mathcal{R}$ by the singleton $\{r\}$, the auxiliary variable Θ_p by θ_{pr}, and the value function $\mathcal{Q}_p$ by $\widetilde{\mathcal{Q}}_{pr}$.

For the MW and closest cuts, we use the procedure proposed by Seo et al. (2022) to identify a guiding point $(x_p^o, \Theta_p^o) \in \mathcal{X}_p^c \times \mathbb{R}_+$ that does not violate any Benders cut. First, we solve the LP relaxation of problem (1) in extensive form, which gives a solution (x_p', y_p') for each period $p \in \mathcal{P}$. We then set Θ_p^o to the objective value $\sum_{r \in \mathcal{R}} c_{pr}^{y\top} y_{pr}'$ of solution y_p' for the operational problem (3), similarly define $\theta_{pr}^o = c_{pr}^{y\top} y_{pr}'$ for single-route problems, and take $x_p^o = x_p' + \epsilon \mathbf{1}$ for $\epsilon = 0.1$. The added perturbation, which slightly relaxes the operational constraints (3b)–(3c), makes it more likely for x^o to be a core point and for the closest cut separation problems (20) to have unique solutions.

In Appendix A.8, we present a computational comparison of the standard cuts, the MIS cuts, the MW cuts, the closest cuts, and the $\ell-1$ deepest cuts as part of our accelerated logic-based Benders decomposition. The closest cuts achieve the best performance and are used throughout Section 5.

4.2.4   Monotone cuts

If a master problem iteration returns a solution $(x', \Theta') \in \mathcal{X} \times \mathbb{R}_+^{\mathcal{P}}$ for which (x_p', Θ_p') violates a Benders cut (19) in at least one period $p \in \mathcal{P}$, one can directly proceed to the next iteration without considering integer subproblems. Monotone cuts are only needed when classical cuts do not suffice to eliminate a solution from the feasible domain of the restricted master problem. In the integer phase of our algorithm, we solve for each period $p \in \mathcal{P}$ the integer subproblem (3) associated with solution x_p'. If this problem is feasible and its optimal value is strictly underestimated at the current master solution, i.e., $\Theta_p' < \mathcal{Q}_p(x_p') < \infty$, then we generate the optimality cut $\Theta_p \geq \mathcal{D}_p^{\mathrm{IP}}(x_p; x_p')$. However, if the integer problem is infeasible, instead of directly generating a generic feasibility cut $\mathbf{1}[x_p \leq x_p'] = 0$, we execute a sequence of tests aiming at identifying a subset of components of x_p' that cause the current solution to be infeasible.

Strengthened monotone feasibility cuts.   First, for each route $r \in \mathcal{R}$ and period $p \in \mathcal{P}$, we verify whether the fleet assignment decisions $\eta_r'^p$ imply the infeasibility of the operational problem by solving the single-route feasibility problem $y_{pr} \in \mathbb{Z}_+^{m_{pr}}, D_r y_{pr} \leq e_{pr} + E \eta_{pr}'$, which relaxes the charging capacity constraints (3b). Since on-route BEBs and conventional buses can be continuously in service in this model, if $\mathcal{Y}_{pr}(\eta_{pr}') := \{y_{pr} \in \mathbb{Z}_+^{m_{pr}} : D_r y_{pr} \leq e_{pr} + E \eta_{pr}'\} = \emptyset$, then we conclude that the fleet of depot BEBs is insufficient to satisfy the service level requirements unless the fleet size of conventional buses and on-route BEBs is increased. We obtain the following feasibility cut:

$$\mathbf{1}[(\bar{\eta}_r^p, \widetilde{\eta}_r^p + \widehat{\eta}_r^p) \leq (\bar{\eta}_r'^p, \widetilde{\eta}_r'^p + \widehat{\eta}_r'^p)] = 0. \tag{23}$$

If the single-route fleet assignment tests are inconclusive, we solve a multi-route feasibility problem that considers the on-route charging infrastructure and the total number of installed depot chargers. We replace the original set $\mathcal{I}$ of depots by a unique aggregate depot i^* and the original set $\mathcal{K}$ of charger types by a unique model k^*. For each tuple $(r, b, s) \in \mathcal{R} \times \mathcal{B} \times [0 : s_b - 1]$, we set the charging time parameter of the aggregate depot to $\kappa_{rbi^*k^*s} = \min_{i \in \mathcal{I}, k \in \mathcal{K}} \kappa_{rbiks}$. This has for effect of aggregating the depot charging capacity constraints and relaxing them by considering optimistic charging times. The feasibility test consists in solving the operational problem of period p for the current fleet and on-route chargers $(\widetilde{\chi}_p', \eta_p')$, with the singletons $\{i^*\}$ and $\{k^*\}$ replacing the original sets $\mathcal{I}$ and $\mathcal{K}$, and the objective of minimizing the number of depot chargers, which becomes a variable. If $\bar{\chi}_p'^{\mathrm{LB}} := \min\{\bar{\chi}_{i^*k^*}^p \in \mathbb{Z}_+ : (3\mathrm{b}) - (3\mathrm{c}), y_{pr} \in \mathbb{Z}_+^{m_{pr}} \; \forall r \in \mathcal{R}\}$

exceeds the number of depot chargers $\sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} \bar{\chi}_{ik}^{\prime p}$ from the current solution, we conclude that the total number of depot chargers or at least one component of the fleet sizing or on-route chargers installation decisions vectors must be increased to recover feasibility. This results in the following feasibility cut:

$$1 \left[\left(\eta^p, \widetilde{\chi}^p, \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} \bar{\chi}_{ik}^p \right) \leq \left(\eta'^p, \widetilde{\chi}'^p, \bar{\chi}_p'^{\mathrm{LB}} - 1 \right) \right] = 0. \tag{24}$$

Finally, if these feasibility tests fail to eliminate x_p' from the feasible domain of the relaxed master problem, we resort to the generic feasibility cut, which states that at least one component of the strategic solution x_p must be strictly higher than in solution x_p':

$$1[x_p \leq x_p'] = 0. \tag{25}$$

It is immediate that $(23) \implies (24) \implies (25)$, with the first implication holding for any route $r \in \mathcal{R}$. In addition to eliminating a larger subset of the master problem's domain, the sparser infeasibility cuts require fewer componentwise comparisons in their indicator function.

Sparse encoding of monotone cuts. The monotone cuts (9b), (23), (24), and (25) are active when a collection of componentwise lower-bounding inequalities on integer vectors is respected. For a decision vector $\nu \in \mathbb{Z}_+^m$ and a fixed solution $\nu' \in \mathbb{Z}_+^m$, the condition $1[\nu \leq \nu'] = 0$ holds if and only if the disjunction $\vee_{l=1}^m (\nu_l \geq \nu_l' + 1)$ is satisfied or, equivalently, if $\sum_{l=1}^m 1[\nu_l \geq \nu_l' + 1] \geq 1$. Adding such disjunctive constraints to the relaxed master problem requires keeping track of each clause using a binary variable (Balas 1979). In our implementation, for each period $p \in \mathcal{P}$ and each component x_{pl} of the vector x_p, we maintain a pool $\mathcal{A}_{pl}$ of indicators $a_{plk} \in \{0, 1\}$, each bounded by a constraint $(k+1) a_{plk} \leq x_{pl}$ enforcing that $a_{plk} \leq 1[x_{pl} \geq k+1]$. Initially, each pool $\mathcal{A}_{pl}$ is empty, and a_{plk} is generated when $x_{pl}' = k$ appears as the threshold value in a monotone cut. A feasibility cut (25) is then reformulated as $\sum_{l=1}^n a_{pl(x_{pl}')} \geq 1$, and an optimality cut (9b) becomes $\Theta_p \geq \mathcal{Q}_p(x_p')(1 - \sum_{l=1}^n a_{pl(x_{pl}')})$. Similarly, a pool of binary variables is maintained for each sum of decision variables appearing in the feasibility cuts (23) and (24).

In contrast with standard implementations (e.g., Hooker and Osorio 1999, Liu et al. 2024), where a new collection of n binary variables is generated each time a monotone cut is added to the master problem, we allow the indicators a_{plk} to appear in all the cuts where the same threshold on x_{pl} is encountered. In practice, this makes a significant difference in the number of variables added to the master problem. Indeed, monotone cuts are mostly needed in late iterations of the algorithm, where near-optimal solutions that only differ in a few components are sequentially visited and bounded by monotone optimality cuts. By reusing previously generated binary indicators, such cuts can be generated by introducing a small fraction of the indicators required by the standard approach.

4.2.5 Outline of the algorithm

We implement our algorithm in a classical Benders decomposition fashion. The initial relaxed master problem is equipped with the disaggregated auxiliary variables of Section 4.2.1 and the relaxed operational model of Section 4.2.2. Each iteration comprises three phases. In the first phase, the current master problem is solved, and the global lower bound is updated. In the second phase, Benders cuts are generated at the current master solution by applying the closest cut selection method from Section 4.2.3 to the LP relaxations of the operational problem (3) and of the single-route relaxations (10). The third phase is entered

if the LP relaxation of each operational problem is feasible for the current strategic solution. It consists of solving the operational problems with integrality constraints, generating the monotone cuts of Section 4.2.4, and updating the incumbent solution. The algorithm iterates until the desired tolerance is reached. The pseudocode and further discussion on implementation are provided in Appendix A.5.

4.3 Arc selection algorithm

Exact and heuristic algorithms serve complementary roles in this work. Our logic-based Benders decomposition algorithm is designed to expand the frontier of instances for which provably optimal solutions can be obtained. Notably, we can achieve optimality for instances of practical interest to transit authorities, who typically plan electrification projects in phases by considering a subset of routes at a time. However, as the planning scope broadens to long-term, citywide electrification, even finding a feasible solution becomes challenging, and heuristic methods become essential.

We propose a simple approach that exploits the structure of the depot BEB schedules as circulations on dense graphs. Since flow variables represent the vast majority of the decision variables in the extensive formulation of the BFEP, sparsifying these graphs can significantly reduce the size of the problem. Our method identifies a restricted set of relevant arcs by considering the active variables in the optimal solutions of a collection of single-route problems and LP relaxations of the BFEP. The algorithm solves a sequence of restricted problems where these arcs are progressively introduced. Finally, the original extensive formulation is solved, subject to a time limit. This last step provides a global optimality gap and makes it possible to use the arc selection algorithm as an exact method. At each step, the incumbent solution is used to warm-start the next problem.

The arc section algorithm comprises four phases. In the first phase, three sets of arcs associated with the flow variables $\{(w_r^p, v_r^p, z_r^p)\}_{(p,r)\in\mathcal{P}\times\mathcal{R}}$ of the depot BEB scheduling model are constructed. The first two sets $\mathcal{E}^{\mathrm{LP1}}$ and $\mathcal{E}^{\mathrm{LP2}}$ correspond to the active arcs from the LP relaxation of the extensive formulation of the problem, solved without and with additional constraints on early charging. To construct the third set $\mathcal{E}^{\mathrm{SR}}$, we solve for each period $p \in \mathcal{P}$, each route $r \in \mathcal{R}$, each type of depot BEBs $b \in \mathcal{B}$, and each fleet size $m \in \{0, 1, \ldots, \max_{t\in\mathcal{T}} d_r^{pt} - 1\}$ of on-route BEBs and conventional buses a single-route scheduling problem in which the charging capacity constraints (3b) are relaxed (i.e., an infinite number of chargers is assumed to be available at each location). Each arc that is active in at least one optimal solution is included in $\mathcal{E}^{\mathrm{SR}}$. In the second and third phases, we solve the extensive formulation of the BFEP associated with the edge set $\mathcal{E}^{\mathrm{LP1}} \cup \mathcal{E}^{\mathrm{LP2}}$, and then with the edge set $\mathcal{E}^{\mathrm{LP1}} \cup \mathcal{E}^{\mathrm{LP2}} \cup \mathcal{E}^{\mathrm{SR}}$. Finally, in the fourth phase, we solve the warm-started extensive formulation without restrictions. The pseudocode of algorithm AS is provided in Appendix A.6.

5 Computational experiments

This computational study evaluates: (i) the impact of our acceleration techniques on the performance of the logic-based Benders decomposition framework; (ii) the ability of our exact algorithm to solve partial network electrification planning instances to optimality; and (iii) the scalability of our heuristic algorithm on citywide electrification planning instances. Section 5.1 studies the acceleration techniques. Sections 5.2 and 5.3 present results for the partial and complete bus networks. Finally, in Section 5.4, we illustrate our model on the Chicago transit system. The experiments were conducted with a Python implementation and

Gurobi 10.0.3 on a computing cluster node equipped with 48 Intel Xeon Platinum 8260 cores @ 2.40 GHz and 192 GB of RAM. The relative optimality tolerance of the algorithms was set to 0.01%.

Our instances are constructed from historical data made available by transit agencies on the Mobility Database. We selected eight US cities, for which we built our instances based on the routes ($\mathcal{R}$), terminals ($\mathcal{J}$), and depots ($\mathcal{I}$) currently in use. We defined the minimum service requirement parameters on each route as the average number of buses in service during each hour ($T=24$) of a typical weekday, and assumed that this level would remain constant throughout the planning horizon ($d^p = d^{p'} \ \forall p, p' \in \mathcal{P}$). More details on the parameters used for our tests are given in Appendix A.7.

5.1 Acceleration techniques

We study the effect of each of our acceleration techniques on the overall performance of our logic-based Benders decomposition framework. To do so, we evaluate the individual impact of replacing the closest cut selection method (CC) of Section 4.2.3 by standard Benders cuts, and of removing the preprocessing (PP) of Section 4.1, the disaggregation method (DA) of Section 4.2.1, the partial decomposition (PD) of Section 4.2.2, and the improved monotone cuts and encoding techniques (MC) of Section 4.2.4. Finally, we consider using none of these acceleration techniques. Except for minor practical enhancements, such as the early-stopping condition applied to our master problem iterations (see Appendix A.5), the unaccelerated version of our logic-based Benders decomposition exactly corresponds to algorithm GBD2^{dis} of Liu et al. (2024), which we take as a baseline.

We consider for each problem size $(|\mathcal{R}|, |\mathcal{P}|) \in \{(3,4), (6,6), (9,6)\}$ a set of 10 instances composed of disjoint subsets of routes from the Atlanta network. We impose a time limit of four hours per instance for each method, with the default optimality tolerance of 0.01%. Table 2 presents average results for the computing time, in seconds, the number of iterations (Iter), the number of Benders cuts (BCuts) and monotone cuts (MCuts) added to the model, and the number of binary indicators (Ind) needed to encode the latter. These metrics are reported as geometric means to better reflect central tendencies. We present the percentage of computing time spent on the master problem (MP), the linear cut generation subproblems (LP), and the integer monotone cut generation subproblems (IP). Finally, the number of instances solved to optimality, and the average optimality gap, taken as 0 for instances that terminated before the time limit, are reported. Instances that reached the time limit contribute to the overall results using the figures recorded at termination.

The results of Table 2 show that our acceleration techniques drastically improve the performance of the logic-based Benders decomposition framework. Indeed, they allow solving 29 instances to optimality within the time limit, compared to 11 for the baseline method. Moreover, for challenging instances, they reduce the average computing time and optimality gap by three orders of magnitude.

Our ablation study reveals that our acceleration techniques act in complementarity to improve the overall performance of our logic-based Benders decomposition.

- (PP): The preprocessing step divides the average solving time and the number of iterations by a factor of two across all the groups of instances.

- (DA): Disaggregating the operational model based on single-route relaxations is our most impactful acceleration technique. It slightly increases the number of generated Benders cuts for smaller instances. However, in the last group, excluding this technique reduces from nine to two the number of solved instances and increases the average computing time by two orders of magnitude.

• (PD): Including a relaxation of the operational model in the master problem formulation is our most efficient acceleration technique for the smallest instances, where it divides by two the number of iterations and by three the number of generated Benders cuts. This technique always remains beneficial, but is less impactful for large instances.

• (CC): Using closest cuts systematically reduces the number of generated cuts. For the last group, it divides the solving time and the number of iterations by almost six and three, respectively.

• (MC): This set of acceleration techniques is the one whose impact on performance is the most subtle. Since the LP relaxation of the operational problem is rather tight and our algorithm relies in priority on classical Benders cuts, the number of monotone cuts required to attain provable optimality is usually very small. Consequently, although the sparse encoding makes a noticeable difference in the number of indicators, the size of the master problem remains manageable even with a naive encoding, and the impact on the algorithm's overall performance is limited. For the same reason, monotone feasibility cuts are rarely needed, and our infeasibility detection tests are rarely executed. In the last group, a single instance stands out as an exception. This instance is solved to optimality by the variants All, All−DA, and All−PD, which all generate two single-route fleet assignment cuts (23). In contrast, it cannot be solved within the time limit with the variant All−MC, which had instead generated five weaker generic infeasibility cuts (25) at termination.

Table 2: Performance of LBBD with different acceleration techniques

Instances ($|\mathcal{R}|, |\mathcal{P}|$)	Accelerations	Summary				Cuts			Time (%)		
		Opt	Gap	Time	Iter	BCuts	MCuts	Ind	MP	LP	IP
(3,4)	All	**10**	**0.0000**	11.1	**5.9**	33.1	**1.7**	**3.1**	10.5	44.7	10.2
	All − PP	**10**	**0.0000**	17.2	10.3	57.3	2.1	3.4	22.8	52.3	10.6
	All − DA	**10**	**0.0000**	11.8	8.4	**19.2**	1.9	3.3	15.8	37.1	12.2
	All − PD	**10**	**0.0000**	18.6	11.4	90.4	1.9	3.3	7.8	65.9	8.0
	All − CC	**10**	**0.0000**	12.1	9.4	58.2	2.1	4.5	29.7	19.6	21.0
	All − MC	**10**	**0.0000**	**10.9**	6.1	33.5	1.9	6.2	11.5	45.0	7.9
	None	**10**	**0.0000**	174.3	65.0	219.3	2.4	27.2	58.3	24.8	15.7
(6,6)	All	**10**	**0.0000**	**56.9**	**10.3**	113.3	1.9	4.9	29.7	38.3	13.3
	All − PP	9	0.0010	133.0	22.5	164.6	2.6	5.6	39.0	35.8	16.3
	All − DA	**10**	**0.0000**	146.8	26.8	**101.8**	2.2	6.6	57.6	28.3	5.6
	All − PD	**10**	**0.0000**	69.8	14.2	241.5	**1.6**	**3.2**	19.9	58.2	8.7
	All − CC	**10**	**0.0000**	108.1	21.0	230.1	2.4	7.2	63.1	13.5	14.6
	All − MC	**10**	**0.0000**	57.9	10.8	114.6	2.1	17.6	31.7	38.3	11.3
	None	1	0.7932	13747.4	209.3	1141.6	2.8	90.2	95.7	2.6	1.5
(9,6)	All	**9**	**0.0037**	**372.8**	**24.0**	**247.9**	3.5	13.1	54.5	29.2	7.7
	All − PP	8	0.0095	659.4	33.2	317.7	5.9	**10.4**	62.4	26.2	7.5
	All − DA	2	0.0498	10306.1	80.3	308.2	4.1	27.5	95.7	2.9	0.9
	All − PD	**9**	0.0038	444.9	26.8	422.0	3.9	13.0	45.1	39.4	8.3
	All − CC	6	0.0334	2137.6	61.3	581.8	5.5	35.5	82.1	6.3	9.8
	All − MC	8	0.0084	398.8	24.9	251.8	3.6	36.4	53.2	29.2	8.8
	None	0	2.6173	14400.0	198.9	1190.7	**1.1**	13.4	91.9	4.8	3.2

5.2 Partial network instances

In this section, we evaluate the ability of our accelerated logic-based Benders decomposition (LBBD) algorithm to solve instances of moderate size to optimality. We compare LBBD to the extensive formulation (EX) and arc selection (AS) approaches. Our benchmark set is generated from the networks of Atlanta, Boston, Chicago, and Dallas. For each city, we consider the electrification of a subset of routes over $|\mathcal{P}| \in \{2, 4, 6, 8, 10\}$ years, with and without on-route chargers and BEBs. We take $|\mathcal{R}| \in \{3, 6, 9, 12, 15\}$ routes when on-route BEBs are allowed, and $|\mathcal{R}| \in \{2, 4, 6, 6, 10\}$ otherwise, for a total of 200 instances. The time limit is set to two hours per instance, and the preprocessing steps of Section 4.1 are applied to all methods. For AS, 75% of the time budget is allocated to the heuristic phase, with 75% of this share being allocated to the second restricted IP. The remaining computing time is used to solve the warm-started extensive formulation.

Figure 1 shows the cumulative number of instances solved to optimality as a function of time and a box plot of the final optimality gaps for each method. Table 3 separates the results by availability of on-route BEBs and number of routes. We report the number of solved instances, the average optimality gap, and the count of instances on which each algorithm provides the best performance.

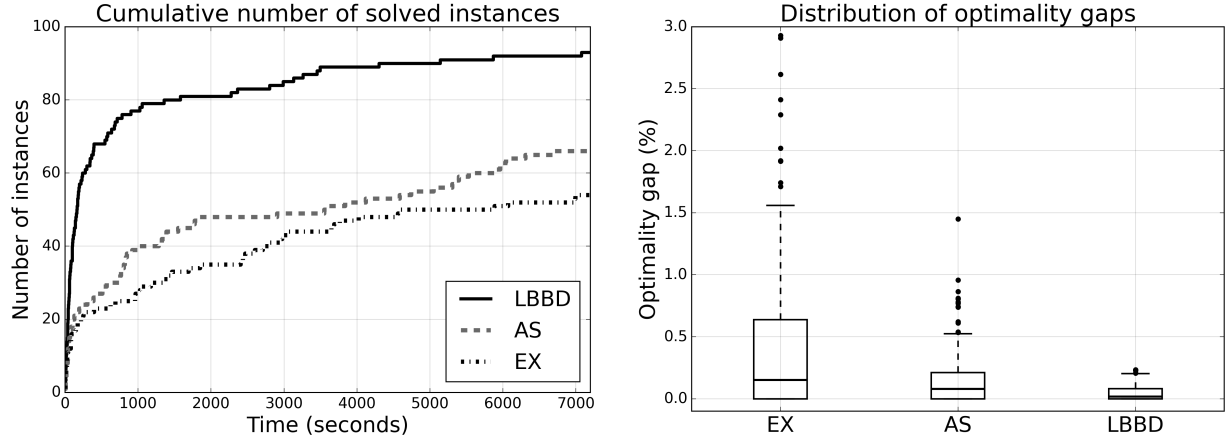Figure 1: Solved instances and optimality gaps for partial networks

Table 3: Results per group of instances for partial networks

Instances with on-route BEBs

$|\mathcal{R}|$	EX			AS			LBBD		
	Opt	Gap	Best	Opt	Gap	Best	Opt	Gap	Best
3	18	0.01	3	18	0.01	0	**19**	**0.00**	**17**
6	13	0.18	1	13	0.09	0	**17**	**0.00**	**19**
9	2	0.29	0	8	0.09	1	**13**	**0.01**	**19**
12	1	0.42	0	3	0.19	1	**7**	**0.04**	**19**
15	1	0.46*	0	**4**	0.22	4	**4**	**0.06**	**16**
Tot/avg	35	0.27	4	46	0.12	6	**60**	**0.02**	**90**

* excludes an instance for which no solution was found

Instances without on-route BEBs

$|\mathcal{R}|$	EX			AS			LBBD		
	Opt	Gap	Best	Opt	Gap	Best	Opt	Gap	Best
2	12	0.04	9	11	0.04	0	**17**	**0.01**	**11**
4	5	0.33	2	6	0.24	0	**11**	**0.03**	**18**
6	1	0.56	0	2	0.20	1	**5**	**0.06**	**19**
8	**1**	0.89	0	**1**	0.25	4	**1**	**0.10**	**16**
10	0	1.12	0	0	0.24	5	0	**0.15**	**15**
Tot/avg	19	0.59	11	20	0.19	10	**34**	**0.07**	**79**

The results show that LBBD significantly outperforms the other methods. It solves 94 instances, compared to 54 for EX and 66 for AS, requires only 2.5% and 5.3% of the time limit to match the number of instances solved by the other two methods, and achieves the best performance in 169 out of 200 instances. The largest optimality gap obtained with LBBD is 0.23%, whereas 50 instances exhibit an optimality gap of 0.64% or higher (up to 2.93%) with EX, and of 0.21% or higher (up to 1.45%) with AS. AS maintains the most stable performance as the number of routes increases, particularly when on-route BEBs are prohibited. In this case, meeting electrification targets requires larger fleets of depot BEBs, and the scheduling simplifications performed in the first heuristic phase greatly reduce the computational effort needed to identify good-quality solutions. Nevertheless, AS provides limited improvements over EX in the number of instances solved to optimality.

5.3 Complete network instances

In this section, we evaluate the ability of our heuristic method to scale to long-term, large-scale electrification planning problems. We construct our instances from the complete bus networks of eight major US cities, with a transition horizon of $|\mathcal{P}| = 10$ years. For comparison purposes, we also consider a simple policy restriction (PR) heuristic, where the arc selection phase of algorithm AS is replaced by the most efficient a priori arc elimination rules we identified in preliminary experiments. In this approach, the depot BEBs are only allowed to charge starting from the fully depleted battery state (i.e., $z^{pt}_{rbiks} = 0$ if $s > 0$), and cannot idle during service times, except in the fully charged state (i.e., $v^{pt}_{rbs} = 0$ if $s < s_b$ and $d^{pt}_r > 0$). The extensive formulation is first solved with these restrictions, and the incumbent solution obtained from the restricted model is then used to warm-start the original formulation. This approach is intended to serve as a fair baseline method against AS. Table 4 reports lower and upper bounds, in millions of dollars, and optimality gaps, in percentage, for EX, PR, and AS, on instances including and excluding on-route BEBs.

Table 4: Results for complete networks

| City | With on-route BEBs | | | | | | Without on-route BEBs | | | | | |
| | EX | | PR | | AS | | EX | | PR | | AS | |
	UB	Gap	UB	Gap	UB	Gap	UB	Gap	UB	Gap	UB	Gap
Atlanta	–	–	964.36	0.63	**962.32**	**0.42**	–	–	1103.09	6.45	**1045.61**	**1.30**
Boston	–	–	2384.71	∗	**2379.87**	**0.81**	∗	∗	2561.54	7.15	**2397.25**	**0.78**
Chicago	–	–	2734.13	0.83	**2727.72**	**0.59**	–	–	2890.68	5.41	**2744.84**	**0.39**
Dallas	–	–	1205.69	0.85	**1200.61**	**0.44**	–	–	1319.92	4.86	**1262.80**	**0.56**
Detroit	439.06	0.83	439.23	0.60	**438.46**	**0.43**	–	–	459.80	2.24	**451.84**	**0.51**
Houston	1972.56	17.51	1644.14	1.03	**1630.17**	**0.18**	–	–	1763.37	6.23	**1659.81**	**0.38**
Las Vegas	–	–	634.02	0.09	**634.01**	**0.08**	–	–	690.75	1.01	**686.69**	**0.43**
Los Angeles	4078.27	18.75	3344.34	0.95	**3325.93**	**0.40**	–	–	3556.79	6.04	**3361.52**	**0.59**

− no feasible solution was found within the eight-hour time limit, ∗ out-of-memory error

The results indicate that sparsifying the depot BEBs scheduling graphs makes it possible to achieve good solutions for otherwise intractable instances. AS systematically dominates, most noticeably when on-route BEBs are unavailable. Since both PR and AS limit the efficiency of the depot BEBs by restricting the flexibility of their schedules but do not affect on-route BEBs, it was expected that they would provide

solutions of lower quality in the second group of instances. The relatively stable optimality gaps of AS, whether on-route BEBs are available or not, confirm that the active arcs from the LP relaxations and integer subproblems considered in our method suffice to construct schedules of high quality. This is not the case for a priori restriction rules.

Interestingly, the lower bound Gurobi achieves at the root node tends to become better as instance size increases, although it usually stops making progress once the solver starts branching on large instances. In Benders decomposition, the initial master problem shows the opposite trend, and solving each cut separation problem (20) can take several minutes. As a consequence, AS can achieve optimality gaps that are not much higher for complete networks than for the partial networks of the previous section, whereas our exact algorithm would not be competitive for complete network instances. We conclude that citywide instances are best approached with primal heuristics, and that relying on a lower bound from the extensive formulation of the problem is sufficient to obtain very good optimality gaps in this context. Further discussion, along with detailed results for each phase of algorithms AS and PR, is provided in Appendix A.9.

5.4   Illustrative example on the Chicago network

We conclude our computational study by discussing high-level managerial insights and illustrating the properties of our model. We consider the best solution identified in the previous section for the complete network instance of Chicago with on-route BEBs. Figure 2 presents the number of buses and chargers in the system in each year of the planning horizon, as well as the yearly investment and operational costs, and Figure 3 summarizes the daily service level of each type of bus in each period. A visualization of the network of chargers throughout the transition is provided in Figure 4.

Figure 2: Planning solution and costs summary – Chicago
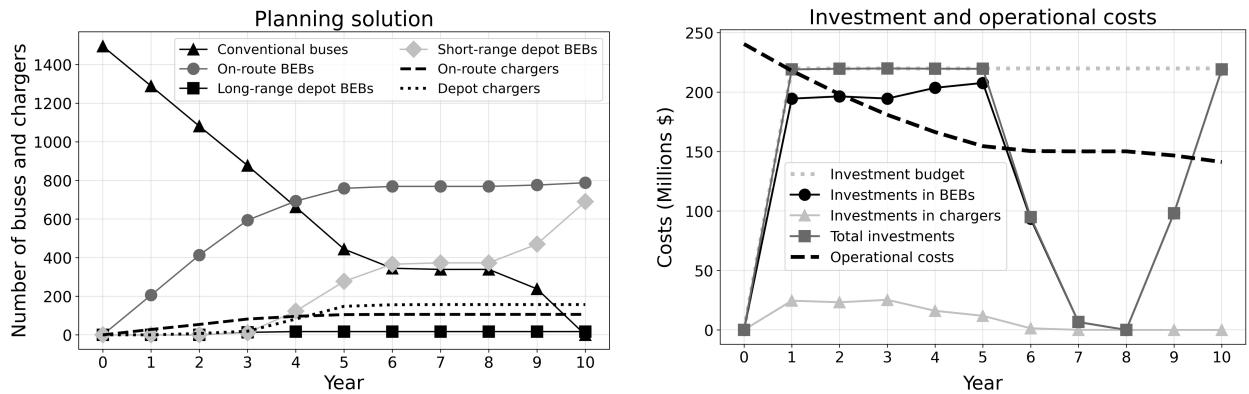
Assuming that the objective coefficients and service level requirements are constant across all investment periods and electrification targets are imposed only in the last planning period, it can be shown (see Appendix A.10) that an optimal solution starts with a phase of investments that yield a least $(1-\gamma)\%$ in yearly operational savings, followed by a phase without investments, and a last phase of investments that provide at most $(1-\gamma)\%$ in yearly operational savings. Since this instance satisfies these assumptions, it is indeed what we observe. The yearly budget is almost fully used from years 1 to 5. In this phase, the return on investment (ROI), defined for period p as $\left(O_p(x_p) - O_p(x_{p-1})\right)/I_p(x_p - x_{p-1})$ ranges from 10.26% in year 1 to 5.38% in year 5. Year 6 marks the end of the first phase, with an ROI of 4.36%, slightly above the threshold of

(1-γ)=4%, on \$94.8 million in new investments. We also observe nonzero investments in year 7, with an ROI of 4.3%, highlighting that this solution is slightly suboptimal. Indeed, advancing these investments to year 6 and adjusting the operations accordingly would improve the value of the solution by \$16.5 thousand (< 0.001% of the overall costs). After a phase where no additional investments are made, we observe a last phase of investments whose purpose is to complete the electrification of the fleet by the end of the planning horizon. These last investments fall below the rentability threshold (ROI of 3.57% and 2.5% for years 9 and 10, respectively), and are thus delayed as much as possible.

Due to the cheaper operating costs of BEBs compared to conventional buses (and following Proposition 7), the operational costs decrease monotonically throughout the transition. The initial annual operational costs of \$240.4 million reach \$150.4 million at the end of the first phase of investments, and \$141.2 million in year 10. Overall, the initial fleet of 1495 conventional buses is replaced by a fleet of the same size consisting of 788 on-route BEBs, 690 short-range depot BEBs, and 17 long-range depot BEBs, supported by 106 on-route chargers and 157 depot chargers, for investment costs totaling \$1.52 billion, and operational costs of \$1.66 billion over 10 years.

Figure 3: Number of buses in service per hour by year – Chicago
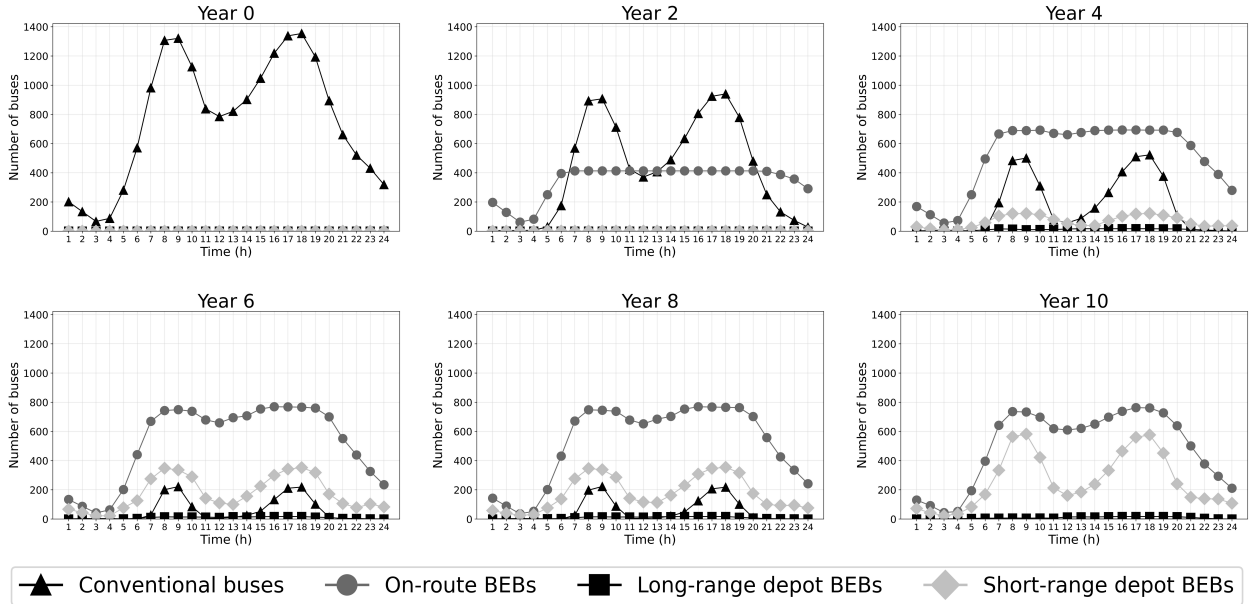
In the first three years, all the investments are made in on-route BEBs and fast chargers. Depot BEBs are progressively introduced starting from year 4, and represent the quasi-totality of the BEBs acquired in the last phase of investments. This can be explained by the usage profile presented in Figure 3. For example, in year 2, the 413 on-route BEBs are in service 84.2% of the time, and are all in service for 14 consecutive hours starting from 5 am. In comparison, depot BEBs can be in service at most 66.7% of the time, since we assumed they require 1 hour of charging for every 2 hours of operation. The hourly operating costs of conventional buses being by far the highest, on-route BEBs are thus preferable to depot BEBs as long as conventional buses are still needed to satisfy the base load. After that, short-range depot BEBs, which are the cheapest to acquire and operate, become the most cost-efficient option. They are mostly used to satisfy the service requirement in peak hours, and recharge around noon and at night.

Figure 4: Chargers deployment – Chicago
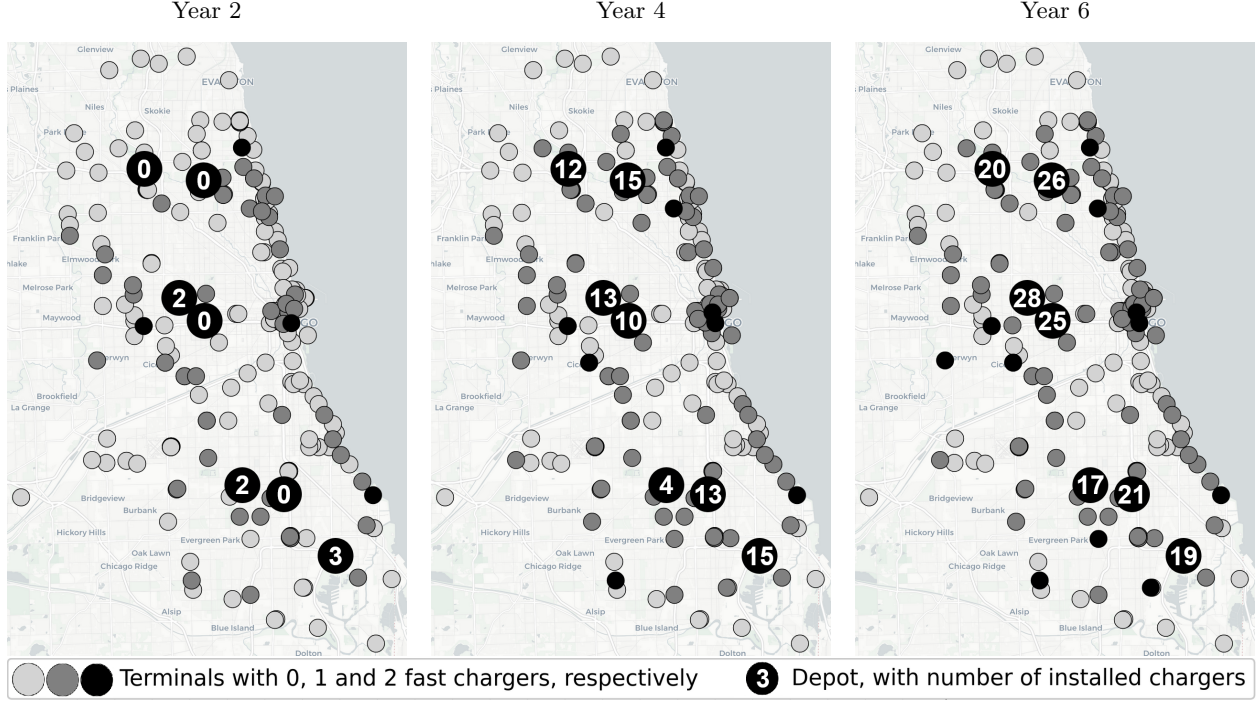
Regarding charger locations, Figure 4 shows that fast chargers are principally deployed in densely connected areas. Terminals located on the city outskirts are never equipped with fast chargers, meaning that the routes outside of the city center are mostly serviced by depot BEBs. Interestingly, no chargers are installed after the first investment phase. Indeed, although more chargers could reduce the operational costs by increasing the scheduling flexibility of the depot BEBs introduced in years 9 and 10, these savings in the final years do not suffice to offset additional capital investment. If operational expenses were taken into account beyond the 10-year planning horizon, the resulting extended amortization period might justify further investments in the last phase.

6 Conclusion

This paper presents a flexible and scalable framework for bus fleet electrification planning, jointly optimizing strategic and operational decisions. The proposed model yields a two-stage integer program with separable subproblems whose optimal values are monotone in the first-stage variables. Building on a logic-based Benders decomposition framework recently proposed for this class of problems, we introduce a suite of acceleration techniques that achieve speedups of up to three orders of magnitude, enabling the solution of practically relevant instances to proven optimality. Additionally, we develop an easily implementable restriction heuristic tailored to very large-scale problems. Experiments on real data demonstrate optimality gaps below 1% can be achieved on citywide electrification problems, providing valuable decision support for transit agencies.

We made simplifying assumptions throughout the paper. At the operational level, we considered a non-preemptive charging policy and assumed that each bus type uses a single charging technology. Importantly, these assumptions are not fundamental to our algorithms. Their relaxation would preserve the monotonicity property leveraged by our logic-based Benders decomposition, as well as the circulation structure of the schedules exploited by the arc selection algorithm. Also, inherently uncertain parameters such as future energy costs, driver wages, and vehicle and charger characteristics were treated as deterministic. As a possible extension, uncertainty in operational parameters could be modeled by scenarios, and uncertainty in strategic parameters could be accounted for by jointly optimizing over different trajectories for future technology. These generalizations of the model would result in additional separable subproblems that could be efficiently handled by our decomposition framework. Crucially, long-term planning models should in practice be applied in a rolling horizon fashion to guide short-term decisions, with parameters updated as new information becomes available and careful sensitivity analyses performed at each stage of deployment.

A promising avenue for future research in integrated bus fleet electrification planning is the modeling of interactions between charging infrastructure and the power grid. At the strategic level, this includes incorporating grid capacity expansion decisions to accommodate increased loads from chargers. At the operational level, integrating vehicle-to-grid scheduling would enable transit agencies to respond to daily price fluctuations and participate actively in electricity markets. From an algorithmic perspective, embedding advanced solution methods for E-VSP within an accelerated logic-based Benders decomposition could facilitate the integration of detailed scheduling and routing models into long-term planning problems or, conversely, enable key strategic decisions to be incorporated into E-VSP formulations.

Acknowledgement

References

Balas E (1979) Disjunctive programming. *Annals of discrete mathematics* 5:3–51.

C40 (2023) Green and healthy streets accelerator report. `https://www.c40.org/wp-content/uploads/2024/03/C40_Green__Healthy_Streets_Accelerator_Progress_Report_2023.pdf`. accessed: February 2025.

CARB (2018) California transitioning to all-electric public bus fleet by 2040. URL `https://ww2.arb.ca.gov/news/california-transitioning-all-electric-public-bus-fleet-2040`, accessed: May 2025.

Conforti M, Wolsey LA (2019) "Facet" separation with one linear program. *Mathematical Programming* 178(1):361–380.

Cornuéjols G, Lemaréchal C (2006) A convex-analysis perspective on disjunctive cuts. *Mathematical Programming* 106(3):567–586.

Crainic TG, Hewitt M, Maggioni F, Rei W (2021) Partial Benders decomposition: general methodology and application to stochastic network design. *Transportation Science* 55(2):414–435.

CTA (2022) Bus electrification plan and near-term bus purchases. `https://www.transitchicago.com/assets/1/6/Mar2022_-_CTA_E-Bus_Board_Presentation.pdf`. accessed: February 2025.

Davidson N (2023) How many electric buses does your city have? (2023 edition). `https://www.govtech.com/biz/data/how-many-electric-buses-does-your-city-have-2023-edition`. accessed: February 2025.

de Vos MH, van Lieshout RN, Dollevoet T (2024) Electric vehicle scheduling in public transit with capacitated charging stations. *Transportation Science* 58(2):279–294.

Dirks N, Schiffer M, Walther G (2022) On the integration of battery electric buses into urban bus networks. *Transportation Research Part C: Emerging Technologies* 139:103628.

Fischetti M, Salvagnin D, Zanette A (2010) A note on the selection of Benders' cuts. *Mathematical Programming* 124:175–182.

Gairola P, Nezamuddin N (2023) Optimization framework for integrated battery electric bus planning and charging scheduling. *Transportation Research Part D: Transport and Environment* 118:103697.

Gendron B, Scutellà MG, Garroppo RG, Nencioni G, Tavanti L (2016) A branch-and-Benders-cut method for nonlinear power design in green wireless local area networks. *European Journal of Operational Research* 255(1):151–162.

Gleeson J, Ryan J (1990) Identifying minimally infeasible subsystems of inequalities. *ORSA Journal on Computing* 2(1):61–63.

Hartmanis J (1982) Computers and intractability: a guide to the theory of np-completeness (michael r. garey and david s. johnson). *Siam Review* 24(1):90.

He Y, Liu Z, Song Z (2023a) Joint optimization of electric bus charging infrastructure, vehicle scheduling, and charging management. *Transportation Research Part D: Transport and Environment* 117:103653.

He Y, Liu Z, Zhang Y, Song Z (2023b) Time-dependent electric bus and charging station deployment problem. *Energy* 282:128227.

Hooker JN, Osorio MA (1999) Mixed logical-linear programming. *Discrete Applied Mathematics* 96:395–442.

Hooker JN, Ottosson G (2003) Logic-based Benders decomposition. *Mathematical Programming* 96(1):33–60.

Hosseini M, Turner J (2024) Deepest cuts for Benders decomposition. *Operations Research* .

Hu H, Du B, Liu W, Perez P (2022) A joint optimisation model for charger locating and electric bus charging scheduling considering opportunity fast charging and uncertainties. *Transportation Research Part C: Emerging Technologies* 141:103732.

Liu B, Bissuel C, Courtot F, Gicquel C, Quadri D (2024) A generalized Benders decomposition approach for the optimal design of a local multi-energy system. *European Journal of Operational Research* 318(1):43–54.

Liu X, Qu X, Ma X (2021) Optimizing electric bus charging infrastructure considering power matching and seasonality. *Transportation Research Part D: Transport and Environment* 100:103057.

Magnanti TL, Wong RT (1981) Accelerating Benders decomposition: Algorithmic enhancement and model selection criteria. *Operations research* 29(3):464–484.

MARTA (2023) Electric bus fleet charging infrastructure. URL `https://itsmarta.com/uploadedFiles/More/About_MARTA/Bus%20Fleet%20Transition%20and%20Infrastructure%20Plan%20RAC.pdf`, accessed: February 2025.

MBTA (2022) Bus electrification plan. `https://cdn.mbta.com/sites/default/files/2023-07/2022-05-bus-electrification-plan_0.pdf`. accessed: February 2025.

MTA (2024a) Metropolitan Transportation Authority bus transformation update part 1: Fleet and facility update. `https://cdn.mbta.com/sites/default/files/2021-04/2021-04-26-fmcb-bus-transformation-update.pdf`. accessed: February 2025.

MTA (2024b) Metropolitan Transportation Authority zero-emission transition plan. `https://www.mta.info/document/138261`. accessed: February 2025.

Papadakos N (2008) Practical enhancements to the Magnanti–Wong method. *Operations Research Letters* 36(4):444–449.

Parmentier A, Martinelli R, Vidal T (2023) Electric vehicle fleets: Scalable route and recharge scheduling through column generation. *Transportation Science* 57(3):631–646.

Pelletier S, Jabali O, Mendoza JE, Laporte G (2019) The electric bus fleet transition problem. *Transportation Research Part C: Emerging Technologies* 109:174–193.

Perumal SS, Lusby RM, Larsen J (2022) Electric bus planning & scheduling: A review of related problems and methodologies. *European Journal of Operational Research* 301(2):395–413.

Rogge M, Van der Hurk E, Larsen A, Sauer DU (2018) Electric bus fleet size and mix problem with optimization of charging infrastructure. *Applied Energy* 211:282–295.

Schrijver A (2003) *Combinatorial optimization: polyhedra and efficiency*, volume A (Springer).

Seo K, Joung S, Lee C, Park S (2022) A closest Benders cut selection scheme for accelerating the Benders decomposition algorithm. *INFORMS Journal on Computing* 34(5):2804–2827.

Wang Y, Liao F, Lu C (2022) Integrated optimization of charger deployment and fleet scheduling for battery electric buses. *Transportation Research Part D: Transport and Environment* 109:103382.

Xylia M, Leduc S, Patrizio P, Kraxner F, Silveira S (2017) Locating charging infrastructure for electric buses in Stockholm. *Transportation Research Part C: Emerging Technologies* 78:183–200.

Zhou Y, Wang H, Wang Y, Yu B, Tang T (2024) Charging facility planning and scheduling problems for battery electric bus systems: A comprehensive review. *Transportation Research Part E: Logistics and Transportation Review* 183:103463.

A Appendix

A.1 Notation

Table A.1: Indices and sets, strategic variables, operational variables, parameters, value functions

Notation	Description
$p \in \mathcal{P}$	Investment period (year); $\mathcal{P} = \{1, 2, \ldots, P\}$
$t \in \mathcal{T}$	Operational time interval (hour); $\mathcal{T} = \mathbb{Z}/T\mathbb{Z}$ the cyclic group of integers modulo T
$r \in \mathcal{R}$	Bus route
$b \in \mathcal{B}$	Type of depot BEB
$i \in \mathcal{I}$	Depot location
$j \in \mathcal{J}$	On-route terminal
$k \in \mathcal{K}$	Type of depot charger
$s \in \mathcal{S}_b$	Charging state for buses of type b; $\mathcal{S}_b = [0 : s_b]$; $\mathcal{S}_b^w = [1 : s_b]$; $\mathcal{S}_b^z = [0 : s_b-1]$
$\mathcal{R}(j)$	Set of routes connected to terminal j
$\mathcal{J}(r)$	Set of terminals connected to route r
x_p	Strategic variables of period p; $x_p = (\chi_p, \{\eta_{pr}\}_{r \in \mathcal{R}}) \in \mathbb{Z}_+^n$
χ_p	Chargers location decisions; $\chi_p = (\bar{\chi}^p, \widetilde{\chi}^p) \in \mathbb{Z}_+^{\mathcal{I} \times \mathcal{K}} \times \mathbb{Z}_+^{\mathcal{J}}$
η_{pr}	Fleet size decisions on route r; $\eta_{pr} = (\bar{\eta}_r^p, \widetilde{\eta}_r^p, \widehat{\eta}_r^p) \in \mathbb{Z}_+^{\mathcal{B}} \times \mathbb{Z}_+ \times \mathbb{Z}_+$
$\bar{\chi}_{ik}^p$	# of chargers of type k at depot i in period p
$\widetilde{\chi}_j^p$	# of fast on-route chargers at terminal j in period p
$\bar{\eta}_{rb}^p$	# of depot BEBs of type b assigned to route r in period p
$\widetilde{\eta}_r^p$	# of on-route BEBs assigned to route r in period p
$\widehat{\eta}_r^p$	# of conventional buses assigned to route r in period p
y_{pr}	Operational variables of period p on route r; $y_{pr} = (w_r^p, v_r^p, z_r^p, \widetilde{w}_r^p, \widehat{w}_r^p) \in \mathbb{Z}_+^{\mathcal{T} \times \mathcal{B} \times \mathcal{S}_b^w} \times \mathbb{Z}_+^{\mathcal{T} \times \mathcal{B} \times \mathcal{S}_b} \times \mathbb{Z}_+^{\mathcal{T} \times \mathcal{B} \times \mathcal{I} \times \mathcal{K} \times \mathcal{S}_b^z} \times \mathbb{Z}_+^{\mathcal{T} \times \mathcal{J}(r)} \times \mathbb{Z}_+^{\mathcal{T}}$
w_{rbs}^{pt}	# of depot BEBs of type b in service in state s on route r during interval t of period p
v_{rbs}^{pt}	# of depot BEBs of type b idling in state s on route r during interval t of period p
z_{rbiks}^{pt}	# of depot BEBs of type b assigned to route r starting to charge in state s at depot i on a type k charger in interval t of period p
$\widetilde{w}_r^{pt}$	# of on-route BEBs in service on route j in interval t of period p
$\widehat{w}_r^{pt}$	# of conventional buses in service on route j in interval t of period p
γ	Yearly discount factor
c_l^u	Coefficient (coefficient vector) of component(s) l of decision variables u
I_p^{UB}	Investment budget in period p
$\widehat{\eta}_p^{\mathrm{UB}}$	Maximum number of conventional buses allowed in period p
$\bar{\chi}_i^{\mathrm{UB}}$	Maximum number of chargers that can be installed in depot i
$\widetilde{\chi}_j^{\mathrm{UB}}$	Maximum number of chargers that can be installed at terminal j
d_r^{pt}	Demand (minimum number of buses in service) on route r in interval t of period p
ρ	# of on-route BEBs allowed to share the same charger during a service interval
κ_{rbiks}	# of time intervals needed to perform a round-trip between route r and depot i and fully charge a BEB of type b starting from state s using charger type k
$I_p(x_p - x_{p-1})$	Investment costs of period p
$O_p(x_p)$	Operational costs of period p; $O_p(x_p) = H_p(x_p) + \mathcal{Q}_p(x_p)$
$H_p(x_p)$	Fixed maintenance costs of period p (do not depend on operational decisions)
$\mathcal{Q}_p(x_p)$	Optimal variable operational costs of period p given x_p (also called scheduling costs)

A.2 Proofs

Proof of Proposition 1. Let $\mathcal{R}$ be a set of demand points and $\mathcal{J}$ be a set of candidate locations, with $\mathcal{J}(r) \subseteq \mathcal{J}$ the subset of candidate locations that cover the demand point $r \in \mathcal{R}$. The set covering problem (SCP) can be formulated as:

$$\min \sum_{j \in \mathcal{J}} u_j \tag{A.1a}$$

$$\text{s.t.} \sum_{j \in \mathcal{J}(r)} u_j \geq 1, \qquad\qquad \forall r \in \mathcal{R}, \tag{A.1b}$$

$$u_j \in \{0, 1\}, \qquad\qquad \forall j \in \mathcal{J}, \tag{A.1c}$$

where we assume without loss of generality that $|\mathcal{J}(r)| \geq 1 \ \forall r \in \mathcal{R}$, and the problem is thus feasible.

By associating each candidate location $j \in \mathcal{J}$ to a terminal location with capacity $\widetilde{\chi}_j^{\mathrm{UB}} = 1$ and each demand point $r \in \mathcal{R}$ to a route, with the same covering locations $\mathcal{J}(r) \subseteq \mathcal{J} \ \forall r \in \mathcal{R}$, the SCP reduces to the BFEP with $P = 1$ investment period ($\mathcal{P} = \{1\}$), $T = 1$ time interval ($\mathcal{T} = \{0\}$), no depot locations ($\mathcal{I} = \emptyset$) and depot BEBs ($\mathcal{B} = \emptyset$), a retirement target $\widehat{\eta}_p^{\mathrm{UB}} = 0$, a fast-charging parameter $\rho = |\mathcal{R}|$, unit lower bounds $d_r^{1,0} = 1 \ \forall r \in \mathcal{R}$ on service levels, trivial operational costs $O_1(x_1) = 0$ for any x_1, investments costs $I_1(x_1 - x_0) = \sum_{j \in \mathcal{J}} \widetilde{\chi}_j^1$ given by the number of installed on-route chargers, a nonbinding budget $I_1^{\mathrm{UB}} = |\mathcal{J}|$, and an initial solution x_0 in which no chargers are installed and only one conventional bus is assigned to each route in period 0. Omitting the variables that must be fixed to 0 and the constraints that are respected by construction of the problem, as well as the indices on period and time interval, this special case of the BFEP can be written as:

$$\min \sum_{j \in \mathcal{J}} \widetilde{\chi}_j \tag{A.2a}$$

$$\text{s.t.} \sum_{j \in \mathcal{J}(r)} \widetilde{w}_{rj} \geq 1, \qquad\qquad \forall r \in \mathcal{R}, \tag{A.2b}$$

$$\sum_{r \in \mathcal{R}(j)} \widetilde{w}_{rj} \leq |\mathcal{R}|\widetilde{\chi}_j, \qquad\qquad \forall j \in \mathcal{J}, \tag{A.2c}$$

$$\sum_{j \in \mathcal{J}(r)} \widetilde{w}_{rj} \leq \widetilde{\eta}_r, \qquad\qquad \forall r \in \mathcal{R}, \tag{A.2d}$$

$$\widetilde{\chi}_j \in \{0, 1\}, \qquad\qquad \forall j \in \mathcal{J}, \tag{A.2e}$$

$$\widetilde{\eta}_r \in \mathbb{Z}_+, \qquad\qquad \forall r \in \mathcal{R}, \tag{A.2f}$$

$$\widetilde{w}_{rj} \in \mathbb{Z}_+, \qquad\qquad \forall r \in \mathcal{R}, \ \forall j \in \mathcal{J}, \tag{A.2g}$$

where (A.2b), (A.2c), and (A.2d) are the service level, on-route charging capacity, and fleet size constraints, respectively.

For $\widetilde{\chi}$ fixed, the objective value of any feasible solution is $\sum_{j \in \mathcal{J}} \widetilde{\chi}_j$, and the restricted problem admits a feasible solution if and only if $\sum_{j \in \mathcal{J}(r)} \widetilde{\chi}_j \geq 1 \ \forall r \in \mathcal{R}$. Indeed, suppose that $\exists\, r \in \mathcal{R}$ such that $\sum_{j \in \mathcal{J}(r)} \widetilde{\chi}_j = 0$. Constraints (A.2c), then imply that $w_{rj} = 0 \ \forall j \in \mathcal{J}(r)$, so that the service level constraint (A.2b) on route r cannot be respected. Otherwise, we can build a feasible solution by setting $\widetilde{\eta}_r = 1 \ \forall r \in \mathcal{R}$ and, for each $r \in \mathcal{R}$, $\widetilde{w}_{rj} = 1$ for an arbitrary $j \in \{j' \in \mathcal{J}(r) : \widetilde{\chi}_{j'} = 1\}$ and $\widetilde{w}_{rj} = 0$ for all the other terminals. We conclude that the BFEP (A.2) can equivalently be expressed as the SCP (A.1), where the decision variable $\widetilde{\chi}_j$ replaces u_j for each $j \in \mathcal{J}$. $\qquad\square$

Proof of Proposition 2. The proof is again by reduction from the SCP (A.1). Let the set of depot locations $\mathcal{I}$ of the BFEP be equal to the set of candidate locations $\mathcal{J}$ of the SCP, and let each depot have unit capacity, i.e., $\bar{\chi}_i^{\mathrm{UB}} = 1 \ \forall i \in \mathcal{I}$. We consider $T = |\mathcal{R}| + 1$ time intervals and retirement target $\hat{\eta}_p^{\mathrm{UB}} = 0$. Each route $r \in \mathcal{R}$ only requires one bus in service in the interval $t = 0$, and none in other intervals. We and consider a single type of depot BEBs $\mathcal{B} = \{b\}$ with battery capacity $s_b = 1$, and a single type of chargers $\mathcal{K} = \{k\}$ that can recharge a depleted depot BEBs in one time interval, taking $\kappa_{rbik0} = 1$ for all $(r, i) \in \mathcal{R} \times \mathcal{I}$. The cost of charging at depot $i \in \mathcal{I}$ for a BEB operating on route r is set to zero if $i \in \mathcal{J}(r)$ and $|\mathcal{I}|$ otherwise. The other operational costs are taken as trivial, leading to the operational costs function $O_1(x_1) = \sum_{r \in \mathcal{R}} \sum_{i \in \mathcal{I} \setminus \mathcal{J}(r)} \sum_{t \in \mathcal{T}} z_{ri}^{1t}$. The investments costs $I_1(x_1 - x_0) = \sum_{i \in \mathcal{I}} \bar{\chi}_j^1$ are given by the number of installed depot chargers. We take a nonbinding investment budget $I_1^{\mathrm{UB}} = |\mathcal{J}|$, and an initial solution x_0 in which no chargers are installed and only one conventional bus is assigned to each route in period 0. Omitting the variables that must be fixed to 0 and the constraints that are trivially respected by construction of the problem, as well as the indices on period, time interval, type of depot BEB, and chargers, this special case of the BFEP can be written as problem (A.3), where (A.3b) are the service level constraints, (A.3c)-(A.3d) are the depot BEB dynamics constraints, (A.3e) are the depot charging capacity constraints, and (A.3f) are the fleet size constraints.

$$\min \sum_{i \in \mathcal{I}} \bar{\chi}_i + |\mathcal{I}| \sum_{r \in \mathcal{R}} \sum_{i \in \mathcal{I} \setminus \mathcal{J}(r)} \sum_{t \in \mathcal{T}} z_{ri}^t \tag{A.3a}$$

$$\text{s.t. } w_{r1}^0 \geq 1, \qquad \forall r \in \mathcal{R}, \tag{A.3b}$$

$$w_{r1}^t + v_{r1}^t = \sum_{i \in \mathcal{I}} z_{ri0}^{t-1} + v_{r1}^{t-1}, \qquad \forall t \in \mathcal{T}, \ r \in \mathcal{R}, \tag{A.3c}$$

$$\sum_{i \in \mathcal{I}} z_{ri0}^t + v_{r0}^t = w_{r1}^{t-1} + v_{r0}^{t-1}, \qquad \forall t \in \mathcal{T}, \ r \in \mathcal{R}, \tag{A.3d}$$

$$\sum_{r \in \mathcal{R}} z_{ri0}^t \leq \bar{\chi}_i, \qquad \forall t \in \mathcal{T}, \forall i \in \mathcal{I}, \tag{A.3e}$$

$$w_{r0}^0 + v_{r0}^0 + v_{r1}^0 + z_{ri0}^{T-1} \leq \bar{\eta}_r, \qquad \forall r \in \mathcal{R}, \tag{A.3f}$$

$$\bar{\chi}_i \in \{0, 1\}, \qquad \forall i \in \mathcal{I}, \tag{A.3g}$$

$$\bar{\eta}_r \in \mathbb{Z}_+, \qquad \forall r \in \mathcal{R}, \tag{A.3h}$$

$$w_{r1}^t, v_{r0}^t, v_{r1}^t, z_{ri0}^t \in \mathbb{Z}_+, \qquad \forall t \in \mathcal{T}, \ \forall r \in \mathcal{R}, \ \forall i \in \mathcal{I}. \tag{A.3i}$$

We consider two cases. First, consider a feasible solution to problem (A.3) in which there is some route $r \in \mathcal{R}$ for which none of the depots permitting free charging trips is equipped with a charger, i.e., $\bar{\chi}_i = 0$ $\forall i \in \mathcal{J}(r)$. For the service level constraint (A.3b) to hold for route r, we need $w_{r1}^0 \geq 1$. As a consequence of the flow balance constraints (A.3c)-(A.3d), this implies that at least one charging variable z_{ri}^t, $t \in \mathcal{T}$, $i \in \mathcal{I}$ takes a strictly positive value. By the charging capacity constraints (A.3e), at least one charging variable z_{ri}^t, $t \in \mathcal{T}$, $i \in \{i' \in \mathcal{I} : \bar{\chi}_i = 1\} \subseteq (\mathcal{I} \setminus \mathcal{J}(r))$ thus takes a strictly positive value. The objective value of such a solution is thus lower bounded by $\sum_{i \in \mathcal{I}} \bar{\chi}_i + |\mathcal{I}| \geq |\mathcal{I}| + 1$.

Second, consider a fixed $\bar{\chi}$ such that $\sum_{i \in \mathcal{J}(r)} \bar{\chi}_i \geq 1$ $\forall r \in \mathcal{R}$. We show that we can build a feasible solution that only relies on free charging trips and dominates any solution from the first case. We start by setting $\bar{\eta}_r = 1$ $\forall r \in \mathcal{R}$. Then, let us index the set of routes as $\mathcal{R} = \{1, 2, \ldots, |R|\}$ and the set of depot locations as $\mathcal{I} = \{1, 2, \ldots, |I|\}$, and assign to each route $r \in \mathcal{R}$ an arbitrary depot $i_r \in \{i \in \mathcal{J}(r) : \bar{\chi}_i = 1\}$. For each route $r \in \mathcal{R}$, we take as only positive operational variables $w_{r1}^0 = 1$, $v_{r0}^t = 1$ $\forall t \in \{1, \ldots, r-1\}$,

$z^t_{r i_r 0} = 1$ for $t = r$, and $v^t_{r1} = 1 \ \forall t \in \{r+1, \ldots, T-1\}$, which yields a schedule in which the BEB is in service during interval $t = 0$, recharges during interval $t = r$ using the charger of depot i_r, and idles during all the other time intervals. This schedule respects the charging dynamics constraints, uses a single vehicle, and meets the service level requirement for interval $t = 0$. Furthermore, since the BEB assigned to route r is only in charge during interval r at a depot location $i \in \mathcal{J}(r)$ equipped with a charger, each charger is used by at most one vehicle at the time, and the charging capacity constraints are thus also respected. The solution is thus feasible, with objective $\sum_{i \in \mathcal{I}} \bar{\chi}_i \leq |\mathcal{I}|$.

We conclude that any optimal solution to problem (A.3) belongs to the second case. Using our construction, this special case of the BFEP thus reduces to finding a solution $\bar{\chi}$ with minimal support such that $\sum_{i \in \mathcal{J}(r)} \bar{\chi}_i \geq 1 \ \forall r \in \mathcal{R}$, which is exactly equivalent to the SCP (A.1), where the decision variable $\bar{\chi}_i$ replaces π_i for each $i \in \mathcal{I} = \mathcal{J}$. $\qquad \square$

Proof of Proposition 3. Consider a solution such that $\widetilde{\chi}^p_j < \widetilde{\chi}^{\mathrm{UB}}_j$ and $\widetilde{\chi}^p_{j'} > 0$ for some $j \in \mathcal{J}$, $j' \in \mathcal{J}(j)$, $p \in \mathcal{P}$. By the monotonicity constraints (1c), these conditions hold for a consecutive sequence of periods $\{p_1, \ldots, p_2\}$, for $p_1 \leq p_2$. Since $\mathcal{R}(j) \supseteq \mathcal{R}(j')$, we can modify the solution without affecting the investment and operational costs by setting $\widetilde{\chi}^{p'}_j \leftarrow \widetilde{\chi}^{p'}_j + 1$ and $\widetilde{\chi}^{p'}_{j'} \leftarrow \widetilde{\chi}^{p'}_{j'} - 1$ for all $p' \in \{p_1, \ldots, p_2\}$, and reassigning to the additional charger of terminal j the charging activities of $\min\left\{\rho, \sum_{r \in \mathcal{R}(j')} \widetilde{w}^{p't}_{rj'}\right\}$ on-route BEBs that relied on terminal j' during each interval $t \in \mathcal{T}$ and period $p' \in \{p_1, \ldots, p_2\}$ in the original solution. These steps can be repeated until no charger remains in locations dominated by terminals with nonzero residual capacity, i.e., until $\widetilde{\chi}^p_{j'} = 0 \ \forall j' \in \mathcal{J}(j)$ if $\widetilde{\chi}^p_j < \widetilde{\chi}^{\mathrm{UB}}_j$ for each period $p \in \mathcal{P}$ and each terminal $j \in \mathcal{J}$. This process terminates after a finite number of steps since the dominance relation of definition 1 is a strict partial order.

From there, we can assume that terminal j has reached its maximum hosting capacity $\widetilde{\chi}^{\mathrm{UB}}_j$ before considering installing chargers at a dominated location $j' \in \mathcal{J}(j)$. The number of vehicles required to satisfy the service requirements (2d) on bus lines connected to terminal j is at most $\max_{p \in \mathcal{P}, t \in \mathcal{T}} \sum_{r \in \mathcal{R}(j)} d^{pt}_r$. The demand for on-route chargers that cannot be satisfied by terminal j is thus upper bounded by:

$$\max\left\{0, \max_{p \in \mathcal{P}, t \in \mathcal{T}} \sum_{r \in \mathcal{R}(j)} d^{pt}_r - \rho \widetilde{\chi}^{\mathrm{UB}}_j\right\}. \tag{A.4}$$

Since dominated terminals are only connected to a subset of the routes $\mathcal{R}(j)$, this upper bound is in particular valid for the residual demand that can be supplied by location $j' \in \mathcal{J}(j)$. The right-hand side of inequality (4) is the smallest integer number of chargers that suffices to satisfy the charging capacity constraints (2c) at terminal j' given this upper bound on charging demand. $\qquad \square$

Proof of Proposition 4. We consider the following generic problem:

$$\min_{x \in \mathcal{X} \subseteq \mathbb{R}^{n_1}_+} \quad f^\top x + \mathcal{Q}(x) \tag{A.5}$$

where the subproblem value function is given by:

$$\mathcal{Q}(x) := \min_{y \in \mathbb{R}^n_+} \quad c^\top y \tag{A.6a}$$

$$\text{s.t. } Ay \geq b - Bx. \tag{A.6b}$$

Problem (A.5) can be expressed in epigraphic form as:

$$\min_{x \in \mathcal{X},\ \theta \in \mathbb{R}_+} \quad f^\top x + \theta \tag{A.7a}$$

$$\text{s.t. } \theta \geq \mathcal{Q}(x). \tag{A.7b}$$

A solution $(x', \theta') \in \mathbb{R}_+^{n_1} \times \mathbb{R}_+$ violates constraint (A.7b) if and only if problem (A.8) is infeasible:

$$\min_{y \in \mathbb{R}_+^m} \quad 0 \tag{A.8a}$$

$$\text{s.t. } Ay \geq b - Bx', \tag{A.8b}$$

$$c^\top y \leq \theta', \tag{A.8c}$$

which equivalently means that the dual problem (A.9) is unbounded:

$$\min_{(\pi, \pi_0) \in \Pi} \quad \pi^\top (b - Bx') - \pi_0 \theta', \tag{A.9}$$

where $\Pi := \{(\pi, \pi_0) \in \mathbb{R}_+^m \times \mathbb{R}_+ : \pi^\top A - \pi_0 c^\top \leq 0\}$.

Problem (A.7) can thus equivalently be expressed as:

$$\min_{x \in \mathcal{X},\ \theta \in \mathbb{R}_+} \quad f^\top x + \theta \tag{A.10a}$$

$$\text{s.t. } \pi^\top (b - Bx) - \pi_0 \theta \leq 0, \quad \forall (\pi, \pi_0) \in \Pi, \tag{A.10b}$$

We consider an arbitrary solution $(x', \theta') \in \mathbb{R}_+^{n_1} \times \mathbb{R}_+$ and a guiding point $(x^o, \theta^o) \in \mathbb{R}_+^{n_1} \times \mathbb{R}_+$ that does not violate any Benders cut (A.10b).

For the pair of points (x', θ'), (x^o, θ^o), the Conforti-Wolsey deepest cut selection problem of Hosseini and Turner (2024) is obtained from problem (A.9) by truncating the dual cone Π with the normalization constraint (A.11b), resulting in the following separation problem:

$$\min_{(\pi, \pi_0) \in \Pi} \quad \pi^\top (b - Bx') - \pi_0 \theta' \tag{A.11a}$$

$$\text{s.t. } \pi^\top B(x^o - x') + \pi_0 (\theta^o - \theta') \leq 1. \tag{A.11b}$$

For the same pair of points (x', θ'), (x^o, θ^o), Seo et al. (2022) formulate the closest cut selection problem as:

$$\min_{(\pi, \pi_0) \in \Pi} \quad \beta = \frac{-\pi^\top (b - Bx^o) + \theta^o \pi_0}{\pi^\top B(x^o - x') + \pi_0 (\theta^o - \theta')} \tag{A.12a}$$

$$\text{s.t. } \pi^\top B(x^o - x') + \pi_0 (\theta^o - \theta') > 0. \tag{A.12b}$$

By maximizing $1 - \beta$ instead of minimizing β, the closest cut selection problem (A.12) can equivalently be expressed as:

$$\max_{(\pi, \pi_0) \in \Pi} \quad 1 - \beta = \frac{\pi^\top (b - Bx') - \pi_0 \theta'}{\pi^\top B(x^o - x') + \pi_0 (\theta^o - \theta')} \tag{A.13a}$$

$$\text{s.t. } \pi^\top B(x^o - x') + \pi_0 (\theta^o - \theta') > 0. \tag{A.13b}$$

Since the objective (A.13a) and the left-hand side of the normalization constraint (A.13b) are both positive homogeneous in the dual variables (π, π_0), the normalization constraint can be replaced by $\pi^\top B(x^o - x') + \pi_0 (\theta^o - \theta') = 1$, which yields the following formulation:

$$\max_{(\pi, \pi_0) \in \Pi} \quad \pi^\top (b - Bx') - \pi_0 \theta' \tag{A.14a}$$

$$\text{s.t. } \pi^\top B(x^o - x') + \pi_0(\theta^o - \theta') = 1. \tag{A.14b}$$

By Proposition 6 of Hosseini and Turner (2024), constraint (A.11b) is binding at optimality in the Conforti-Wolsey deepest cut selection problem (A.11), which is thus equivalent to the closest cut selection problem (A.14). This concludes the proof. $\qquad\square$

As a direct consequence of Proposition 4, the Conforti-Wolsey deepest cuts inherit the theoretical properties demonstrated by Seo et al. (2022) for the closest cuts. Indeed, in addition to supporting the feasible region defined by all Benders cuts, the closest cuts are either facet or improper when the separation problem admits a unique optimal solution, and any optimality cut generated by the closest cut approach is Pareto-optimal if x^o is a core point.

A.3 Illustration of the operational model

We consider a problem with an operational horizon of $T = 6$ intervals, one route, one BEB type, one depot, and one charger type, i.e., $\mathcal{R} = \{r\}$, $\mathcal{B} = \{b\}$, $\mathcal{I} = \{i\}$, and $\mathcal{K} = \{k\}$. In period $p \in \mathcal{P}$, we assume that the fleet is composed of $\bar{\eta}_{rb}^p = 3$ depot BEBs with battery capacity $s_b = 3$, and that the depot is equipped with $\bar{\chi}_{ik}^p = 2$ chargers. Starting from state $s = 2$, one interval suffices to fully recharge the bus, i.e., $\kappa_{rbik2} = 1$, compared to two intervals for states $s \in \{1, 0\}$, i.e., $\kappa_{rbik1} = \kappa_{rbik0} = 2$. The service level requirements are given by $d_r^p = (d_r^{p0}, d_r^{p1}, d_r^{p2}, d_r^{p3}, d_r^{p4}, d_r^{p5}) = (2, 3, 2, 1, 1, 1)$.

Figure A.1: Solution consisting of two cyclic schedules, with periods of one day (left) and two days (right)
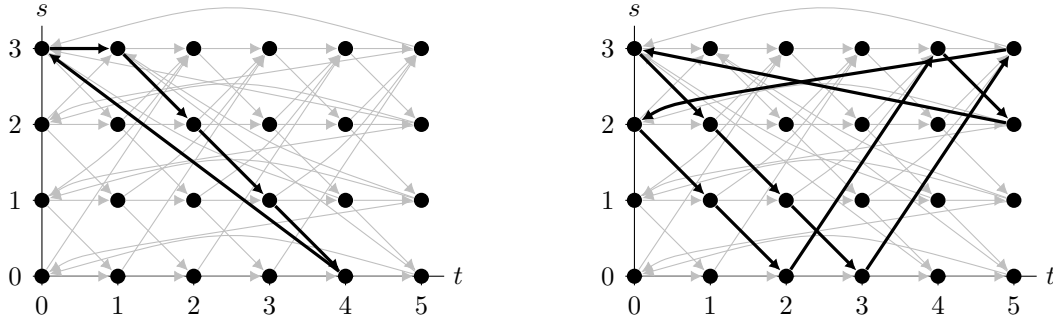

Figure A.1 depicts two cycles in which the arcs in gray have flow 0, and the arcs in black have flow 1. In the first schedule, a bus is in the fully charged state $s = 3$ at time $t = 0$ and idles for one interval, reaching state $(t, s) = (1, 3)$. It is then in service for three intervals, reaching state $(t, s) = (4, 0)$. Finally, it charges for two intervals, completing the cycle by returning to state $(t, s) = (0, 3)$. The period of this cycle is T intervals. This means that a single bus can repeat this schedule daily. The second schedule can be interpreted similarly. However, since it spans $2T$ intervals, it must be executed by two buses staggered by T intervals. At time $t = 0$, one bus is thus in state $s = 2$, whereas the second bus is in state $s = 3$. After one day, their role are inverted. It follows that each arc along the cycle is traversed once per day, and the operations repeat daily at the fleet level.

The circulation obtained by summing these two cycles is a feasible solution for the operational problem (2) associated with the above parameters. It corresponds to setting the service variables w_{rbs}^{pt} such that $(t, s) \in \{(1, 3), (2, 2), (3, 1), (0, 3), (1, 2), (2, 1), (5, 3), (0, 2), (1, 1), (4, 3)\}$, the idling variables v_{rbs}^{pt} such that $(t, s) = (0, 3)$, and the charging variables z_{rbiks}^{pt} such that $(t, s) \in \{(4, 0), (2, 0), (3, 0), (5, 2)\}$ to 1, and all the other operational variables to 0. The service level in this solution matches the minimum requirements

imposed by constraints (2d). The charger usage is 0 in intervals $t \in \{0,1\}$, 1 in interval $t = 2$, and 2 in intervals $t \in \{3,4,5\}$, which respects the depot capacity constraints (2b) for $\bar{\chi}_{ik}^p = 2$. The flow balance constraints (2e)-(2h) are satisfied since the solution is composed of a set of cycles with unit flows. The active variables that contribute to the fleet size constraint (2h) are w_{rb3}^{p0}, w_{rb2}^{p0}, and v_{rb3}^{p0}, and the limit of $\bar{\eta}_{rb}^p = 3$ vehicles is thus respected. Finally, constraints (2c) and (2i) are trivially satisfied due to the absence of conventional buses and on-route BEBs in the solution.

This example illustrates that our model allows generating schedules that could not be achieved if, as customary in the literature, each vehicle had to perform the same schedule daily. This increased flexibility enables better coordination of the operations and allows meeting the service level requirements with fewer BEBs. Indeed, in our example, one can verify that any cycle that repeats daily contains at most three service arcs. Since the total demand is $\sum_{t \in \mathcal{T}} d_r^{pt} = 10$, four BEBs would be needed instead of three if each vehicle had to perform the same schedule daily. This modeling choice can therefore result in significant savings in many instances.

A.4 Approximate operational model

To approximate the operational decisions using a reduced number of variables, we introduce for each route $r \in \mathcal{R}$ and each period $p \in \mathcal{P}$ a collection $\omega_r^p = (\bar{\omega}_r^p, \widetilde{\omega}_r^p, \widehat{\omega}_r^p) \in \mathbb{R}_+^{|\mathcal{B}|+2}$ of auxiliary variables representing the average service level provided by each type of bus:

$$\bar{\omega}_{rb}^p = \frac{1}{T} \sum_{t \in T} \sum_{s \in \mathcal{S}_b^w} w_{rbs}^{pt}, \qquad\qquad \forall b \in \mathcal{B}, \qquad\qquad\qquad (A.15a)$$

$$\widetilde{\omega}_r^p = \frac{1}{T} \sum_{t \in T} \sum_{j \in \mathcal{J}(r)} \widetilde{w}_{rj}^{pt}, \qquad\qquad\qquad\qquad\qquad\qquad (A.15b)$$

$$\widehat{\omega}_r^p = \frac{1}{T} \sum_{t \in T} \widehat{w}_r^{pt}. \qquad\qquad\qquad\qquad\qquad\qquad\qquad (A.15c)$$

Averaging the service level constraints (2d) of each period and each route over the time intervals $t \in \mathcal{T}$ yields a surrogate relaxation that can be expressed using the auxiliary variables (A.15a)–(A.15c):

$$\sum_{b \in \mathcal{B}} \bar{\omega}_{rb}^p + \widetilde{\omega}_r^p + \widehat{\omega}_r^p \geq \frac{1}{T} \sum_{t \in \mathcal{T}} d_r^{pt}, \quad \forall r \in \mathcal{R}, \ p \in \mathcal{P}. \qquad\qquad (A.16)$$

For constraints (A.16) to serve as nontrivial feasibility cuts in the master problem, the auxiliary variables are then upper bounded by linear functions of the strategic variables. We obtain a first set of bounds by averaging the fleet size constraints (2h)–(2i) over the operational time intervals:

$$\bar{\omega}_{rb}^p \leq \alpha_{rb} \bar{\eta}_{rb}^p, \qquad\qquad \forall p \in \mathcal{P}, \ r \in \mathcal{R}, \ b \in \mathcal{B}, \qquad\qquad (A.17a)$$

$$\widetilde{\omega}_r^p \leq \widetilde{\eta}_r^p, \qquad\qquad \forall p \in \mathcal{P}, \ r \in \mathcal{R}, \qquad\qquad\qquad (A.17b)$$

$$\widehat{\omega}_r^p \leq \widehat{\eta}_r^p, \qquad\qquad \forall p \in \mathcal{P}, \ r \in \mathcal{R}, \qquad\qquad\qquad (A.17c)$$

where $\alpha_{rb} := \max_{i \in \mathcal{I}, k \in \mathcal{K}, s \in \mathcal{S}_b^z} \frac{s_b - s}{s_b - s + \kappa_{rbiks}}$ is an upper bound on the fraction of the time intervals during which depot BEBs of type $b \in \mathcal{B}$ can be in service on route $r \in \mathcal{R}$ while respecting the conservation constraints (2e)–(2g). This bound is attained by alternating indefinitely between working and recharging from the state $s \in \mathcal{S}_b^z$ that maximizes the ratio of regained battery level $s_b - s$ to charging time $\min_{i \in \mathcal{I}, k \in \mathcal{K}} \kappa_{rbiks}$ achieved by using the fastest charger type at the nearest depot.

The average service level can also be bounded by the installed charging infrastructure. To do so, we average the on-route charging capacity constraints (2c) of period $p \in \mathcal{P}$ over the time intervals $t \in \mathcal{T}$. From there, summing over the set of terminals $j \in \mathcal{J}(r)$ connected to a route $r \in \mathcal{R}$, and over all the terminals $j \in \mathcal{J}$, gives inequalities (A.18a) and (A.18b), respectively.

$$\widetilde{\omega}_r^p \leq \sum_{j \in \mathcal{J}(r)} \rho \widetilde{\chi}_j^p, \quad \forall p \in \mathcal{P}, \ r \in \mathcal{R}, \tag{A.18a}$$

$$\sum_{r \in \mathcal{R}} \widetilde{\omega}_r^p \leq \sum_{j \in \mathcal{J}} \rho \widetilde{\chi}_j^p, \quad \forall p \in \mathcal{P}. \tag{A.18b}$$

Constraints (A.18a) enforce that the chargers installed at terminals connected to a route $r \in \mathcal{R}$ suffice to satisfy the average energy requirements of the on-route BEBs assigned to this route. In contrast, constraints (A.18b) ensure that the total charging capacity can satisfy the energy requirements of the fleet over the complete network.

Similar bounds can be devised for depot BEBs. Let $\beta_{ik} := \max_{r \in \mathcal{R}, b \in \mathcal{B}, s \in \mathcal{S}_b^z} \frac{s_b - s}{\kappa_{rbiks}}$ denote the maximum number of battery states that can be regained per time interval using a charger of type $k \in \mathcal{K}$ installed at depot $i \in \mathcal{I}$. By the depot charging capacity constraints (2b), which limit to $\bar{\chi}_{ik}^p$ the number of vehicles simultaneously using chargers of type $k \in \mathcal{K}$ at depot $i \in \mathcal{I}$ during period $p \in \mathcal{P}$, it follows that no more than $\sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} \beta_{ik} \bar{\chi}_{ik}^p$ units of charge can be regained by the fleet of depot BEBs in each time interval. From there, since the flow conservation constraints (2e)–(2g) imply that the total flow $\sum_{t \in \mathcal{T}} \sum_{r \in \mathcal{R}} \sum_{b \in \mathcal{B}} \sum_{s \in \mathcal{S}_b^w} w_{rbs}^{pt}$ on service arcs equals the number of units of charge regained by the fleet over the operational horizon, the following inequalities hold:

$$\sum_{r \in \mathcal{R}} \sum_{b \in \mathcal{B}} \bar{\omega}_{rb}^p \leq \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} \beta_{ik} \bar{\chi}_{ik}^p, \quad \forall p \in \mathcal{P}. \tag{A.19}$$

Upper bounds on the average service level of on-route BEBs and conventional buses can also be devised from the service level constraints. Indeed, since, for each $(p, r, t) \in \mathcal{P} \times \mathcal{R} \times \mathcal{T}$, the variables $\widetilde{w}_r^{pt}$ and $\widehat{w}_r^{pt}$ are assumed to have positive objective coefficients in model (3), and the service level constraint (2d) is the only one that impose a lower bound on their value, the level of service of conventional buses and on-route BEBs never exceeds the demand d_r^{pt} in an optimal solution. From there, given that $m := \widetilde{\eta}_r^p + \widehat{\eta}_r^p$ conventional buses and on-route BEBs are assigned to route r in period p, their average service level $\widetilde{\omega}_r^p + \widehat{\omega}_r^p$ can be upper bounded by $\sigma_{pr}(m) := \frac{1}{T} \sum_{t \in \mathcal{T}} \min\{d_r^{pt}, m\}$. Analogously to the lower bounds (6), these upper bounds can be imposed by constructing the piecewise-linear upper envelope of the set of points $\mathcal{M}'_{pr} := \{(m, \sigma_{pr}(m))\}_{m=0}^{\max_{t \in \mathcal{T}} d_r^{pt}}$, giving:

$$(\widetilde{\eta}_r^p + \widehat{\eta}_r^p, \widetilde{\omega}_r^p + \widehat{\omega}_r^p) \in \text{conv}\left(\text{hypo}\left(\mathcal{M}'_{pr}\right)\right). \tag{A.20}$$

Finally, for $(p, r) \in \mathcal{P} \times \mathcal{R}$, let $c_{pr}^{\widehat{\omega}}/T := \min_{t \in \mathcal{T}} c_{ptr}^{\widehat{w}}$ and $c_{pr}^{\widetilde{\omega}}/T := \min_{t \in \mathcal{T}, j \in \mathcal{J}(r)} c_{ptrj}^{\widetilde{w}}$ denote the smallest cost of operating a conventional bus and an on-route BEB for one interval, respectively. Similarly, let $c_{prb}^{\bar{\omega}}/T := \min_{t \in \mathcal{T}, s \in \mathcal{S}_b^w} c_{ptrbs}^w + \min_{t \in \mathcal{T}, i \in \mathcal{I}, k \in \mathcal{K}, s \in \mathcal{S}_b^z} \frac{c_{ptrbiks}^z}{s_b - s}$ denote the smallest cost of operating a depot BEB of type $b \in \mathcal{B}$ for one time interval, where the second term in the definition represents the minimal charging costs needed to replenish one unit of charge. The scheduling costs can be bounded by the average service level variables as follows:

$$\theta_{pr} \geq \sum_{b \in \mathcal{B}} c_{prb}^{\bar{\omega}} \bar{\omega}_{rb}^p + c_{pr}^{\widetilde{\omega}} \widetilde{\omega}_r^p + c_{pr}^{\widehat{\omega}} \widehat{\omega}_r^p. \tag{A.21}$$

We approximate the operational model by including in the master problem the auxiliary variables $\bar{\omega}_r^p \in \mathbb{R}_+^{\mathcal{B}}$, $\widetilde{\omega}_r^p \in \mathbb{R}_+$, and $\widehat{\omega}_r^p \in \mathbb{R}_+$ for each $(p,r) \in \mathcal{P} \times \mathcal{R}$, along with constraints (A.16)–(A.21). The approximation of the operational problem provided by this method is weaker than that of the semi-relaxation approach, which retains all the original variables in the master. However, since our approach relies on a very small number of variables and constraints to approximate the operational model, it only marginally increases the computational cost of solving the master problem. Our preliminary experiments showed that it consistently outperforms the semi-relaxed formulation and the standard approach that does not include an approximate operational model in the master.

A.5   Pseudocode of the accelerated logic-based Benders decomposition

---

Algorithm 1 Accelerated logic-based Benders decomposition algorithm

---

1: Initialize LB $\leftarrow 0$;   UB $\leftarrow +\infty$; $l \leftarrow 1$;

2: Initialize the relaxed master problem MP^l:

3: $\displaystyle \min_{x \in \mathcal{X}, \Theta \in \mathbb{R}_+^{\mathcal{P}}, \theta \in \mathbb{R}_+^{\mathcal{P} \times \mathcal{R}}, \omega \in \mathbb{R}_+^{\mathcal{P} \times \mathcal{R} \times (|\mathcal{B}|+2)}} \sum_{p \in \mathcal{P}} f_p^\top x_p + \sum_{p \in \mathcal{P}} \gamma^{p-1} \Theta_p$ s.t. (11), (A.16)–(A.21)

4: **while** $(\text{UB} - \text{LB})/\text{UB} > \text{rel\_tol}$ **do**

5:    Solve MP^l and save the components (x', Θ', θ') of its optimal solution

6:    Update LB $\leftarrow \sum_{p \in \mathcal{P}} f_p^\top x_p' + \sum_{p \in \mathcal{P}} \gamma^{p-1} \Theta_p'$

7:    **for** $p \in \mathcal{P}$ **do**

8:       Solve the cut generation problem (20), with $h = h_{cc}$, and get the solution $(\bar{\pi}, \bar{\pi}_0)$

9:       **if** a (x_p', Θ_p') violates the unified cut (19) associated with $(\bar{\pi}, \bar{\pi}_0)$ **then**

10:          Generate the unified cut (19)

11:          **for** $r \in \mathcal{R}$ **do**

12:             Repeat steps 8 to 10, with $\{r\}$ replacing $\mathcal{R}$ and θ_{pr} replacing Θ_p

13:          **end for**

14:       **end if**

15:    **end for**

16:    **if** the LP relaxation of (3) at x_p' is feasible for each $p \in \mathcal{P}$ **then**

17:       **for** $p \in \mathcal{P}$ **do**

18:          Solve the operational problem (3), with optimal value $\mathcal{Q}_p(x_p')$

19:          **if** $\Theta_p' < \mathcal{Q}_p(x_p') < \infty$ **then**

20:             Generate the optimality cut $\Theta_p \geq \mathcal{D}_p^{\text{IP}}(x_p; x_p')$

21:          **else if** $\mathcal{Q}_p(x_p') = \infty$ **then**

22:             For each $r \in \mathcal{R}$ such that $\mathcal{Y}_{pr}(\eta_{pr}') = \emptyset$, generate the violated cut (23). If none is violated, generate (24) if $\bar{\chi}_p^{\text{LB}}(\widetilde{\chi}_p', \eta_p') > \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} \bar{\chi}_{ik}'^p$, and (25) otherwise.

23:          **end if**

24:       **end for**

25:       **if** $\sum_{p \in \mathcal{P}} f_p^\top x_p' + \sum_{p \in \mathcal{P}} \gamma^{p-1} \mathcal{Q}_p(x_p') < \text{UB}$ **then**

26:          Update UB $\leftarrow \sum_{p \in \mathcal{P}} f_p^\top x_p' + \sum_{p \in \mathcal{P}} \gamma^{p-1} \mathcal{Q}_p(x_p')$; Set x' as the incumbent solution

27:       **end if**

28:    **end if**

29: **end while**

30: **return** the incumbent solution, with optimal value UB

Algorithm 1 provides the detailed outline of our logic-based Benders decomposition approach. The global bounds and iteration counter are initialized in step 1. The relaxed master problem, equipped with the disaggregated auxiliary variables of Section 4.2.1 and the approximated operational model described in Section 4.2.2, and detailed in Appendix A.4, is initialized in step 2. Each iteration of the algorithm comprises three phases. In the first phase (steps 5-6), the current master problem is solved, and the global lower bound is updated. In the second phase (steps 7-15), multi-route and single-route Benders cuts are generated at the current master solution using the closest cut selection method from Section 4.2.3. In the third phase (steps 16-28), the operational problems associated with the current master solution are solved (step 18), which allows generating optimality cuts (step 20) and feasibility cuts (step 22) as described in Section 4.2.4, and updating the incumbent solution (step 26). These steps are repeated until the relative tolerance rel_tol (e.g. 0.01%) is reached.

This implementation reflects a classical Benders decomposition framework, where the master problem is solved to optimality at each iteration. Most recent applications of Benders decomposition instead use a Branch-and-cut framework, in which violated Benders cuts are separated at each integer node (and optionally at fractional nodes) of the branch-and-bound tree (Rahmaniani et al. 2017). The Branch-and-Benders-cut framework is not directly applicable in the context of our logic-based Benders algorithm, since adding monotone cuts requires introducing new indicator variables to the model, and MILP solvers usually do not support adding columns to a model inside callbacks. A possible alternative is to generate the Benders cuts in a Branch-and-Benders-cut fashion before resorting to monotone cuts. This way, at each iteration, the identified master solution satisfies all the Benders cuts (in particular, the first iteration provides an optimal solution for the BFEP without integrality constraints on the operational variables), and the integer phase is then executed as in Algorithm 1 before moving to the next iteration. In preliminary computational experiments, we observed that this hybrid approach can outperform the classical framework for small instances. However, it produces significantly more Benders cuts at each iteration compared to the classical approach, which leads to very expensive master problem iterations.

We obtained our best results by instead adding an early-stopping condition on the master problem in the classical framework, a technique that has previously been leveraged in the context of network design problems (Geoffrion and Graves 1974, Kewcharoenwong and Üster 2014). To limit the computing time spent on the master problem iterations, we exit step 5 early when more than two seconds have been spent on the current master problem, the lower bound LB^l on the current master problem is larger than or equal to the global lower bound LB, and the upper bound UB^l on the current master problem and the global upper bound respect $(\text{UB}-\text{UB}^l)/\text{UB} > \text{rel\_tol}$. The respective objective of these three conditions is to disable early stopping when the master problem is very inexpensive to solve, to ensure that some progress is made on the global lower bound at each iteration, and to avoid stopping if the current relaxed master problem may suffice to show that the incumbent global solution attains the desired optimality gap. When early-stopping is used, the only other change that must be applied to Algorithm 1 is to replace the update rule in step 6 by $\text{LB} \leftarrow \text{LB}^l$. The experiments of Section 5 are based on this version of the algorithm.

A.6 Pseudocode of the arc selection algorithm

Algorithm 2 Arc selection algorithm

1: Solve the LP relaxation of the extensive formulation, let $\mathcal{E}^{\mathrm{LP1}}$ be the set of active arcs (w, v, z)

2: Solve the LP relaxation of the extensive formulation without early charging, i.e. with $\mathcal{S}_b^z = \{0\}\ \forall b \in \mathcal{B}$, let $\mathcal{E}^{\mathrm{LP2}}$ be the set of active arcs (w, v, z)

3: **for** $(p, r, b) \in \mathcal{P} \times \mathcal{R} \times \mathcal{B}$ **do**

4:     **for** $m \in \{0, 1, \ldots, \max_{t \in \mathcal{T}} d_r^{pt} - 1\}$ **do**

5:         With only depot BEBs of type b, i.e. $\mathcal{B} \leftarrow \{b\}$, solve (5), with optimal value $\bar{\eta}_{pr}^{\mathrm{LB}}(m)$

6:         Solve the following single-route depot BEB scheduling problem:

7:         $$\min_{\widetilde{\eta}_r^p, \widehat{\eta}_r^p \in \mathbb{Z}_+,\ \bar{\eta}_r^p \in \mathbb{Z}_+^{\mathcal{B}},\ y_{pr} \in \mathbb{Z}_+^{m_{pr}}} c_{pr}^{y\top} y_{pr}$$

8:         s.t. $D_r y_{pr} \leq e_{pr} + E \eta_{pr}, \quad \widehat{\eta}_r^p + \bar{\eta}_r^p \leq m, \quad \bar{\eta}_{rb}^p = \bar{\eta}_{pr}^{\mathrm{LB}}(m), \quad \bar{\eta}_{rb'}^p = 0\ \forall b' \in \mathcal{B} \setminus \{b\}$

9:         Let $\mathcal{E}_{prbm}^{\mathrm{SR}}$ be the set of active arcs $(w_{rb}^p, v_{rb}^p, z_{rb}^p)$ in the optimal solution

10:     **end for**

11:     Collect the active arcs for bus type b on route r in period p as $\mathcal{E}_{prb}^{\mathrm{SR}} := \bigcup_{m=0}^{\max_{t \in \mathcal{T}} d_r^{pt} - 1} \mathcal{E}_{prbm}^{\mathrm{SR}}$

12: **end for**

13: Collect the active arcs from the single-route problems as $\mathcal{E}^{\mathrm{SR}} := \bigcup_{(p,r,b) \in \mathcal{P} \times \mathcal{R} \times \mathcal{B}} \mathcal{E}_{prb}^{\mathrm{SR}}$

14: Solve the extensive formulation with positive flows allowed on the arcs $\mathcal{E}^{\mathrm{LP1}} \cup \mathcal{E}^{\mathrm{LP2}}$. Save the incumbent solution (x^1, y^1) when the time limit is reached

15: Solve the extensive formulation with positive flows allowed on the arcs $\mathcal{E}^{\mathrm{LP1}} \cup \mathcal{E}^{\mathrm{LP2}} \cup \mathcal{E}^{\mathrm{SR}}$, warm-started at (x^1, y^1). Save the incumbent solution (x^2, y^2) when the time limit is reached

16: Solve the extensive formulation, warm-started at (x^2, y^2). Save the incumbent solution (x^3, y^3) when the time limit is reached, along with the bounds LB and UB on the optimal value

17: **return** the incumbent solution (x^3, y^3), and the bounds LB and UB

A.7 Description of instances

Table A.2: Characteristics of complete network instances

City	# Depots	# Terminals	# Routes	# Buses	Service rate (%)
Atlanta	5	115	110	441	68.5
Boston	9	214	166	1286	53.1
Chicago	7	235	126	1495	51.3
Dallas	3	152	149	623	55.8
Detroit	4	90	42	214	60.6
Houston	8	135	117	941	47.3
Las Vegas	2	84	39	280	71.4
Los Angeles	12	188	140	1817	51.9

We assume that each terminal location has an installation capacity of $\widetilde{\chi}_j^{\mathrm{UB}} = 2$ on-route chargers that can serve up to $\rho = 8$ on-route BEB each hour. We used geospatial images to estimate the maximum installation capacity of depot chargers in each depot. All the cost coefficients we use are estimated using the baseline

parameters presented in Johnson et al. (2020) and the references therein. We consider a single type of depot charger ($|\mathcal{K}|=1$) reflecting typical specifications for a 70kW AC charger, with a cost of \$60.05k, including lifetime maintenance. The on-route chargers represent 325kW DC chargers with an acquisition cost of \$877.59k. Also, we consider $|\mathcal{B}| = 2$ models of depot BEBs, respectively with $s_1 = 6$ and $s_2 = 12$ hours of operational capacity, all requiring $\kappa_{rbiks} = \lceil \frac{s - s_b}{2} \rceil$ time intervals to travel between the route and any depot and fully charge before reentering service. The unit cost of short-range depot BEBs is set to \$943k per unit, compared to \$1093k for long-range depot BEBs and on-route BEBS. Salvage revenues from the retirement of conventional buses are ignored. The operating costs of conventional buses were estimated to \$50 per hour of service, compared to \$29 and \$31 for depot BEBs and on-route BEBs after factoring charging and deadhead costs, the later depending on the distance from each route to the selected charging location in the case of depot BEBs. In addition, a fixed yearly maintenance cost of \$10k per unit is applied to the conventional buses. The operational costs are computed based on 250 days of service per year, and a yearly discount factor of 4% ($\gamma = 0.96$) is used. In all the experiments, we assume the initial fleet to be composed exclusively of conventional buses, which must all be retired by the end of the planning horizon ($\widehat{\eta}_P^{\mathrm{UB}} = 0$). For each instance, we estimate the minimum investment budget needed to satisfy the electrification targets by solving the LP relaxation of the problem with an alternative objective. This budget is then multiplied by 1.5 (2.5 for instances where no fast chargers are allowed), and equally divided over the P investment periods to obtain the yearly budget parameters I_p^{UB}. Table A.2 summarizes the characteristics of each city. The number of buses refers to the initial state of the system (year $p = 0$), and the service rate corresponds to the ratio of average service requirements $\frac{1}{T} \sum_{r \in \mathcal{R}} \sum_{t \in \mathcal{T}} d_r^{0t}$ to fleet size in year 0.

A.8 Comparison of Benders cut selection methods

Instances ($|\mathcal{R}|, |\mathcal{P}|$)	Cuts type	Summary				Cuts			Time (%)		
		Opt	Gap	Time	Iter	BCuts	MCuts	Ind	MP	LP	IP
(3,4)	Standard	**10**	**0.0000**	12.1	9.4	58.2	2.1	4.5	29.7	19.6	21.0
	MIS	**10**	**0.0000**	17.5	7.4	40.3	1.8	3.3	9.7	63.3	8.6
	Closest	**10**	**0.0000**	**11.1**	**5.9**	33.1	**1.7**	**3.1**	10.5	44.7	10.2
	MW	**10**	**0.0000**	11.2	6.1	**30.9**	1.9	4.0	11.2	37.6	15.2
	Deepest	**10**	**0.0000**	16.9	7.3	39.4	1.9	3.3	8.9	62.3	9.7
(6,6)	Standard	**10**	**0.0000**	108.1	21.0	230.1	2.4	7.2	63.1	13.5	14.6
	MIS	**10**	**0.0000**	79.7	12.6	137.0	2.0	4.9	26.5	55.3	8.7
	Closest	**10**	**0.0000**	**56.9**	**10.3**	113.3	1.9	4.9	29.7	38.3	13.3
	MW	**10**	**0.0000**	61.7	11.2	**105.6**	2.1	6.4	31.4	37.7	13.2
	Deepest	**10**	**0.0000**	76.5	11.8	131.4	**1.7**	**3.3**	25.7	55.3	9.2
(9,6)	Standard	6	0.0334	2137.6	61.3	581.8	5.5	35.5	82.1	6.3	9.8
	MIS	**9**	0.0051	601.3	27.9	296.9	**3.3**	**9.4**	53.5	35.3	6.8
	Closest	**9**	**0.0037**	**372.8**	**24.0**	**247.9**	3.5	13.1	54.5	29.3	7.7
	MW	8	0.0071	474.9	27.3	257.0	4.2	17.8	53.6	27.1	10.9
	Deepest	**9**	0.0048	567.3	27.2	286.4	4.1	14.6	52.3	38.0	5.4

Table A.3: Performance of LBBD with different linear Benders cuts selection methods

In Table A.3, we evaluate the performance of our accelerated logic-based Benders decomposition algorithm (LBBD) with each Benders cut selection technique from Section 4.2.3, namely the standard Benders cuts (Benders 1962), the MIS cuts (Fischetti et al. 2010), the MW cuts (Magnanti and Wong 1981), the closest cuts (Seo et al. 2022), and the $\ell-1$ deepest cuts (Hosseini and Turner 2024). We also performed preliminary experiments with the $\ell-2$ deepest cuts, but solving the resulting nonlinear cut selection problems proved to be prohibitively expensive. The experiments are performed on the instances presented in the ablation study of Section 5.1, and the same performance metrics are reported.

The results show that the Benders cut selection method has a significant impact on the overall performance of the algorithm. For $(|\mathcal{R}|, |\mathcal{P}|) = (9, 6)$, nine instances are solved to optimality with the unified cut selection methods, compared to eight with the MW cuts, and six with the standard cuts. The closest cuts provide the best performance across all groups of instances. They divide the average number of iterations and generated cuts by more than two compared to standard cuts, and provide a speedup factor exceeding two orders of magnitude for some instances. Although separating stronger Benders cuts requires more computational effort than standard cuts, solving the master problem is the most computationally expensive step of our algorithm for challenging instances, hence the importance of limiting the number of iterations and generated cuts.

A.9 Detailed results on restriction heuristics

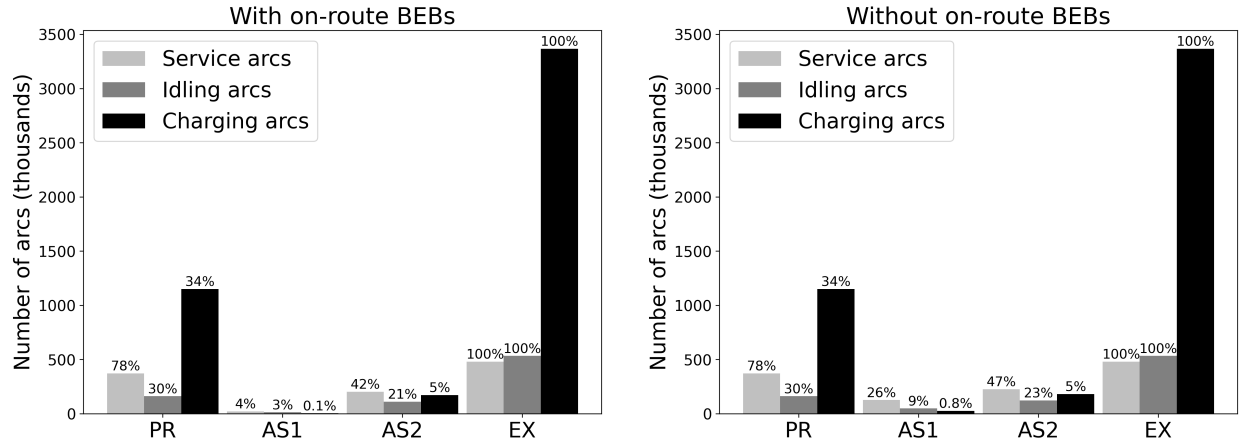Figure A.2: Average number of arcs in restriction heuristics models - Complete networks

Figure A.2 presents the total number of arcs in the depot BEB scheduling graphs for the extensive formulation (EX), the policy restriction model (PR), and the first (AS1) and second (AS2) restricted models solved in the arc selection algorithm (steps 14 and 15 of Algorithm 2). The values are averaged over the complete network instances of Section 5.3. The fraction of each type of arcs (service arcs w, idling arcs v, charging arcs z) retained in each model is also reported. Figure A.3 presents the average lower and upper bounds on the objective value of each model solved in the PR and AS algorithms. The values are normalized based on the best-known lower bound for each instance. Note that time limits of 1.5 hours and 4.5 hours are respectively given to models AS1 and AS2, whereas 6 hours are given to model PR. In both the AS and PR algorithms, the warm-started EX model is then solved for two hours.

40

Figure A.3: Average objective bounds of restriction heuristics models - Complete networks
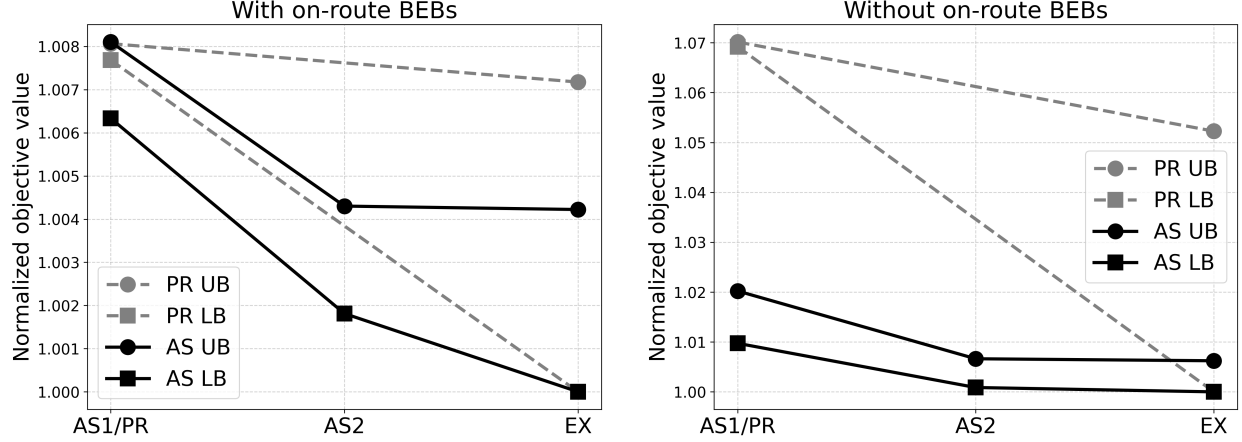
The results show that, although models AS1 and AS2 are much sparser, they capture better solutions than PR, especially for instances that rely exclusively on depot charging. AS1 retains 4.6% and 0.9% of the arcs in the first and second groups of instances, respectively. These values are 12.1% and 11.1% for AS2, and 38.4% in both cases for PR, whose a priori selection rules retain the same arcs whether on-route BEBs are available or not. In the second group of instances, AS1 provides feasible solutions with optimality gaps of 2% for the unrestricted model, compared to 7% for PR. We observe that the PR model can consistently be solved to near-optimality, confirming that the a priori restriction rules eliminate the best feasible solutions for the feasible set. In contrast, models AS1 and AS2 are more difficult to solve, but the good upper bounds they provide confirm the existence of high-quality solutions that use only their sparse depot BEB operational graphs. In algorithm PR, the two-hour computing budget allocated to the warm-started model EX suffices to improve the heuristic solution provided by the restricted model. In contrast, the best solution identified by model AS2 is almost never improved in the last phase of algorithm AS. In this case, the sole purpose of solving the warm-started model EX is to obtain a valid optimality gap for the original problem.

A.10 Properties of optimal sequences of investments

In this section, we consider a generic multi-period integrated planning problem defined on a static system. Studying the properties of this problem allows us to devise structural results on the sequence of optimal investments for important special cases of the BFEP.

We consider problems of the following form:

$$\min_{\substack{x_p \in \mathcal{X}_p,\ p \in [P] \\ x_p \in \mathcal{X}'_p(x_{p-1}),\ p \in [P]}} \quad \sum_{p=1}^{P} \gamma^{p-1} \left(\bar{T}(x_{p-1}, x_p) + S(x_p) \right), \tag{A.22}$$

where $\mathcal{X}_p$ denotes the feasible states for period $p \in [P]$, $\mathcal{X}'_p(x_{p-1})$ is the set of states that can be visited in period p given that the system is in state x_{p-1} in period $p-1$, and $\mathcal{X}_\bullet \supseteq \cup_{p=1}^{P} \mathcal{X}_p$ contains all the states the system can visit. We assume that the transition cost function $\bar{T} : \mathcal{X}_\bullet \times \mathcal{X}_\bullet \to \mathbb{R}$ satisfies $\bar{T}(x_p, x_p) = 0$ $\forall x_p \in \mathcal{X}_\bullet$ and respects the triangle inequality $\bar{T}(x_p, x_{p+2}) \leq \bar{T}(x_p, x_{p+1}) + \bar{T}(x_{p+1}, x_{p+2})$, $\forall x_p, x_{p+1}, x_{p+2} \in \mathcal{X}_\bullet$. The cost of visiting a state is defined by the function $S : \mathcal{X}_\bullet \to \mathbb{R}$, and the initial state is the element of the singleton $\mathcal{X}_0 = \{x_0\}$.

In Proposition 5, we show that, if a sequence of transitions can be advanced and delayed while maintaining the feasibility of an optimal solution in this context, then applying either of these changes also preserves optimality.

Proposition 5. *For any optimal solution x to problem (A.22) and any pair of periods $p' \leq p''$, $p',p'' \in \{2,\ldots,P-1\}$, if solutions x^1 and x^2, defined as:*

$$x_p^1 = \begin{cases} x_p & ,\forall p \in \{0,\ldots,p'-2\} \cup \{p'',\ldots,P\} \\ x_{p+1} & ,\forall p \in \{p'-1,\ldots,p''-1\} \end{cases}, \tag{A.23}$$

$$x_p^2 = \begin{cases} x_p & ,\forall p \in \{0,\ldots,p'-1\} \cup \{p''+1,\ldots,P\} \\ x_{p-1} & ,\forall p \in \{p',\ldots,p''\} \end{cases}, \tag{A.24}$$

are both feasible, then they are both optimal.

Proof. Proof. Advancing by one period the sequence of transitions made in periods $\{p',\ldots,p''\}$ affects the transition costs of period $p'-1$, which change from $\bar{T}(x_{p'-2},x_{p'-1})$ to $\bar{T}(x_{p'-2},x_{p'})$, and of period p'', which change from $\bar{T}(x_{p''-1},x_{p''})$ to $\bar{T}(x_{p''},x_{p''}) = 0$. In addition, in each period $p \in \{p'-1,\ldots,p''-1\}$, solution x^1 visits state x_{p+1} instead of state x_p. The difference Δ_1 between the objective value of solutions x^1 and x is thus:

$$\Delta_1 = \gamma^{p'-2}\left(\bar{T}(x_{p'-2},x_{p'}) - \bar{T}(x_{p'-2},x_{p'-1})\right) - \gamma^{p''-1}\bar{T}(x_{p''-1},x_{p''}) + \sum_{i=0}^{p''-p'} \gamma^{p'-2+i}\left(S(x_{p'+i}) - S(x_{p'-1+i})\right)$$

$$= \gamma^{p'-2}\left[\bar{T}(x_{p'-2},x_{p'}) - \bar{T}(x_{p'-2},x_{p'-1}) - \gamma^{p''-p'+1}\bar{T}(x_{p''-1},x_{p''}) + \sum_{i=0}^{p''-p'} \gamma^{i}\left(S(x_{p'+i}) - S(x_{p'-1+i})\right)\right]$$

$$\leq \gamma^{p'-2}\left[\bar{T}(x_{p'-1},x_{p'}) - \gamma^{p''-p'+1}\bar{T}(x_{p''-1},x_{p''}) + \sum_{i=0}^{p''-p'} \gamma^{i}\left(S(x_{p'+i}) - S(x_{p'-1+i})\right)\right] \tag{A.25}$$

$$=: \Delta,$$

where the inequality follows from the property of the transition cost function, which satisfies $\bar{T}(x_{p'-2},x_{p'}) \leq \bar{T}(x_{p'-2},\bar{x}_{p'-1}) + \bar{T}(\bar{x}_{p'-1},x_{p'})$ for any $\bar{x}_{p'-1} \in \mathcal{X}_\bullet$.

Similarly, deferring by one period the sequence of transitions made in periods $\{p',\ldots,p''\}$ affects the transition costs of period $p''+1$, which change from $\bar{T}(x_{p''},x_{p''+1})$ to $\bar{T}(x_{p''-1},x_{p''+1})$, and of period p', which change from $\bar{T}(x_{p'-1},x_{p'})$ to $\bar{T}(x_{p'-1},x_{p'-1}) = 0$. Also, in each period $p \in \{p',\ldots,p''\}$, solution x^2 visits state x_{p-1} instead of state x_p, so that the difference Δ_2 between the objective value of solutions x^2 and x is:

$$\Delta_2 = \gamma^{p''}\left(\bar{T}(x_{p''-1},x_{p''+1}) - \bar{T}(x_{p''},x_{p''+1})\right) - \gamma^{p'-1}\bar{T}(x_{p'-1},x_{p'}) + \sum_{i=0}^{p''-p'} \gamma^{p'-1+i}\left(S(x_{p'-1+i}) - S(x_{p'+i})\right)$$

$$= \gamma^{p'-1}\left[\gamma^{p''+1-p'}\left(\bar{T}(x_{p''-1},x_{p''+1}) - \bar{T}(x_{p''},x_{p''+1})\right) - \bar{T}(x_{p'-1},x_{p'}) + \sum_{i=0}^{p''-p'} \gamma^{i}\left(S(x_{p'-1+i}) - S(x_{p'+i})\right)\right]$$

$$\leq \gamma^{p'-1}\left[\gamma^{p''+1-p'}\bar{T}(x_{p''-1},x_{p''}) - \bar{T}(x_{p'-1},x_{p'}) + \sum_{i=0}^{p''-p'} \gamma^{i}\left(S(x_{p'-1+i}) - S(x_{p'+i})\right)\right] \tag{A.26}$$

$$= -\gamma\Delta.$$

By optimality of x, we conclude that $\Delta = 0$, and thus that x^1 and x^2 are both optimal. $\qquad\square$

We now consider the case where all the periods share the same feasible transitions, and all the nonterminal periods share the same feasible set. Under these assumptions, Proposition 6 shows that there exists an optimal solution whose periods can be partitioned into three consecutive, possibly empty phases, covering periods $\{1, \ldots, \bar{p}\}$, $\{\bar{p} + 1, \ldots, \bar{q} - 1\}$, and $\{\bar{q}, \ldots, P\}$, respectively, where $0 \leq \bar{p} < \bar{q} \leq P + 1$. The first phase is formed of nontrivial transitions that cannot be advanced, the second phase comprises only trivial transitions, and the last phase contains nontrivial transitions that cannot be delayed.

Proposition 6. *If $\mathcal{X}_1 = \mathcal{X}_2 = \cdots = \mathcal{X}_{P-1}$ and $\mathcal{X}'_1(x_p) = \mathcal{X}'_2(x_p) = \cdots = \mathcal{X}'_P(x_p) =: \mathcal{X}'(x_p) \ \forall x_p \in \mathcal{X}_\bullet$, then there exists an optimal solution x to problem (A.22) that satisfies $x_p \notin \mathcal{X}'(x_{p-2}) \ \forall p \in \{2, \ldots, \bar{p}\} \cup \{\bar{q} + 1, \ldots, P\}$ and $x_{p_1} = x_{p_2} \ \forall \bar{p} \leq p_1 \leq p_2 \leq \bar{q} - 1$, for some $\bar{p} < \bar{q}$, $\bar{p}, \bar{q} \in \{0, \ldots, P + 1\}$.*

Proof. Proof. We consider an arbitrary optimal solution x. First, assume that x does not admit any trivial state transition, i.e., $x_p \neq x_{p-1}$, $\forall p \in \{1, \ldots, P\}$. If there is some $\bar{p} \in \{1, \ldots, P - 1\}$ such that $x_p \notin \mathcal{X}'(x_{p-2}) \ \forall p \in \{2, \ldots, \bar{p}\} \cup \{\bar{p} + 2, \ldots, P\}$, then we set $\bar{q} = \bar{p} + 1$, and x satisfies the conditions of the corollary. Otherwise, there exists $p' \leq p''$, $p', p'' \in \{2, \ldots, P - 1\}$ such that $x_{p'} \in \mathcal{X}'(x_{p'-2})$ and $x_{p''+1} \in \mathcal{X}'(x_{p''-1})$. Since $\mathcal{X}_1 = \mathcal{X}_2 = \cdots = \mathcal{X}_{P-1}$, this is sufficient to conclude that solutions x^1 and x^2, as defined in equations (A.23)–(A.24), are both feasible and, by Proposition 5, optimal. From there, we set $x = x_1$, and notice that we have recovered an optimal solution that contains a trivial transition in period p''.

Now, we assume that x is an arbitrary optimal solution that admits at least one trivial transition. Assume that x admits nonconsecutive trivial state transitions, i.e., $x_{p'-2} = x_{p'-1} \neq \ldots \neq x_{p''} = x_{p''+1}$, for some $p' \leq p''$, $p', p'' \in \{2, \ldots, P - 1\}$. Again, we consider solutions x^1 and x^2, as defined in equations (A.23)–(A.24). To show that all the transitions in solutions x^1 and x^2 are feasible, it suffices to verify their nontrivial transitions that are not present in solution x. We notice that there are no such transitions. Indeed, the only candidates are $(x^1_{p'-2}, x^1_{p'-1}) = (x_{p'-2}, x_{p'}) = (x_{p'-1}, x_{p'})$ for solution x^1, and $(x^2_{p''}, x^2_{p''+1}) = (x_{p''-1}, x_{p''+1}) = (x_{p''-1}, x_{p''})$ for solution x^2. Therefore, they are both feasible and thus optimal. We set $x = x_1$, and repeat the same procedure, where nontrivial state transitions are iteratively advanced by one period while preserving optimality, until x no longer contains nonconsecutive trivial transitions.

From there, the only remaining case that needs to be considered is that of an arbitrary optimal solution x with one or more trivial transitions, all of which take place in consecutive periods. Let $\bar{p} + 1, \ldots \bar{q} - 1$ be these periods, for $\bar{p} \leq \bar{q} - 2$, $\bar{p}, \bar{q} \in \{0, \ldots, P + 1\}$, so that $x_{p_1} = x_{p_2} \ \forall \bar{p} \leq p_1 \leq p_2 \leq \bar{q} - 1$. If there is some $p \in \{2, \ldots, \bar{p}\}$ such that $x_p \in \mathcal{X}'(x_{p-2})$, we set $p' = p$, $p'' = \bar{p}$, obtain that solutions x^1 and x^2, as defined in equations (A.23)–(A.24), are both optimal, set $x = x_1$, which now contains trivial transitions in periods $\bar{p}, \ldots \bar{q} - 1$, and return to the general case. Otherwise, if there is some $p \in \{\bar{q} + 1, \ldots, P\}$ such that $x_p \in \mathcal{X}'(x_{p-2})$, we set $p' = \bar{q}$, $p'' = p - 1$, obtain that solutions x^1 and x^2, as defined in equations (A.23)–(A.24), are both optimal, set $x = x_2$, which now contains trivial transitions in periods $\bar{p} + 1, \ldots \bar{q}$, and return to the general case. When no such p remains, we conclude that x satisfies the conditions of the corollary. $\qquad\square$

A consequence of Proposition 6 is that, assuming period-invariant service level constraints and cost coefficients, the BFEP reduces to a two-period problem when the investment budget applies to the full planning horizon and retirement targets are only imposed in the last period.

Corollary 1. *If the yearly budget constraints (1b) are replaced by a shared budget constraint $\sum_{p \in \mathcal{P}} I_p(x_p - x_{p-1}) \leq I^{UB}$, the retirement constraints (1f) apply to period P only, and the costs functions are the same*

for all periods, i.e., $I_1(\cdot) = \cdots = I_P(\cdot) =: I(\cdot)$ and $O_1(\cdot) = \cdots = O_P(\cdot) =: O(\cdot)$, then the BFEP can be reformulated as the following two-period problem:

$$\min_{x_1, x_P} \; I(x_1 - x_0) + O(x_1) \sum_{p=1}^{P-1} \gamma^{p-1} + \gamma^{P-1} \left(I(x_P - x_1) + O(x_P) \right) \tag{A.27a}$$

$$s.t. \quad I(x_P - x_0) \leq I^{UB}, \; (1c)\text{–}(1i). \tag{A.27b}$$

Proof. Proof. In this proof, we express the BFEP under formulation (A.22). By linearity of the investment cost function $I(\cdot)$, the shared budget constraint $\sum_{p \in \mathcal{P}} I_p(x_p - x_{p-1}) \leq I^{UB}$ simplifies to $I(x_P - x_0) \leq I^{UB}$, which only involves the last period state x_P. Since we assumed that the retirement constraints (1f) only apply to period P, the feasible set $\mathcal{X}_p$ of each period $p \in \{1, \ldots, P-1\}$ is thus only constrained by the upper bounds (1g) on chargers, so that $\mathcal{X}_1 = \cdots = \mathcal{X}_{P-1}$. Furthermore, since each period is subject to the same monotonicity constraints (1d)–(1e), the BFEP satisfies $\mathcal{X}'_1(\cdot) = \mathcal{X}'_2(\cdot) = \cdots = \mathcal{X}'_P(\cdot) =: \mathcal{X}'(\cdot)$.

From there, under the assumption that the investment cost coefficients are the same in each period, we can define the transition cost function of the BFEP as $\bar{T}(x_{p-1}, x_p) = I(x_p - x_{p-1})$. Since the same holds for the operational costs, we can define the state cost function of the BFEP as $S(\cdot) = O(\cdot)$, and conclude that the problem satisfies the conditions of Proposition 6.

Since the only linking constraints between periods for the BFEP are the monotonicity constraints (1d)–(1e), any feasible solution respects $x_p \in \mathcal{X}'(x_{p-2}) \; \forall p \in \{2, \ldots, P\}$. Therefore, Proposition 6 allows to conclude that there exists an optimal solution x whose critical periods satisfy $\bar{p} \leq 1$ and $\bar{q} \geq P$, implying that $x_1 = x_2 = \cdots = x_{P-1}$. Enforcing these equalities and omitting the redundant periods $p \in \{2, \ldots, P-1\}$ yields formulation A.27. $\qquad \square$

The two investment phases identified in Proposition 6 thus collapse on the first and last investment periods in the special case of the BFEP considered in Corollary 1. More generally, Corollary 2 shows that these two phases are characterized by the return on investment (ROI) of their transition costs.

Corollary 2. *Given the assumptions of Proposition 6, and that $\mathcal{X}_P \subseteq \mathcal{X}_{P-1}$, there exists an optimal solution x to problem* (A.22) *whose critical periods $\bar{p}$ and $\bar{q}$, defined as in Proposition 6, respect:*

$$(1 - \gamma)\bar{T}(x_{\bar{p}-1}, x_{\bar{p}}) \leq S(x_{\bar{p}-1}) - S(x_{\bar{p}}), \quad \text{if } \bar{p} \geq 1, \tag{A.28}$$

$$(1 - \gamma)\bar{T}(x_{\bar{q}-1}, x_{\bar{q}}) \geq S(x_{\bar{q}-1}) - S(x_{\bar{q}}), \quad \text{if } \bar{q} \leq P. \tag{A.29}$$

Proof. Proof. Let x be an optimal solution that satisfies the conditions of Proposition 6. First, assume that x does not admit any trivial state transition, i.e., $x_p \neq x_{p-1}, \; \forall p \in \{1, \ldots, P\}$, meaning that $\bar{p}$ and $\bar{q}$ must be consecutive. There are two subcases to consider. First, assume that there is a period $\hat{p} \in \{2, \ldots, P\}$ that satisfies $x_{\hat{p}} \in \mathcal{X}'(x_{\hat{p}-2})$. In this case, the condition $x_p \notin \mathcal{X}'(x_{p-2}) \; \forall p \in \{2, \ldots, \bar{p}\} \cup \{\bar{q}+1, \ldots, P\}$ implies that $\bar{p} = \hat{p} - 1$ and $\bar{q} = \hat{p}$. Let $p' = p'' = \bar{p}$. Solution x^2, as defined in equation (A.24), is feasible. By inequality (A.26), the difference Δ_2 between the objective value of solutions x^2 and x respects:

$$\Delta_2 \leq -\gamma^{\bar{p}-1} \left[(1 - \gamma)\bar{T}(x_{\bar{p}-1}, x_{\bar{p}}) + S(x_{\bar{p}}) - S(x_{\bar{p}-1}) \right].$$

By optimality of x, we conclude that $\Delta_2 \geq 0$, hence $(1 - \gamma)\bar{T}(x_{\bar{p}-1}, x_{\bar{p}}) \leq S(x_{\bar{p}-1}) - S(x_{\bar{p}})$. Similarly, taking $p' = p'' = \bar{q}$, solution x^1, as defined in equation (A.23), is feasible. Note that in the case $\hat{p} = P$, this would

not be true in general without the assumption $\mathcal{X}_P \subseteq \mathcal{X}_{P-1}$. By inequality (A.25), the difference Δ_1 between the objective value of solutions x^1 and x respects:

$$\Delta_1 \leq \gamma^{\bar{q}-2} \left[(1-\gamma)\bar{T}(x_{\bar{q}-1}, x_{\bar{q}}) + S(x_{\bar{q}}) - S(x_{\bar{q}-1}) \right].$$

By optimality of x, we conclude that $\Delta_1 \geq 0$, hence $(1-\gamma)\bar{T}(x_{\bar{q}-1}, x_{\bar{q}}) \geq S(x_{\bar{q}-1}) - S(x_{\bar{q}})$.

Now, assume that there is no period $\hat{p} \in \{2, \ldots, P\}$ that satisfies $x_p \in \mathcal{X}'(x_{p-2})$, i.e., no pair of consecutive transitions can be replaced by a single transition. If $(1-\gamma)\bar{T}(x_0, x_1) \geq S(x_0) - S(x_1)$, then we set $\bar{p} = 0$ and $\bar{q} = 1$, and the conditions are respected. Otherwise, we set $\bar{q} = \bar{p} + 1$ for:

$$\bar{p} = \max\{p \in \{1, \ldots, P\} : (1-\gamma)\bar{T}(x_{p-1}, x_p) \leq S(x_{p-1}) - S(x_p)\},$$

and the conditions are respected.

Now, we consider the case where x admits trivial state transitions. If $x_1 = x_0$, then $\bar{p} = 0$, in which case (A.28) is respected. Otherwise, $\bar{p} = \min\{p \in \{1, \ldots, P-1\} : x_p = x_{p+1}\}$ and solution x^2, as defined in equation (A.24) for $p' = p'' = \bar{p}$ is feasible. As before, by optimality of x, we conclude that $\Delta_2 \geq 0$, hence $(1-\gamma)\bar{T}(x_{\bar{p}-1}, x_{\bar{p}}) \leq S(x_{\bar{p}-1}) - S(x_{\bar{p}})$. Condition (A.28) is thus always satisfied.

If $x_{P-1} = x_P$, then $\bar{q} = P + 1$, in which case (A.29) is respected. Otherwise, $\bar{q} = \min\{p \in \{1, \ldots, P\} : x_{p-2} = x_{p-1} \neq x_p\}$ and solution x^1, as defined in equation (A.23) for $p' = p'' = \bar{q}$ is feasible. As previously, by optimality of x, we conclude that $\Delta_1 \geq 0$, hence $(1-\gamma)\bar{T}(x_{\bar{q}-1}, x_{\bar{q}}) \geq S(x_{\bar{q}-1}) - S(x_{\bar{q}})$. Condition (A.29) is thus also always satisfied. $\qquad\square$

For a discount factor $\gamma < 1$, and assuming that the transition function $\bar{T}(\cdot)$ and state function $S(\cdot)$ respectively reflect investment and operational costs, equations (A.28) and (A.29) admit an intuitive economic interpretation. They indicate that the last investments of the first phase will return at least $1 - \gamma$ times their costs in operational savings over one time period, a return that the last phase of investments does not achieve. In the context of the BFEP, this suggests that early investments will be driven by operational savings, whereas less profitable investments that are required to achieve electrification targets will be postponed to the end of the planning horizon.

We conclude this section by showing that, if the parameters of the operational problem do not change between periods, the investment costs are period-invariant and include the maintenance costs of new assets, and retirement targets are only imposed in the last period, then the optimal operational costs are nonincreasing in nonterminal periods for the BFEP.

Proposition 7. *If the retirement constraints* (1f) *apply to period P only, the investment functions satisfy $I_p(\cdot) = I(\cdot) \ \forall p \in \mathcal{P}$ and do not depend on the $\hat{\eta}_r^p$ variables, and the operational costs respect $O_p(\cdot) = H(\cdot) + \mathcal{Q}(\cdot) \ \forall p \in \mathcal{P}$, with $H(x_p) = \hat{c} \sum_{r \in \mathcal{R}} \hat{\eta}_r^p$, for some constant $\hat{c} \geq 0$, then any optimal solution to the BFEP satisfies $O(x_0) \geq O(x_1) \geq \cdots \geq O(x_{P-1})$.*

Proof. Proof. Let $(\{x_p\}_{p \in \mathcal{P}}, \{y_{pr}\}_{p \in \mathcal{P}, r \in \mathcal{R}})$ be an optimal solution for which there exists a period $p \in \{1, \ldots, P-1\}$ such that $O(x_{p-1}) < O(x_p)$. We consider a modified solution in which we use in period p the same assignment of conventional buses to routes as in period $p-1$, i.e., $\hat{\eta}_r^p = \hat{\eta}_r^{p-1} \ \forall r \in \mathcal{R}$. By the assumption that period p is not subject to stricter retirement constraints than period $p - 1$, this preserves feasibility. From there, the monotonicity constraints (1d) ensure that we can reassign depot and on-route BEBs in period p so that the number of each type of bus on each route is nondecreasing from period $p-1$ to period p, i.e., $\bar{\eta}_{rb}^p \geq \bar{\eta}_{rb}^{p-1}$ and $\tilde{\eta}_r^p \geq \tilde{\eta}_r^{p-1} \ \forall b \in \mathcal{B}, \ r \in \mathcal{R}$. Finally, the monotonicity constraints (1c) on chargers

allow us to conclude that $x_p \geq x_{p-1}$ after modification. Noting that the optimal scheduling cost $\mathcal{Q}(\cdot)$ is a nonincreasing function of the strategic variables x_p (see Remark 2), and using that $H(x_p) = H(x_{p-1})$ since $\widehat{\eta}^p = \widehat{\eta}^{p-1}$, we obtain that the operational costs of period p now satisfy $O(x_p) \leq O(x_{p-1})$. As the other terms of the objective remain unchanged, we conclude that the original solution was suboptimal. $\square$

References

Benders JF (1962) Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik* 4(1):238–252.

Fischetti M, Salvagnin D, Zanette A (2010) A note on the selection of Benders' cuts. *Mathematical Programming* 124:175–182.

Geoffrion AM, Graves GW (1974) Multicommodity distribution system design by Benders decomposition. *Management science* 20(5):822–844.

Hosseini M, Turner J (2024) Deepest cuts for Benders decomposition. *Operations Research* .

Johnson C, Nobler E, Eudy L, Jeffers M (2020) Financial analysis of battery electric transit buses. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States).

Kewcharoenwong P, Üster H (2014) Benders decomposition algorithms for the fixed-charge relay network design in telecommunications. *Telecommunication Systems* 56:441–453.

Magnanti TL, Wong RT (1981) Accelerating Benders decomposition: Algorithmic enhancement and model selection criteria. *Operations research* 29(3):464–484.

Rahmaniani R, Crainic TG, Gendreau M, Rei W (2017) The Benders decomposition algorithm: A literature review. *European Journal of Operational Research* 259(3):801–817.

Seo K, Joung S, Lee C, Park S (2022) A closest Benders cut selection scheme for accelerating the Benders decomposition algorithm. *INFORMS Journal on Computing* 34(5):2804–2827.