Renewable Energy, Environment, and Power Quality Journal (REE&PQJ)
ISSN:3020-531X
Volume 26, July 2026

Smart Charging of Urban Electric Bus Fleets Using Proximal Policy Optimization

A. Martínez[1], M. Alonso[1], H. Amaris[1] and M. Cordeiro-Costas[2]

[1] Department of Electrical Engineering
Universidad Carlos III de Madrid, 28911, Leganés (Spain),
100440775@alumnos.uc3m.es, monica.alonso@uc3m.es, hortensia.amaris@uc3m.es

[2] CINTECX, Universidade de Vigo, 36310, Vigo (Spain), moises.cordeiro.costas@uvigo.gal

Abstract. The large-scale deployment of electric buses (e-buses) requires intelligent charging hubs capable of ensuring cost-effective operation while maintaining fleet availability and preventing grid congestion. This paper presents a reinforcement learning–based energy management system for an urban e-bus logistics depot centre operating under technical and economic constraints, including uncertainty in battery state of charge and operational schedules. The charging problem is formulated as a Markov Decision Process and solved using Proximal Policy Optimization (PPO). The agent determines charger allocation and hourly energy delivery based on bus-level variables (state of charge, remaining energy, time to departure, priority) and global variables (time, accumulated depot consumption, dynamic electricity prices, and charger availability). The reward function balances timely charging, grid compliance, priority service, and cost minimization. Validation on a realistic depot with 16 chargers and 25 e-buses demonstrates that the proposed controller satisfies fleet requirements while respecting network limits and shifting demand away from high-price periods.

Key words. Smart charging, reinforcement learning, e-bus fleet charging, grid-constraint optimization, PPO

1. Introduction

The decarbonization of urban transport is a key pillar of current energy transition strategies, with electric buses (e-buses) playing a central role in reducing greenhouse gas emissions and improving air quality. However, large-scale fleet electrification introduces significant operational challenges, particularly in the coordinated management of depot charging under grid, economic, and scheduling constraints.

Recent research highlights the growing use of optimization and learning-based methods to improve electric fleet utilization and charging coordination under uncertainty [1]–[3]. Reinforcement learning (RL) approaches have demonstrated strong potential for handling high-dimensional decision spaces and stochastic operating conditions in fleet-centric applications [3], [4], while robust optimization frameworks increasingly address uncertainty and battery degradation effects [5]. Nevertheless, many existing approaches do not fully integrate economic signals, grid limitations, and fleet priorities within a unified decision-making architecture.

This paper proposes a reinforcement learning–based energy management system (EMS) for intelligent charging coordination in an urban e-bus depot centre. The problem is formulated as a Markov Decision Process and solved using Proximal Policy Optimization (PPO), enabling adaptive and stable policy learning. Simulation results demonstrate improved operational efficiency, reduced penalty and energy costs, and full compliance with power grid constraints. The main contributions are summarized as follows:

(1) the formulation of the e-bus smart charging problem as an MDP under grid, operational, and economic constraints.
(2) the development of a PPO-based EMS for coordinated charger assignment and power allocation.
(3) the application of a dedicated hyperparameter optimization strategy to improve PPO performance; and
(4) a benchmarking analysis against baseline PPO, SAC, and DDPG under a realistic depot scenario.

The content of the paper is organized as follows. Section 2 defines the problem and modelling framework. Section 3 presents the case study, PPO tuning, and performance results, including comparisons with other RL approaches. Section provides the main conclusions.

2. PPO-Based e-Bus Fleet Energy Management Framework

The energy management system (EMS) of the logistics depot is responsible for allocating charging points to electric buses and determining the charging power supplied to each vehicle. Due to the complexity and high dimensionality of the decision-making process, a PPO algorithm is adopted. PPO is a RL technique that enables stable and efficient policy updates by maximizing the expected cumulative reward while constraining policy deviations to prevent instability during training.

The charging management problem is formulated as a Markov Decision Process (MDP). At each simulation step, the RL agent interacts with an environment that represents the operation of the charging infrastructure and its interaction with the fleet. Based on the observed system state, the agent makes decisions aimed at optimizing infrastructure utilization, minimizing energy costs, and ensuring fleet availability.

The state space is defined as a vector combining bus-level and global variables. Bus-level variables include state of charge (SoC), remaining energy demand, time to full charge, charging status, bus priority level, and remaining time until departure. Global variables comprise time of day, cumulative daily energy consumption of the depot centre, hourly electricity price, and charger availability. The action space is continuous and includes the selection of the bus to be charged and the hourly energy to be delivered. A summary of the observation, action, and state space parameters is provided in Table 1.

Table 1. Definition of the variables included in the observation, action, and state spaces of the proposed RL model.

Space	Variable	Range
Observation	State of charge of bus i	[0–1]
Observation	Remaining energy demand of bus i	[0–1]
Observation	Time to full charge of bus i	[0–1]
Observation	Priority level of bus i	Binary
Observation	Remaining time until departure of bus i	[0–1]
Observation	Time of day	[0–1]
Observation	Grid power limit	[0–1]
Observation	Electricity price	[0–1]
Observation	Charger availability	Binary vector (N_ch)
Action	Bus selection	Continuous (N_ch)
Action	Charging power	Continuous (N_ch)
State	Charger assigned	Binary (N_bus)
State	Waiting for charger	Binary (N_bus)
State	In transit	Binary (N_bus)
State	Charging completed	Binary (N_bus)

The reward function integrates multiple criteria reflecting both operational and economic objectives. These include costs associated with dynamic electricity prices (eq. 1); incentives to ensure that e-buses reach the required state of charge before departure (eq. 2); penalties associated with bus waiting times or unmet charging demand (eq. 3); and penalties for exceeding the maximum power available at the grid connection point (eq. 4).

$$\boldsymbol{R_{cost}(t) = -\lambda_c \cdot p_t \cdot \sum_{i=1}^{N} P_i(t) \cdot \Delta t} \qquad \text{Eq. 1}$$

where $\boldsymbol{p_t}$ denotes the electricity price at time step $\boldsymbol{t}$, $\boldsymbol{P_i(t)}$ is the charging power assigned to bus $\boldsymbol{i}$, N is the number of buses, $\boldsymbol{\Delta t}$ is the simulation time step, and $\boldsymbol{\lambda_c}$ is the corresponding weighting factor.

$$\boldsymbol{R_{soc}(t) = -\lambda_s \cdot \sum_{i=1}^{N} \max(0, SoC_i^{req} - SoC_i(t))} \qquad \text{Eq. 2}$$

where $\boldsymbol{SoC_i(t)}$ denotes the state of charge of bus $\boldsymbol{i}$ at time step $\boldsymbol{t}$, $\boldsymbol{SoC_i^{req}}$ is the required state of charge before departure, and $\boldsymbol{\lambda_s}$ is the corresponding weighting factor.

$$\boldsymbol{R_{wait}(t) = -\lambda_w \cdot N_{wait}(t)} \qquad \text{Eq. 3}$$

where $\boldsymbol{N_{wait}(t)}$ denotes the number of buses waiting for charging at time step $\boldsymbol{t}$, and $\boldsymbol{\lambda_w}$ is the corresponding penalty weight.

$$\boldsymbol{R_{grid}(t) = -\lambda_g \cdot \left[\max\left(0, \sum_{i=1}^{N} P_i(t)\right) - P_{grid}^{max}\right]^2} \qquad \text{Eq. 4}$$

where $\boldsymbol{P_{grid}^{max}}$ is the maximum grid connection power, and $\boldsymbol{\lambda_g}$ is the corresponding penalty weight.

The overall reward is defined as the weighted combination of the previously described components (eq. 5).

$$\boldsymbol{R(t) = R_{cost}(t) + R_{soc}(t) + R_{wait}(t) + R_{grid}(t)} \qquad \text{Eq. 5}$$

The grid is modelled through a set of aggregated power constraints at the connection point. A maximum power limit of 2,200 kW is imposed at the point of common coupling, representing the capacity of the grid connection. Operational constraints associated with the charging infrastructure and the e-buses are considered. Each charging point is subject to a maximum power rating, which limits the instantaneous charging power per vehicle, while the total number of available chargers restricts the number of e-buses that can charge simultaneously. Furthermore, each e-bus is modelled with charging and discharging power limits, as well as SoC bounds, ensuring physically consistent battery operation.

Algorithm 1 summarizes the proposed PPO-based EMS, including state observation, action selection, constraint handling, and policy update.

Algorithm 1. PPO-based EMS for e-bus charging coordination

Input: Fleet data, charger data, grid constraints, PPO hyperparameters
Output: Charging assignment and power schedule
1. Initialize simulation environment
2. Initialize PPO agent (policy πθ and value function Vφ)
3. For each training episode:
  3.1 Reset environment
  3.2 Observe initial state s_0
  3.3 For each time step t:
    a) Build observation vector o_t including:
      - Bus variables: SoC_i, energy $demand_i$, time-to-$charge_i$, $priority_i$, time-to-$departure_i$
      - Global variables: time, energy consumption, electricity price, grid limit, charger availability
    b) Select action a_t:
      - Bus assignment per charger
      - Charging power per charger
    c) Apply constraints:
      - Charger availability
      - Charger power limits
      - Grid power limit
      - Battery SoC constraints
    d) Execute action in environment
    e) Compute reward:
      R(t) = R_cost + R_soc + R_wait + R_grid
    f) Observe next state s_{t+1}
    g) Store transition (s_t, a_t, R(t), s_{t+1})
    h) If update condition is met:
      - Compute advantages
      - Update PPO policy (clipped objective)
      - Update value function
    i) If terminal condition: break
Evaluate trained model

To assess the effectiveness of the proposed approach, a benchmarking framework is defined including multiple learning-based strategies. The performance of the optimized PPO configuration is compared against a non-optimized PPO setup, corresponding to the default hyperparameter configuration, as well as 2 additional RL algorithms, namely soft actor critic (SAC) and deep deterministic policy gradient (DDPG). The comparison is conducted in terms of energy cost and computational cost.

The energy cost reflects the operational performance of each strategy, while the computational cost accounts for the resources required during the training process.

3. Results

A. E-bus system description and operating conditions

For the validation of the proposed energy management system, a logistics depot centre is considered as the case study. The charging infrastructure comprises a total of 16 chargers distributed across eight different types, enabling the realistic representation of technological diversity in slow, fast, and ultra-fast charging systems. The technical specifications of these chargers are presented in Table 2.

Table 2. Power ratings of charging infrastructure

	Power rate (kW)
Slow charging	40
Medium charging	140
Fast charging	200
Ultra-fast charging	450

The electricity prices considered in the study correspond to real market data from Spain [6], enabling a realistic representation of time-varying energy costs. The training process is conducted over a one-month period, capturing different operational patterns across weekdays, Saturdays, and Sundays and holidays, each characterized by distinct demand and mobility behaviours. The performance of the trained models is then evaluated on representative days from these categories, allowing the assessment of the robustness and generalization capability of the proposed approach under varying operating conditions.

The fleet managed by the charging strategy consists of 25 electric buses with a battery capacity of 324 kWh. The e-bus fleet covers 1 route with a daily distance of 180 km. The frequency of service varies throughout the day, resulting in different bus circulation patterns and charging opportunities. This variability is illustrated in Fig. 1

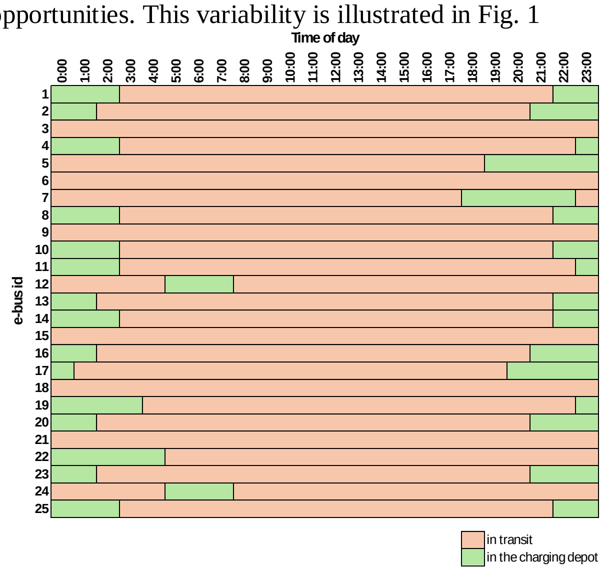

Fig. 1. Temporal distribution of e-bus operational states.

Fig. 1 highlights the variability in bus availability for charging, distinguishing between periods when vehicles are in transit and when they are located at the charging depot. As a result, charging opportunities are not evenly distributed over time or across vehicles, adding complexity to the energy management problem.

The route has medium demand and is modelled using a time-dependent service schedule. The frequency of service varies throughout the day, resulting in different bus circulation patterns and charging opportunities. From this schedule, key operational variables are derived on an hourly basis, including the dwell time at charging locations, the initial SoC of each e-bus upon arrival, and the number of EVs charging simultaneously. These variables are summarised in Table 3.

Table 3. Descriptive statistics of operational e-bus variables.

	Min	Max	Avg
Dwell time (h)	1	9	4.41
Initial SoC (%)	5.69	44.30	31.68
Simultaneous charging buses	0	16	7.78

B. PPO tuning

The hyperparameter optimization of the PPO agent is carried out using the Optuna framework, which enables an efficient exploration of the search space. The selected hyperparameters and their corresponding ranges are presented in Table 3. These ranges are defined based on commonly adopted values in the literature, while allowing sufficient flexibility to capture the specific dynamics of the problem. The search space includes key learning parameters such as the learning rate (α), discount factor (γ), and GAE parameter (λ), as well as architectural choices related to the policy and value networks, including the number of layers and neurons. In addition, optimization-related parameters such as batch size, rollout length, clipping parameter (ε), gradient norm, and loss coefficients (c1 and c2) are considered. The explored ranges are selected to ensure a trade-off between stability and adaptability of the learning process, avoiding excessively narrow configurations that could limit performance.

Table 4. Hyperparameter search space for PPO optimization.

Parameter	Range
Learning rate (α)	[0.00001-0.0003]
Discount factor (γ)	[0.95-0.99]
Batch size	[32-128]
Clip parameter (ε)	[0.10-0.25]
Value function (c_1)	[0.3-0.7]
Entropy (c_2)	[0.0001-0.01]
GAE lambda (λ)	[0.90-0.98]
Rollout	[512-2048]
Gradient	[0.5-2.5]
Policy Layers	[2-3]
Policy Neurons	[64-128]
Value Layers	[2-3]
Value Neurons	[64-128]

The weighting factors were selected to ensure a balanced contribution of economic, operational, and technical objectives, avoiding over-prioritization of any single reward component. These are presented in Table 5.

Table 5. Weighting parameters of the reward function.

	Value
λ_c	5
λ_s	50
λ_w	20
λ_g	5

C. Optimal PPO configuration

The selected configuration provides a balanced trade-off between training stability, exploration capability, and critic accuracy. The optimal setup includes a low-to-moderate learning rate (2.92×10^{-5}), a relatively high value function coefficient (0.795), an intermediate rollout length (512 steps), and a reduced but non-zero entropy coefficient (0.008), promoting controlled exploration. The adopted neural network architecture presents medium complexity, with two hidden layers of 64 neurons each ([64, 64]). Table 6 summarises the performance of the different hyperparameter configurations evaluated, including the comparison with the baseline setup.

Table 6. Performance comparison of optimal PPO configuration and baseline.

Parameter	Optimal	Baseline
Learning rate (α)	0.0000292	0.0003
Discount factor (γ)	0.9256	0.99
Batch size	32	64
Clip parameter (ε)	0.237	0.2
Value function (c_1)	0.795	0.5
Entropy (c_2)	0.008	0.0
GAE lambda (λ)	0.9388	0.95
Rollout	512	2048
Gradient	2.12	0.5
Policy Layers	2	2
Policy Neurons	64, 64	64, 64
Value Layers	2	2
Value Neurons	64, 64	64, 64

The PPO agent maximizes the clipped surrogate objective, constraining policy updates to ensure stable learning. The total loss combines actor loss, critic loss, and entropy regularization to balance exploitation and exploration (Fig. 2).

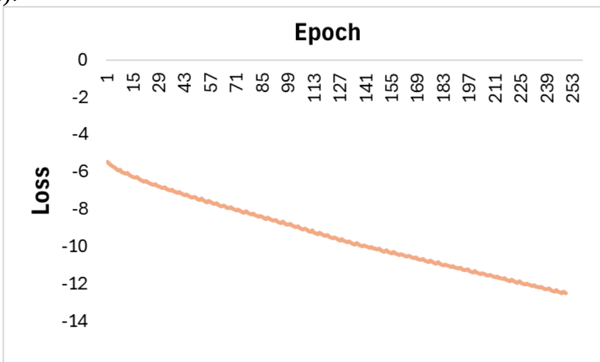

Fig. 2. Total losses actor-critic.

Fig. 3 shows the evolution of the average reward during training. In epoch 25, the loss curve begins to stabilize, indicating that the agent starts identifying effective charging strategies. From epoch 80 onward, the loss converges to values close to -100 with minimal fluctuations, suggesting a stable training process and policy convergence. The reduced variance in later epochs reflects increased consistency in decision-making and improved robustness of the learned policy.

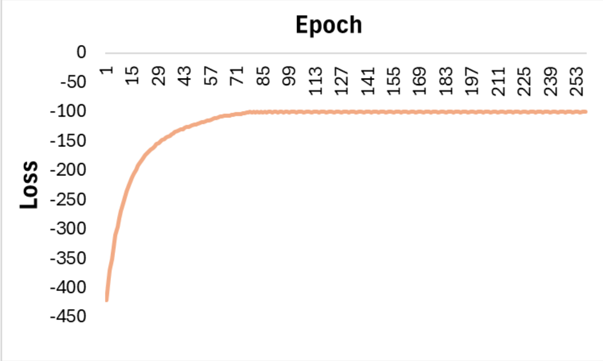

Fig. 3. Average reward during the training phase.

C. Results

The EMS coordinates charger allocation and charging schedules according to fleet requirements and maximum energy available at the power grid connection point. Fig. 4 illustrates the temporal utilization of the charging infrastructure, showing the connection status of each charger throughout the day. The charger usage is concentrated in specific time intervals, particularly when a higher number of e-buses are available at the depot. During daytime hours, a significant number of chargers remain unused, which is consistent with fleet operation patterns (Fig. 1) where vehicles are predominantly in transit. In contrast, higher utilization is observed during early morning and late evening periods, when more vehicles are present at the depot and available for charging.

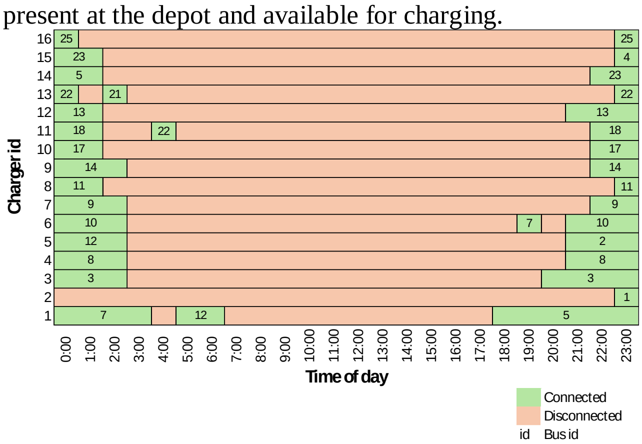

Fig. 4. Temporal evolution of charger usage and bus assignment throughout the day

Fig. 5 shows the aggregated charging power profile over a day, disaggregated by individual chargers. The results indicate that charging activity is concentrated during early hours and late evening periods, when a higher number of e-buses are available at the depot. In contrast, limited charging activity is observed during daytime hours, as most vehicles are in transit. The charging schedule is primarily driven by vehicle availability rather than solely by electricity price signals. Nevertheless, these periods of higher charging activity often coincide with lower electricity prices, enabling the strategy to take advantage of economically favorable conditions whenever operational constraints allow it. The total power demand remains below the maximum grid capacity, represented by the upper limit line, confirming that the grid constraint is consistently satisfied. Additionally, the distribution of power across chargers reflects a coordinated allocation of resources, contributing to a balanced utilization of the charging infrastructure.

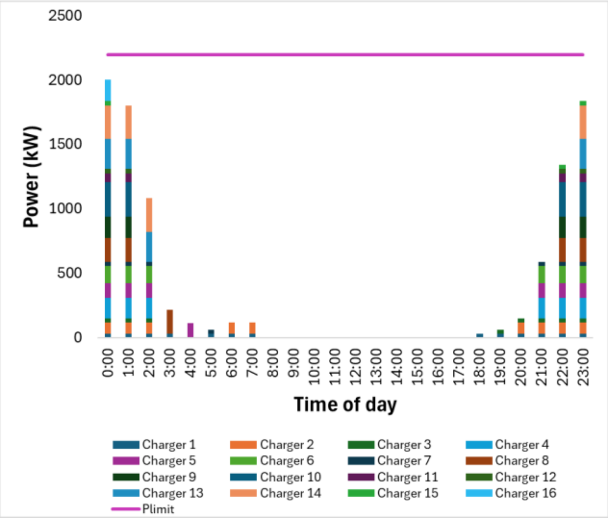

Fig. 5. Daily charging power profile by charger with grid power limit.

Table 7 compares the performance of the optimized PPO against standard PPO, SAC, and DDPG. The results show that the optimized PPO achieves the lowest operational cost (2565.97 €) and the shortest computation time (77.23 s), slightly outperforming the standard PPO in both metrics. However, this improvement comes at the expense of a marginal increase in energy consumption compared to PPO. In contrast, SAC and DDPG exhibit significantly higher costs and energy usage, highlighting a clear trade-off between optimization efficiency and energy consumption, where the optimized PPO provides the best overall balance. These results confirm that PPO, particularly when properly tuned, is the most suitable algorithm for the case study, providing the best trade-off between economic and operational performance.

Table 7. Performance comparison of RL algorithms.

	Comp time	Elect cost	Energy consumed
Opt PPO	77.23 s	2565.97 €	147959.04 kWh
PPO	80.03 s	2572.46 €	147374.75 kWh
SAC	87.46 s	3728.18 €	200557.71 kWh
DDPG	83.78 s	3607.77 €	196472.70 kWh

4. Conclusions

This work presents a reinforcement learning–based energy management system for intelligent charging coordination of electric bus fleets in logistics depots. The proposed MDP formulation, solved using PPO, enables adaptive decision-making under technical, economic, and operational uncertainty.

The results demonstrate that the proposed optimization methodology significantly enhances PPO performance, achieving the lowest operational cost and computational time among the evaluated algorithms. While a slight increase in energy consumption is observed compared to the baseline PPO, the overall cost-efficiency improvement validates the effectiveness of the proposed tuning strategy. Furthermore, the comparison with SAC and DDPG highlights that PPO is the most suitable algorithm for the case study, providing the best trade-off between economic and operational performance.

The developed controller successfully satisfies fleet energy requirements while respecting grid constraints and minimizing cost-related penalties. In addition, the integrated architecture, combining simulation, data management, and RL-based control, ensures modularity and scalability.

Acknowledgements

This work was supported by grant PID2024-157191OB-C21 funded by Agencia Estatal de Investigación (AEI).

References

[1] J. Guanetti, Y. Kim, X. Shen, J. Donham, S. Alexander, B. Wootton and F. Borrelli, "Increasing Electric Vehicles Utilization in Transit Fleets using Learning, Predictions, Optimization, and Automation," *arXiv preprint arXiv:2305.14732*, 2023.

[2] J. Fan, H. Wang and A. Liebman, "MARL for Decentralized Electric Vehicle Charging Coordination with V2V Energy Exchange," *arXiv preprint arXiv:2308.14111*, 2023.

[3] J. Qi, L. Lei, T. Jonsson and D. Niyato, "Optimizing Electric Bus Charging Scheduling with Uncertainties Using Hierarchical Deep Reinforcement Learning," *arXiv preprint*, 2025.

[4] P. Michailidis, T. Stergiou and G. Ingramidis, "Reinforcement Learning for Electric Vehicle Charging Systems: Fleet-Centric Approaches," *Energies*, vol. 18, 2025.

[5] J. A. Manzolli, J. P. F. Trovão and C. H. Antunes, "Electric Bus Fleet Charging Management: A Robust Optimisation Framework Addressing Battery Ageing and Uncertainty," *Energy*, 2025.

[6] Red Eléctrica de España. Markets and prices. e·sios [online] (2026).