


Article

Reinforcement Learning for Transit Signal Priority with Priority Factor

Hoi-Kin Cheng ^{1,*}, Kun-Pang Kou ¹ and Ka-Io Wong ² ¹ Department of Civil and Environmental Engineering, University of Macau, Macao, China; kpkou@um.edu.mo² Department of Transportation and Logistics Management, National Yang Ming Chiao Tung University, Hsinchu City 300093, Taiwan; kiwong@nycu.edu.tw

* Correspondence: yb97479@um.edu.mo

Highlights:**What are the main findings?**

- **Effective Reduction of Bus Travel Times:** The study's proposed transit signal priority (TSP) control system that incorporates a priority factor (PF) significantly reduces average travel times for buses—17% in the arterial network and 25% in the grid network.
- **Mitigation of Conflicting Requests:** The RL-based TSP with PF effectively resolves conflicting priority requests when multiple buses approach an intersection from different directions.
- **Minimal Impact on Passenger Cars:** While the average travel time for passenger cars increased (by up to 12% in the arterial network and 7% in the grid network), the impact is not significant compared to the improvements in bus travel times.
- **Dynamic Assignment of PF:** The PF can be dynamically assigned based on the number of passengers on each bus, enhancing the system's adaptability to varying traffic conditions.
- **Longer Cycle Length and Signal Splits:** An increase in buses with higher PF leads to longer cycle lengths and extended signal phase durations for directions with buses.

What is the implication of the main finding?

- **Improved Public Transit Efficiency:** The findings suggest that integrating advanced control strategies like RL and dynamic PFs can enhance the efficiency of public transportation systems, making them more appealing and potentially increasing ridership.
- **Balanced Traffic Management:** The study indicates a feasible approach for balancing the needs of both transit and non-transit users at signalized intersections, which is crucial for urban mobility and reducing overall congestion.
- **Scalability and Adaptability:** The ability to dynamically adjust the PF based on passenger load suggests that this approach can be tailored to different traffic conditions and demands, making it adaptable for various urban environments.
- **Future Research Directions:** The findings underscore the need for further exploration into real-time PF determination and the integration of additional factors (like passenger waiting times) to enhance the operational efficiency and reliability of transit systems.

Abstract: Public transportation has been identified as a viable solution to mitigate traffic congestion. Transit signal priority (TSP) control, which is widely used at signalized intersections, has been recognized as a practical strategy to improve the efficiency and reliability of bus operations. However, traditional TSP control may fall short of efficiency and is facing several challenges of negative externalities for non-transit users and the need to handle conflicting priority requests. Recent studies have proposed the use of reinforcement learning (RL) methods to identify efficient traffic signal control (TSC). Some of these studies on RL-based TSC have incorporated the concept of max-pressure (MP), which is a maximal weight-matching algorithm to minimize queue sizes. Nevertheless, the existing RL-based TSC methods focus on private vehicles and cannot adequately distinguish between buses and private vehicles. In prior research, RL-based control has been implemented within the context of bus rapid transit (BRT) systems. This study proposes a novel RL-based TSC strategy that



Citation: Cheng, H.-K.; Kou, K.-P.; Wong, K.-I. Reinforcement Learning for Transit Signal Priority with Priority Factor. *Smart Cities* **2024**, *7*, 2861–2886. <https://doi.org/10.3390/smartcities7050111>

Academic Editor: Pierluigi Siano

Received: 1 September 2024

Revised: 2 October 2024

Accepted: 3 October 2024

Published: 6 October 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

leverages the MP concept and extends it to incorporate TSP control. This is the first implementation of RL-based TSP control within the mixed-traffic road network. A significant innovation of this research is the introduction of the priority factor (PF), which is designed to prioritize bus movements at signalized intersections. The proposed RL-based TSP with PF control seeks to balance the competing objectives of enhancing bus operations while mitigating adverse impacts on non-transit users. To evaluate the performance of the proposed TSP method with the PF mechanism, simulations were conducted on an arterial and a grid network under dynamic traffic conditions. The simulation results demonstrated that the proposed TSP with PF not only reduces bus travel times and resolves conflicts between priority requests but also does not make a significant negative impact on passenger car operations. Furthermore, the PF can be dynamically assigned according to the number of passengers on each bus, suggesting the potential for the proposed approach to be applied in various traffic management scenarios.

Keywords: transit signal priority; reinforcement learning; traffic signal control; max-pressure control; priority factor

1. Introduction

Public transportation has been widely accepted as one of the most effective strategies to mitigate traffic congestion. To improve bus service levels and attract more passengers, many strategies have been developed, such as transit signal priority (TSP) [1]. To improve the operational regularity and punctuality of transit systems, transportation researchers and transit agencies have devoted great efforts to the development of advanced transit systems during the past decades. TSP control is an operational strategy that facilitates the movement of transit vehicles through signalized intersections. TSP can be categorized into passive priority, active priority, and adaptive priority. TSP systems encounter various challenges, including the negative externalities imposed on non-transit users and the necessity to manage conflicting priority requests. An effective TSP strategy minimizes interruptions to other traffic while attempting to provide priority to transit vehicles and addresses the problem of conflicting TSP requests [1,2].

Although optimization-based methods have been extensively studied, reinforcement learning (RL) based control and max-pressure (MP) control have received significant attention in recent years [3]. RL involves an agent that learns how to interact dynamically with its environment to achieve a long-term maximum reward. Studies [4–6] used RL to obtain efficient traffic signal control (TSC). They applied RL to an isolated intersection, a multi-intersection network, and a large-scale network. Those studies proved that RL-based TSC is better than conventional deterministic traffic management systems. MP, on the other hand, is a new approach to signal timing with mathematically proven network throughput properties [7]. Studies, such as [3,8,9], proposed MP control and worked on incorporating more realistic assumptions and various situations. They considered phase switching loss, oversaturated networks, and a signal cycle constraint. However, most RL-based TSC studies designed rewards and states in a heuristic way, whereas MP control selects phases in real time based on the most recent traffic conditions and is not seeking long-term optimization. To avoid the heuristic design of the RL element and to address the disadvantages of MP, recent studies propose to connect RL with MP. Wei et al. designed an RL-based TSC well supported by the theory in MP control and proposed a well-designed RL approach to solve the network-level TSC problem [10,11].

RL methods can bring the TSC research into a new frontier [12]. Nevertheless, few of the existing studies that applied RL to TSC addressed transit [13]. Those studies mainly focus on the efficient control of private vehicles without considering the difference between transit vehicles and private vehicles in a mixed-traffic environment. Even when accounting for the differences between regular traffic and public transit, existing studies do not specifically target TSP or implement solutions solely within the context of bus rapid transit

(BRT) systems, which have dedicated bus lanes that are separated from general traffic. On the other hand, the standard MP controller is more likely to actuate phases during high-demand approaches, which may end up ignoring the arrival of public transit [14].

Therefore, this study aims to fill this research gap in public transit. The goals of this study are as follows:

- First, develop a TSP control method that utilizes RL-based strategies in a mixed-traffic environment.
- Second, provide a comprehensive theoretical framework for designing the key components of the RL-based system, including the state, action, and reward structures.
- Third, address the negative externalities experienced by non-transit users and the necessity of effectively managing conflicting priority requests.

By integrating the principles of RL with TSP, the proposed method aims to enhance the operational efficiency of transit systems. This approach addresses the limitations of previous studies that utilized heuristic designs for reward and state configurations. The theoretical design emphasizes the dynamic interaction between the RL agent and its environment, allowing for real-time adjustments based on traffic conditions. This is particularly significant given the recent advancements in RL and MP control, as it seeks to optimize control strategies beyond traditional deterministic methods. The framework presented in this paper serves as a foundation for developing more adaptive and responsive traffic management systems that can effectively balance the needs of both transit and private vehicles.

To mitigate the negative externalities of TSP, this study introduces the PF as a mechanism to prioritize bus movements over other vehicles at signalized intersections in a mixed-traffic environment. The PF is designed to enhance the effectiveness of the TSP control method by assigning a higher priority to buses, thereby facilitating their movement through intersections while minimizing disruptions to general traffic flow. This innovative approach addresses the challenges identified in the existing literature, where traditional TSP methods often fail to adequately differentiate between transit vehicles and private vehicles in mixed-traffic environments. By dynamically adjusting the PF based on real-time conditions, such as the number of passengers on each bus or current traffic demand, the proposed mechanism not only improves the reliability of bus services but also mitigates the negative externalities associated with TSP for non-transit users. Furthermore, this prioritization mechanism aligns with contemporary transportation goals of increasing public transit ridership and enhancing overall transit system performance, making it a significant contribution to the field of traffic signal control.

The experimental environment for this research utilized VISSIM, a state-of-the-art microscopic multi-modal traffic flow simulation platform, to obtain compelling empirical results. The proposed RL-based TSP method was evaluated within both an arterial network and a grid network. The results of this study clearly demonstrate several key contributions of the proposed TSP control approach:

1. **Reduction in Bus Travel Times:** The method significantly decreases bus travel times while maintaining minimal negative impacts on passenger car operations, thereby enhancing overall traffic efficiency.
2. **Conflict Resolution:** The approach effectively addresses and resolves conflicting priority requests between buses through the implementation of the PF, ensuring smoother transit operations at signalized intersections.
3. **Dynamic Assignment of Priority Factor:** The research reveals that the PF can be dynamically assigned to individual buses based on their passenger loads. This adaptability indicates the potential for the proposed method to be tailored to various traffic management scenarios, accommodating the specific needs of different transit agencies.
4. **Examines signal timing:** the paper examines signal timing within the framework of the proposed RL-based TSP control method. The analysis reveals that an increase in the number of buses assigned a higher PF is associated with longer cycle lengths. Furthermore, with respect to signal splits, the duration of the traffic signal phases is

extended when servicing directions that include buses with PF. This outcome aligns with expectations because the RL agent prioritizes the swift passage of buses through intersections. Examining signal timing provides insights into the control behavior of RL agents.

2. Literature Review

2.1. Transit Signal Priority (TSP)

TSP is an operational strategy that facilitates the movement of in-service transit vehicles through traffic-signal-controlled intersections. The objectives of TSP include improving schedule adherence and improved transit travel time efficiency while minimizing impacts on normal traffic operations. Particularly, TSP control, developed in the late 1960s, has been recognized as one of the most promising ways to reduce transit travel time at local arterials [15]. By reducing transit vehicles' delay time at intersection queues, TSP can reduce transit delay and travel time and improve transit service reliability, thereby enhancing transit quality of service. It has the potential for reducing overall delay at the intersection on a per-person basis. At the same time, TSP attempts to provide these benefits with a minimum impact on other facility users, including cross-traffic and pedestrians [16]. Therefore, a proper TSP strategy can effectively reduce transit travel time at local arterials without significantly bringing negative impacts to the network [17].

TSP strategies are mostly categorized into the following three categories: passive priority, active priority, and traffic-adaptive priority [18]. Passive priority operates continuously, regardless, based on knowledge of transit route and ridership patterns, and does not require a transit detection/priority request generation system. Active priority strategies provide priority treatment to a specific transit vehicle following detection and subsequent priority request activation. Traffic-adaptive or real-time priority is a strategy that considers the trade-offs between transit and traffic delay and allows graceful adjustments of signal timing by adapting the movement of the transit vehicle and the prevailing traffic condition [17]. Lin et al. [15] divided active priority into rule-based and model-based approaches. The rule-based priority approach grants the signal priority based on the actual presence of transit at the signalized junctions without considering transit conditions or with specific criteria of transit vehicles. Model-based priority approach grants priority to specific transit in terms of transit conditions and traffic conditions; it aims to reduce bus passenger delay or total person delay. This method needs to detect transit locations, transit operations conditions, and traffic volume/queue length of cars.

In general, passive-priority strategies can be an efficient form of TSP when transit operations are predictable with a good understanding of routes, passenger loads, schedule, and/or dwell times. It is used to implement TSP by predetermining signal timings without explicitly recognizing actual bus presence. However, it cannot handle a dynamic traffic environment. Active-priority strategies take advantage of bus detection sensors located upstream of intersections. When a transit vehicle approaches the sensors, the signal controller would provide the designated TSP strategies, such as green extension and red truncation. Nevertheless, active priority may fall short of efficiency and bring negative impacts to non-priority approaches due to overusing priority grants in high transit demand. Adaptive-priority strategies use real-time data to adjust traffic signal timings in response to changing traffic conditions. Adaptive priority is a very sophisticated and complex system and, therefore, not yet common. Table 1 summarizes the TSP types and key characteristics.

TSP systems face several challenges, such as negative externalities for non-transit users and the need to handle conflicting priority requests [19]. Negative externalities refer to the negative impacts that TSP has on non-transit users of the road network. The negative externalities of TSP result in increased waiting times for non-transit users at signalized intersections, especially on the cross streets. It also leads to increased queue lengths and spillbacks for non-transit users, which can block upstream intersections or reduce green time for other phases. Conflicting priority requests between buses can occur when multiple buses request priority simultaneously [20]. It is necessary to measure the priority

level of transit to maximize intersection efficiency. Ma et al. [21] formulated a dynamic programming model to optimize the serving sequence for multiple priority requests as well as the corresponding signal timing plans under various levels of bus occupancy, schedule deviation, and traffic demand. The results showed that the dynamic programming model outperforms the first-come-first-serve policy in terms of reducing bus delays, improving schedule adherence, and minimizing the impacts on other vehicular traffic. Xu et al. [22] proposed a bi-level optimization model to resolve the problem of conflicting TSP requests at arterial corridors. The upper level of the model is progression control and the lower level is intersection control. At the progression-control level, the characteristics of bus operation are considered. At the intersection-control level, a linear programming model is developed to resolve conflicting TSP requests.

Table 1. Summary of TSP key characteristics.

TSP Types	Advantages	Disadvantages	Control Strategy
Passive	Does not require a transit detection system and is easy to implement	Not suitable for high fluctuation traffic situation	Predetermining signal timings
Active	Considers transit conditions or specific criteria	May fall short of efficiency under high transit volumes situation	Adjustments of signal timing; common strategies are green extension and red truncation
Adaptive	Considers trade-offs between transit and traffic conditions	Sophisticated and complex systems	Signal control algorithm that provides priority while explicitly considering the impacts on the rest of the traffic

In the existing literature on transit signal priority (TSP) methods, recent studies have proposed novel approaches to enhance the efficiency and reliability of bus operations. In previous studies, max-pressure control only considered the private vehicle network. Xu et al. [14] combined MP control with TSP for the first time. That study proposed a modified MP control policy considering the TSP of the bus rapid transit system to improve the scope of the application of the MP control policy. With the development of technology, connected vehicle technology can be used to improve transit signal priority by providing real-time data on the location and speed of transit vehicles. This data can be used to adjust traffic signal timings in real time, allowing transit vehicles to move through intersections more quickly and efficiently. Zeng et al. [23] proposed a TSP system with connected-vehicle technology named Route-based TSP. That system uses enriched bus data enabled by connected-vehicle technologies to optimize signal timing and improve bus service reliability via transit signal priorities. The Route-based TSP model is formulated based on route-level transit signal priority, which is more beneficial than providing bus priority on a signal-by-signal basis.

2.2. Reinforcement Learning (RL)-Based Traffic Signal Control

RL is learning what to do—how to map situations to actions—so as to maximize a numerical reward signal. The essence of RL is learning through interaction [24]. An agent interacts with its environment and, upon observing the consequences of its actions, can learn to alter its behavior in response to rewards received. This paradigm of trial-and-error learning has its roots in behaviorist psychology and is one of the main foundations of RL. One of the early breakthroughs in RL was the development of an algorithm known as Q-learning. The Q-learning algorithm estimates the expected reward for each action in each state. The agent uses these estimates to choose the best action to take in each state. While RL has enjoyed some successes in the past, earlier approaches were often limited by a lack of scalability and were inherently constrained to relatively low-dimensional problems. However, the recent rise of deep learning has provided researchers with powerful tools

to overcome these limitations. Deep learning enables RL to scale to problems that were previously intractable, such as learning to play video games directly from pixels [25].

The existing literature on TSC has explored the application of RL as a promising approach to search for more efficient control policies. The fundamental premise of RL is the trial-and-error learning process, where agents iteratively modify signal plans and learn from the observed outcomes. For instance, [4] proposed a Q-learning-based RL method for adaptive traffic control at isolated intersections. Meanwhile, [5] introduced a multi-agent RL framework, where a central agent and outbound agents collaborate to control a network of five intersections. The outbound agents provide relative traffic flow values as local traffic statistics to the central agent, which incorporates this information into its decision-making process. Furthermore, [6] presented a fully scalable and decentralized multi-agent RL algorithm with deep learning for adaptive TSC, evaluating the proposed approach in both a large synthetic traffic grid and a real-world traffic network in Monaco. The results demonstrated the optimality, robustness, and efficiency of the proposed method.

To address the potential limitations of the heuristic design of RL elements, recent studies have explored the integration of RL with the max-pressure (MP) control theory. MP has been an active research topic in TSC, with proven properties of maximizing the throughput of the traffic network. PressLight [10] connected RL with MP, where the reward function is well-supported by the theoretical foundations of MP, which can be proved to be maximizing the throughput of the traffic network. Through comprehensive experiments, that study demonstrated that the proposed method outperforms both conventional transportation approaches and existing learning-based methods. Furthermore, CoLight [11] introduced a graph-attentional network-based model to facilitate communication and cooperation between traffic signals. This approach not only incorporates the temporal and spatial influences of neighboring intersections but also enables index-free modeling of the neighboring intersections. These advancements in the integration of RL with MP and the development of decentralized, scalable multi-agent RL frameworks demonstrate the potential of RL-based approaches to address the complexities and challenges in traffic signal control optimization.

2.3. TSP with RL

Noaen et al. [13] conducted a systematic literature review to dissect all the existing research that applied RL in the network-level TSC covering 160 articles. Among 160 articles, only 2 of them addressed public transit. Chanloha et al. proposed a CTM–BRT framework applied to a road network and compared the obtained data from the real scenarios and the Q-learning [26]. The result outperformed preemptive priority (giving priority to the bus whenever approaching an intersection together with constraints) and differential priority (giving priority to the late bus only) control methods because of the improved awareness of the signal switching cost. Shabestrav et al. [27] proposed a multimodal deep RL-based traffic signal controller that combines both regular traffic and public transit and minimizes the overall travelers' delay through the intersection. Cheng et al. apply deep RL and adopt the pressure concept for TSP control [28].

2.4. Summary

This comprehensive literature review has meticulously examined the current state of knowledge at the intersection of TSP and RL-based TSC optimization. The main objectives of TSP are to improve schedule adherence and transit travel-time efficiency while minimizing impacts on normal traffic operations. However, TSP faces challenges such as negative externalities for non-transit users and the need to manage conflicting priority requests among buses.

Recent advancements in RL have become a promising approach for TSC optimization. To address potential limitations of heuristic RL design, recent studies have integrated RL with the MP control theory. This integration, along with the development of decentralized, scalable multi-agent RL frameworks, demonstrates the potential of RL-based approaches to address the complexities of TSC optimization.

However, the existing body of research has predominantly focused on the efficient control of private vehicles, while underexploring the distinct considerations and challenges presented by the integration of transit vehicles. A systematic literature review conducted by Noaen et al. [13] showed that among 160 articles, only 2 of them applied RL with public transit. Furthermore, these two articles did not advance the application of TSP within mixed-traffic road networks. This gap in the scholarly discourse represents a compelling opportunity for further research and investigation.

To address the identified gap, this study seeks to develop a TSP control method that employs RL-based strategies within a mixed-traffic environment. In order to mitigate the negative externalities experienced by non-transit users and effectively manage conflicting priorities, this research also aims to introduce an innovative mechanism within the RL-based TSP control framework. By addressing these challenges, this study endeavors to contribute to the advancement of knowledge in this critical area of transit signal management.

3. Methodology

This research introduces the PF as an innovative mechanism within an RL-based control framework to prioritize bus movements at signalized intersections in the context of TSP in mixed traffic. The PF facilitates dynamic adjustments based on real-time conditions, such as passenger loads, enabling a more responsive and effective management of bus traffic. This novel contribution enhances the traditional TSP framework by ensuring that buses receive the necessary priority while minimizing disruptions to general traffic flow. Furthermore, it addresses the challenges of conflicting priority requests and improves overall transit efficiency, showcasing the potential of integrating RL-based control with advanced priority mechanisms.

3.1. State, Action and Reward Design

TSC and TSP are two concepts that are related to the management of traffic flow. TSC refers to the use of traffic signals to regulate the flow of traffic at intersections. TSP, on the other hand, is a strategy that gives priority to transit vehicles at signalized intersections. An RL problem involves the interaction between a learning agent and its environment in terms of states, actions, and rewards. TSC and TSP can be regarded as an RL problem because they involve detecting traffic (states) and relying on this information to seek an efficient schedule for traffic signal settings (actions) at intersections, with the goal of maximizing traffic flow while considering various factors (rewards).

RL is the process of creating an agent that learns to make decisions based on feedback from its environment. The agent is a learner and decision-maker, whereas the environment is everything outside the agent, which includes state, action, and reward design. The state is a representation of the environment at a particular time. It includes all the information that is relevant to the agent's decision-making process. The action is a decision made by the agent in response to a state. The agent's goal is to learn a policy that maps states to actions in a way that maximizes the expected return. The reward is a scalar feedback signal that indicates how well the agent is doing at a particular time step. The goal of the agent is to learn a policy that maximizes the expected cumulative reward over time.

3.1.1. State Design

In this study, the state includes three key elements: the number of vehicles on each lane, the current phase, and the elapsed time of the current phase. Let's explore these elements in more detail.

The network operates in a slotted time $t \in \{0, 1, 2, \dots\}$. Let $x_l(t)$ be the number of vehicles of lane l at time t . The number of vehicles on lane l from the current time t to the next time $t + 1$ satisfies the following:

$$x_l(t+1) = x_l(t) - \min[x_l(t), s_l(t)c_l(t)] + \min \left[\sum_{k \in In(l)} \min [s_k(t)c_k(t)r_{(k,l)}(t), x_k(t)r_{(k,l)}(t)] + d_l(t), x_{max,l} - x_l(t) \right] \quad (1)$$

The control signal of lane l at time t is denoted by $s_l(t) \in \{0, 1\}$. $s_l(t) = 1$ indicates that lane l at time t receives a green light, while $s_l(t) = 0$ implies that lane l at time t gets a red light. The variable $c_l(t)$ represents the discharge rate of lane l at time t , which determines the rate at which vehicles can depart the lane when the signal is green. The variable $r_{(k,l)}(t)$ denotes the turning ratio from lane k to lane l at time t . This parameter captures the proportion of vehicles from lane k that are turning into lane l . The variable $d_l(t)$ corresponds to the exogenous flow of vehicles entering lane l at time t . It is stated that the variables $c_l(t)$, $r_{(k,l)}(t)$, and $d_l(t)$ are non-negative bounded i.i.d. (independent and identically distributed) random variables over time, suggesting that these parameters are stochastic in nature and follow a bounded probability distribution that may change over time. For each lane and each discrete time slot, the $x_l(t+1)$ captured by Equation (1) represents the combined effect of leaving vehicles and arriving vehicles. The number of leaving vehicles should not exceed the number of vehicles currently on lane l . Regarding the third term in Equation (1), all arriving vehicles are constrained by the physical space available on lane l . The arriving vehicles originate from a set of input lanes, denoted as $In(l)$, and the exogenous flow $d_l(t)$ entering lane l .

Equation (1) provides a comprehensive representation of vehicle evolution, capturing the key elements of the system including signal states, departure rates, turning ratios, and exogenous traffic inputs. This information is crucial for developing and analyzing RL-based signal control strategies.

The evolution Equation (1) presented in this work exhibits similarities to the traffic models reported in prior studies [10,29]. However, the key distinction explored in the current paper is the adoption of a lane-based evolution approach in contrast to a movement-based formulation. Employing a lane-based perspective offers practical advantages. A single lane can accommodate both straight-through and turning movements, and traffic lights are typically implemented at the lane group level, regulating whether a set of lanes should be granted the right-of-way or be required to stop.

In the context of traffic control, the turning ratios and exogenous flow rates are typically considered as known or given inputs to the system. For a signalized intersection, the discharge rate of vehicles across the stop-line is not constant during the green phase but rather exhibits temporal variations.

Consequently, the state representation employed in the RL-based control framework proposed in this study encompasses the number of vehicles present in each lane, the current signal phase, and the elapsed time within the ongoing phase. This comprehensive state definition aims to capture the dynamic nature of traffic conditions at the intersection, which is crucial for the effective optimization of signal control policies.

3.1.2. Action Design

In certain studies on RL-based TSC, such as [6,7], the selection of signal phases by the agent is stochastic and does not adhere to a predetermined cycle structure. This lack of structure may lead to two significant implementation challenges in practice [7].

First, drivers accustomed to observing established traffic signal cycles may experience confusion when they notice that their phase is skipped due to the agent's alternating selection of other phases. Second, individual turning movements may face excessively long wait times before receiving a green light, resulting in prolonged delays for specific vehicles. Both scenarios could lead to increased driver complaints and, in extreme cases, may cause drivers to perceive the signal controller as malfunctioning, potentially resulting in the dangerous behavior of running a red light.

Common strategies for TSP operating in mixed flow conditions include green extension and red truncation [4]. In this approach, the green-light duration is extended to provide additional time for buses to pass through the intersection, while the red-light duration is truncated to minimize delays for these vehicles.

As previously mentioned, this study employs a framework in which each agent has two permissible actions: either to maintain the current signal indication or to transition to the next phase in a cyclic manner. This structured approach aims to mitigate the potential issues associated with the stochastic phase selection, thereby enhancing the overall effectiveness of the traffic signal control system.

3.1.3. Reward Design

One significant challenge associated with current RL-based TSC approaches is that their configurations are often heuristic and lack robust theoretical justification derived from the transportation literature [10]. A well-structured reward function enhances the learning efficiency of the agent; however, there are no absolute guidelines for constructing such a function. The primary objective of TSC is to minimize travel time, a goal that is inherently complex and difficult to achieve directly [30]. Consequently, a crucial question when applying RL to TSC is how to effectively define the reward function.

This study proposes a reward function grounded in the principles of MP control. The fundamental concept of MP is to minimize the "pressure" at an intersection, which can be broadly defined as the difference between the number of vehicles on incoming lanes and the number on outgoing lanes [10]. Incoming lanes refer to those from which traffic enters the intersection, while outgoing lanes are those from which traffic exits. Over the past decade, MP control has evolved from a novel mathematical concept within a simple store-and-forward queuing model to encompass practical applications [7]. The appealing characteristics of MP control include (i) mathematical proofs of maximum throughput, (ii) real-time phase selection based on actual traffic conditions, and (iii) a decentralized control policy.

The following sections will outline the theoretical framework for designing the reward function, emphasizing its alignment with the objectives of MP control.

- Land-based MP control derivation

MP control is a method for directing traffic around a queuing network that achieves maximum network throughput, which is established using concepts of a backpressure routing algorithm. MP control in the previous studies, which is based on traffic movements, has two problems. The first problem is that the movement-based MP control aggregates the characteristics of lane groups, ignoring individual differences. The second problem is that the movement-based MP control assumes that different movements are separate and do not block each other. To address these issues, this study proposes a lane-based MP control.

Since the first term and second term of Equation (1) can be rewritten as follows:

$$x_l(t) - \min[x_l(t), s_l(t)c_l(t)] = \max[x_l(t) - s_l(t)c_l(t), 0] \quad (2)$$

$$\min \left[\sum_{k \in \ln(l)} \min \left[s_k(t)c_k(t)r_{(k,l)}(t), x_k(t)r_{(k,l)}(t) \right] + d_l(t), x_{max,l} - x_l(t) \right] \leq \sum_{k \in \ln(l)} s_k(t)c_k(t)r_{(k,l)}(t) + d_l(t) \quad (3)$$

Using Equations (2) and (3), Equation (1) can be rewritten:

$$x_l(t + 1) \leq \max[x_l(t) - s_l(t)c_l(t), 0] + \sum_{k \in \ln(l)} s_k(t)c_k(t)r_{(k,l)}(t) + d_l(t) \quad (4)$$

For each slotted time t defines the non-negative Lyapunov function:

$$L(t) \triangleq \frac{1}{2} \sum_l (q_l x_l(t))^2 \quad (5)$$

The Lyapunov function is a scalar measure of the total number of vehicles in the network. The positive value q_l^2 in Equation (5) represents the coefficient of vehicles on lane l . $\mathbf{x}(t)$ and \mathbf{q} represent the vector of number of vehicles ($x_l(t)$) and positive value (q_l^2), respectively. Plugging Equation (4) into (5) yields the following:

$$\begin{aligned} L(t + 1) &= \frac{1}{2} \sum_l (q_l x_l(t + 1))^2 \\ &\leq \frac{1}{2} \sum_l q_l^2 \left(\max[x_l(t) - s_l(t)c_l(t), 0] + \sum_{k \in \ln(l)} s_k(t)c_k(t)r_{(k,l)}(t) + d_l(t) \right)^2 \end{aligned} \quad (6)$$

The conditional Lyapunov drift is defined as the change in a Lyapunov function from one time step to the next:

$$\Delta(t) \triangleq \mathbb{E}[L(t + 1) - L(t) | \mathbf{q}\mathbf{x}(t)] \quad (7)$$

It is defined as the change in a Lyapunov function from one time step to the next. In the derivation of the MP control, the key is to minimize this Lyapunov drift $\mathbb{E}[\Delta(t)]$. The goal is to take actions that make the Lyapunov drift in the negative direction toward zero.

Set $\alpha = x_l(t)$, $\beta = s_l(t)c_l(t)$, and $\gamma = \sum_{k \in \ln(l)} s_k(t)c_k(t)r_{(k,l)}(t) + d_l(t)$. It follows that

$$\begin{aligned} (\max(\alpha - \beta, 0) + \gamma)^2 &= (\max(\alpha - \beta, 0))^2 + \gamma^2 + 2\max(\alpha - \beta, 0)\gamma \\ &\leq (\alpha - \beta)^2 + \gamma^2 + 2\alpha\gamma \\ &= \alpha^2 + \beta^2 + \gamma^2 + 2\alpha(\gamma - \beta) \end{aligned} \quad (8)$$

To simplify the formula in order to make it easier to understand, write $\Delta = \Delta(t)$, $x_l = x_l(t)$, $c_l = c_l(t)$, $s_l = s_l(t)$, $r_{(k,l)} = r_{(k,l)}(t)$, $d_l = d_l(t)$, $\mathbf{x} = \mathbf{x}(t)$. Next plugging Equations (5) and (6) into Equation (7) and using Equation (8) yields the following:

$$\begin{aligned} \Delta &\leq \frac{1}{2} \mathbb{E} \left[\sum_l q_l^2 \left(x_l^2 + (s_l c_l)^2 + \left(\sum_{k \in \ln(l)} s_k c_k r_{(k,l)} + d_l \right)^2 + 2x_l \left(\sum_{k \in \ln(l)} s_k c_k r_{(k,l)} + d_l - s_l c_l \right) \right) - \sum_l q_l^2 x_l^2 \middle| \mathbf{q}\mathbf{x} \right] \\ &= \frac{1}{2} \mathbb{E} \left[\sum_l q_l^2 \left((s_l c_l)^2 + \left(\sum_{k \in \ln(l)} s_k c_k r_{(k,l)} + d_l \right)^2 \right) + 2 \sum_l q_l^2 x_l \left(\sum_{k \in \ln(l)} s_k c_k r_{(k,l)} + d_l - s_l c_l \right) \middle| \mathbf{q}\mathbf{x} \right] \\ &= \mathbb{E} \left[\frac{1}{2} \sum_l q_l^2 \left((s_l c_l)^2 + \left(\sum_{k \in \ln(l)} s_k c_k r_{(k,l)} + d_l \right)^2 \right) + \sum_l q_l^2 x_l \left(\sum_{k \in \ln(l)} s_k c_k r_{(k,l)} + d_l - s_l c_l \right) \middle| \mathbf{q}\mathbf{x} \right] \end{aligned} \quad (9)$$

Suppose the second moments of leaving vehicles and the second moments of arriving vehicles in each queue are bounded so that there is a finite constant $B > 0$ and $B < \infty$ such that for all time t the following property holds:

$$\mathbb{E} \left[(s_l c_l)^2 + \left(\sum_{k \in In(l)} s_k c_k r_{(k,l)} + d_l \right)^2 \middle| \mathbf{q}\mathbf{x} \right] \leq B \tag{10}$$

Plugging Equation (10) into Equation (9) yields the following:

$$\Delta \leq \frac{B}{2} \sum_l q_l^2 + \mathbb{E} \left[\sum_l q_l^2 x_l \left(\sum_{k \in In(l)} s_k c_k r_{(k,l)} + d_l - s_l c_l \right) \middle| \mathbf{q}\mathbf{x} \right] \tag{11}$$

Switching the sums of the second term in Equation (11) obtains the following:

$$\Delta \leq \frac{B}{2} \sum_l q_l^2 + \sum_l q_l^2 x_l d_l + \mathbb{E} \left[\sum_l \left[s_l c_l \left(\sum_{m \in Out(l)} q_m^2 x_m r_{(l,m)} - q_l^2 x_l \right) \right] \middle| \mathbf{q}\mathbf{x} \right] \tag{12}$$

where $Out(l)$ is the set of lanes output from lane l .

Define the weight $w_l(t)$ of each lane l as follows:

$$w_l(t) = q_l^2 x_l - \sum_{m \in Out(l)} q_m^2 x_m r_{(l,m)} \tag{13}$$

Let $\mathbf{s} = (s_l)$ be the network signal control vector and $\Gamma = \{\mathbf{s}\}$ be a set of all network signal control vectors. Define the pressure of \mathbf{s} as follows:

$$p(\mathbf{s} \in \Gamma) = \sum_l s_l c_l w_l \tag{14}$$

MP control policy is to minimize the right-hand side of Equation (12). This amounts to selecting the MP control u^* at time t as follows:

$$u^* = \operatorname{argmax}_{\mathbf{s} \in \Gamma} \{p(\mathbf{s})\} \tag{15}$$

- Determination of positive coefficient values

The following work needs to determine the positive coefficient values q_l^2 and q_m^2 in $w_l(t)$.

When addressing traffic control at intersections, it is essential to consider not only the maximization of throughput but also the issues of spillover and gridlock. Spillover occurs when vehicles making turning movements occupy the available storage length of turning bays, thereby obstructing through movements. Gridlock occurs when vehicles enter an intersection on a green light but do not have enough room on the other side of the intersection to make it all the way through. To mitigate these challenges, this study proposes a reward function for intersections that incorporates key factors such as phase, lane length, and the equilibrium between incoming and outgoing lanes.

To prevent spillover, it is crucial to take lane length into account, as longer lanes provide additional capacity for vehicles waiting to proceed. Specifically, the model considers the maximum permissible number of vehicles on a lane, employing a normalized count rather than an absolute figure in the reward function. In addition, by observing Equation (13), the agent is more likely to activate a phase that allows for more lane goings because this leads to a higher reward, but this can result in short green times for certain phases.

To avoid gridlock, when $w_l(t) < 0$ (indicating negative equilibrium), the agent should be discouraged from activating the corresponding phase to prevent overflow from downstream traffic.

As previously discussed, this study defined the weight of phase ϕ :

$$w_{\phi}(t) = \max \left(\frac{\sum_{l \in \text{In}(\phi)} x_l(t)}{\sum_{l \in \text{In}(\phi)} x_{\max,l}} - \frac{\sum_{m \in \text{Out}(\phi)} x_m(t)}{\sum_{m \in \text{Out}(\phi)} x_{\max,m}}, 0 \right) \quad (16)$$

where $\text{In}(\phi)$ and $\text{Out}(\phi)$ indicate a set of incoming lanes and outgoing lanes of phase ϕ , respectively. The weight of phase $w_{\phi}(t)$ represents the difference in the weighted number of vehicles between the incoming lanes and the outgoing lanes of the phase ϕ at time t .

Equation (16) can be expressed as follows:

$$w_{\phi}(t) = \max \left(\sum_{l \in \text{In}(\phi)} q_l^2 x_l(t) - \sum_{m \in \text{Out}(\phi)} q_m^2 x_m(t), 0 \right) \quad (17)$$

where $q_l^2 = \frac{1}{\sum_{l \in \text{In}(\phi)} x_{\max,l}}$ and $q_m^2 = \frac{1}{\sum_{m \in \text{Out}(\phi)} x_{\max,m}}$

- Reward function

Finally, the reward function in this study is defined as follows:

$$R_i(t) = - \sum_{\phi \in \Phi_i} w_{\phi}(t) \quad (18)$$

where Φ_i is a set of phases of intersection i .

3.2. Justification for Reward Function

There are two primary reasons supporting the efficacy of the proposed reward function.

First, RL serves as a computational framework for understanding and automating goal-directed learning and decision-making. This notable characteristic facilitates the agent's ability to learn optimal behaviors within an environment, prioritizing long-term maximum rewards rather than acting greedily at each time step. The agent's objective, denoted as $G(t)$ in Equation (19), where γ is the discount rate between 0 and 1, is to maximize the total cumulative reward it receives over time. This approach emphasizes the importance of maximizing not only immediate rewards but also cumulative rewards in the long run. It effectively addresses the limitations associated with MP control.

$$G(t+1) = \sum_{k=0}^{\infty} \gamma^k R(t+k+1) \quad (19)$$

Second, selecting the appropriate action requires consideration of the discharge rate under MP control. Traffic engineers have indicated that the discharge rate of vehicles crossing the stop line at signalized intersections is not constant and fluctuates throughout the green phase. In the context of RL, the agent is not explicitly informed about which actions will yield the highest rewards. Instead, it must identify optimal actions through experimentation. This intrinsic property of RL negates the need to inform the agent about the specific discharge flow rates in a variable environment prior to learning how to effectively control traffic signals. Such a characteristic enhances the potential for improved outcomes.

3.3. Learning Process

The learning process involves updating the policy, which serves as a mapping from the perceived states of the environment to the corresponding actions to be executed in those states. This update is informed by feedback received in the form of rewards from the environment. Q-learning is the most widely utilized method for RL-based TSC [13]. In recent years, some studies have employed Deep Q-Networks (DQNs) as function approximators, utilizing a class of artificial neural networks (ANNs) to estimate the Q-value function.

Both Q-learning and DQNs are RL algorithms aimed at identifying the optimal policy for an agent operating within a specific environment. Q-learning is a simpler, traditional algorithm that performs effectively in smaller, less complex environments. In contrast, the DQN is a more sophisticated algorithm that excels in larger and more complex environments. This study adopts a DQN as it overcomes several limitations inherent in Q-learning, enabling the handling of high-dimensional state spaces, reducing the variance of Q-value estimates, and enhancing the overall stability of the learning process. Figure 1 illustrates the general process of a DQN.

Each iteration of the learning process lasts for 4000 s. Each iteration always starts with an empty network and, therefore, must be pre-loaded for training. This initial warm-up period allows the system to stabilize and ensures that the simulation results are not influenced by transient conditions. The initial 400 s serve as a warm-up period for the road network, while the subsequent 3600 s (equivalent to one hour) are allocated for training. To facilitate the agents' adaptation to varying environments, the volume and turning ratios at each intersection approach are randomly reset at the beginning of each iteration.

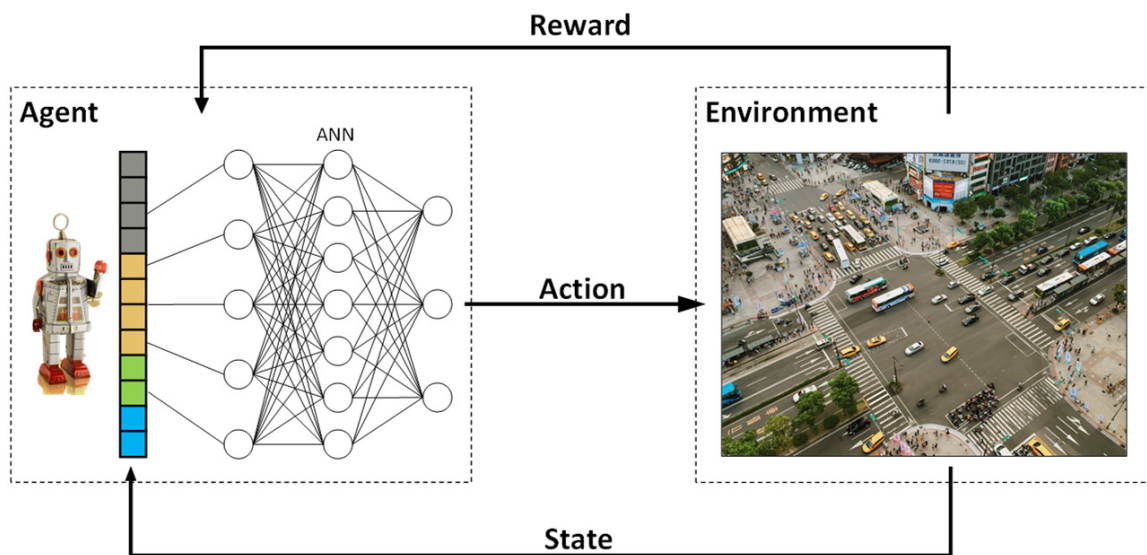


Figure 1. General process of a DQN.

3.4. TSP Control with Priority Factor (PF)

In scenarios involving multiple vehicles in an incoming lane, the weight of phase $w_{\varphi}(t)$ increases. To mitigate this effect, the agent prioritizes phases with a higher volume of vehicles. Consequently, this study introduces a factor termed the priority factor (PF), which enables agents at intersections to give precedence to buses over other vehicles. The agent should prioritize the lane with buses by increasing the count of vehicles in that lane. The number of vehicles on lane l is represented as follows:

$$x_l = x_{car,l} + PF \times x_{bus,l}, \text{ where } PF \geq 1 \quad (20)$$

In the context of TSP, two primary challenges arise: conflicting priority requests and negative externalities. Conflicting priority requests may occur when multiple buses simultaneously seek priority. Negative externalities refer to the adverse impacts that TSP can impose on non-transit users of the roadway. The proposed TSP control framework adopts an RL approach with the inclusion of PF, offering a potential solution to address both conflicting priority requests and the negative externalities associated with TSP.

The coordination of conflicting priority requests can be managed by assigning varying PF values to buses. For instance, arterial buses may receive a higher PF to facilitate efficient passage through intersections compared to side-street buses. Furthermore, PF can be dynamically adjusted based on the status of the bus, such as its passenger load. Under this relative PF framework, buses with higher passenger loads are accorded greater priority. Consequently, while the optimal control objective for all users remains a priority for the agent, the performance of other vehicles is also taken into account. This approach ensures that prioritizing buses does not significantly disrupt overall traffic flow.

For each intersection and in each time slot, the procedure in Figure 2 implements PF to do TSP control in real time.

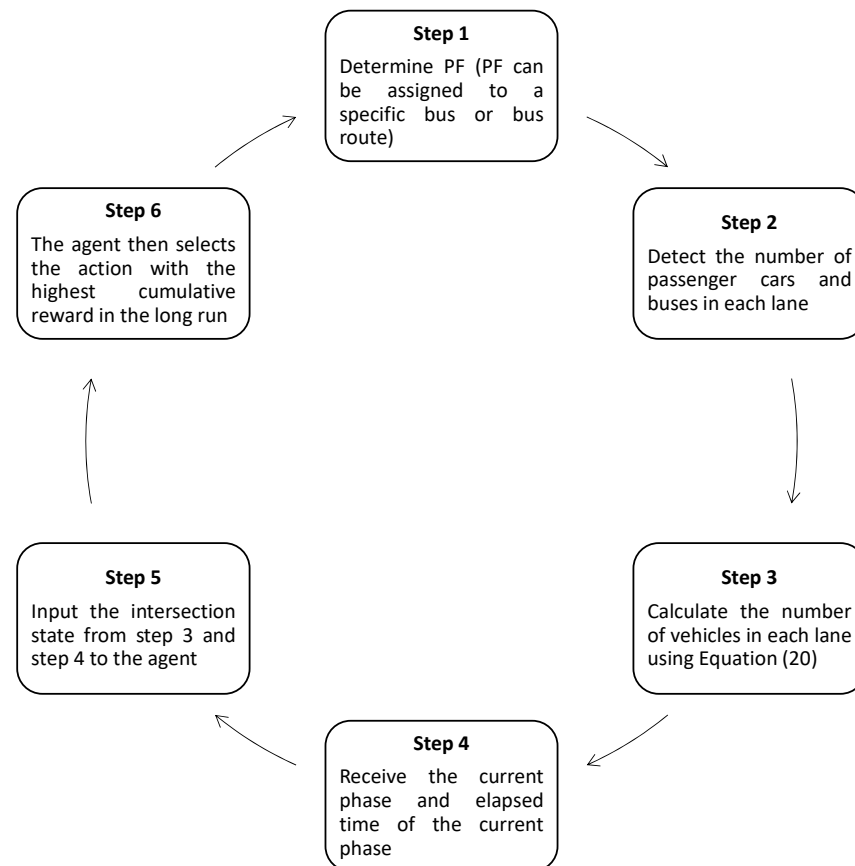


Figure 2. Procedure to TSP control with PF.

4. Experimental Setting

4.1. Simulator

The experiments were conducted using VISSIM [31], a widely recognized microscopic traffic simulation software that facilitates external control of traffic signals through its COM interface. This interface allows users to dynamically access and manipulate VISSIM objects during the simulation process, thereby enabling the integration of custom algorithms.

Figure 3 illustrates the simulation procedure when employing the COM interface in the control study. Upon activation, VISSIM transmits the network state to the COM interface at each time interval. The agents process these states, select appropriate control signals, and communicate their decisions back to the COM interface. Subsequently, the COM interface relays the updated signal status to VISSIM for the subsequent simulation period. This seamless interaction underscores the flexibility and adaptability of VISSIM in accommodating advanced traffic-control strategies.

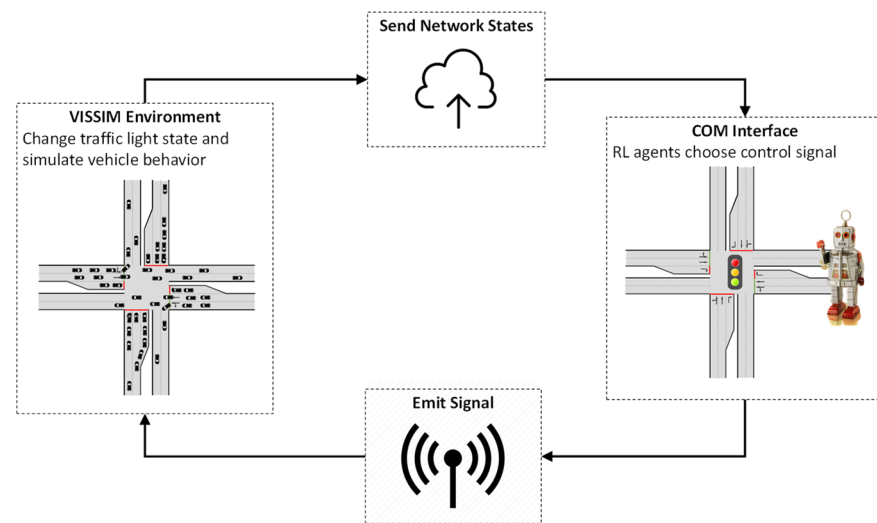


Figure 3. Simulation procedure.

4.2. Road Networks

Two common types of road networks are arterial networks and grid networks, each with unique characteristics and functions. Arterial networks are designed to facilitate the efficient movement of large volumes of traffic across longer distances. They consist of major roads that connect different areas. Grid networks consist of a series of intersecting streets that form a grid-like pattern. This type of network is common in urban planning and is designed to provide multiple routes for travel. Urban planners often consider a combination of both types to create a comprehensive transportation system that meets the needs of the community.

Based on the above, this study evaluated two distinct types of road networks: Arterial_{1×3}, which is a three-intersection arterial network, and Grid_{2×2}, which consists of a grid layout featuring four intersections. Figure 4 presents the configurations of both Arterial_{1×3} and Grid_{2×2}. Each intersection within these networks includes four approaches, with each approach comprising one lane designated for straight-through or left-turn movements, one lane for straight-through traffic, and one exclusive lane for right turns.

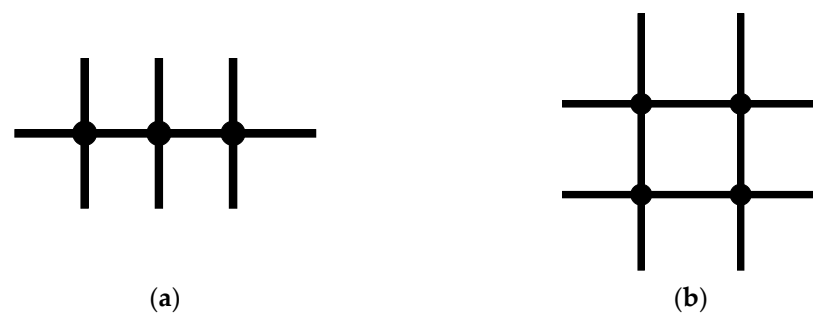


Figure 4. This study evaluated two distinct types of road networks: (a) Arterial_{1×3} network and (b) Grid_{2×2} network.

4.3. Traffic Light Signal Plan

The traffic signal plan consists of four phases organized in a cyclic manner. The sequence of these phases is depicted in Figure 5. This configuration represents a standard traffic signal plan for intersections with four approaches.

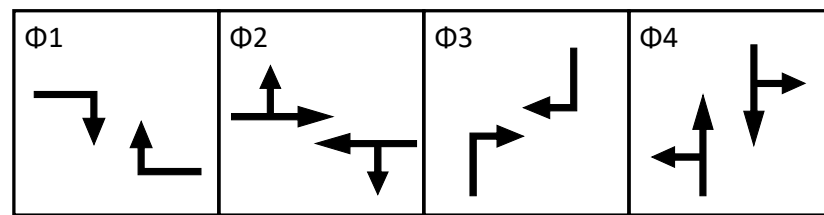


Figure 5. Traffic light signal plan.

4.4. Traffic Volume and Turning Ratio

In real-world scenarios, traffic environments are characterized by constant variability. To effectively simulate these dynamic conditions, both traffic volume and turning ratios are treated as non-static parameters. There are two classifications of traffic volume inputs for the networks: light-to-medium volume (LM), defined as ranging from 500 to 750 passenger cars per hour, and medium-to-heavy volume (MH), which spans from 750 to 1000 passenger cars per hour.

Regarding turning ratios, each approach accommodates three possible driving directions: straight, right, and left. The ratio for straight movements is typically between 0.5 and 0.7.

4.5. Bus Routes

Different streets, neighborhoods, and cities exhibit distinct transportation needs, necessitating a diverse array of service types to address them. Bus routes are typically designed in accordance with the characteristics of the road network, which may consist of arterial roads or grid configurations.

Bus routes that align with arterial roads are intended to facilitate efficient transportation along major thoroughfares. These routes often connect key destinations, including commercial areas, residential neighborhoods, and transportation hubs. Conversely, bus routes that operate within a grid network are structured to provide comprehensive coverage across a city or neighborhood, thereby ensuring easy access to various locations within the grid and enhancing convenience for residents and commuters.

In this study, four specific bus route types—B310, B320, B330, and B340—are designated within the Arterial_{1×3} and Grid_{2×2} configurations. These designations serve to identify and differentiate the various bus routes operating within these networks. Figure 6 illustrates the routing of these bus routes in both Arterial_{1×3} and Grid_{2×2}.

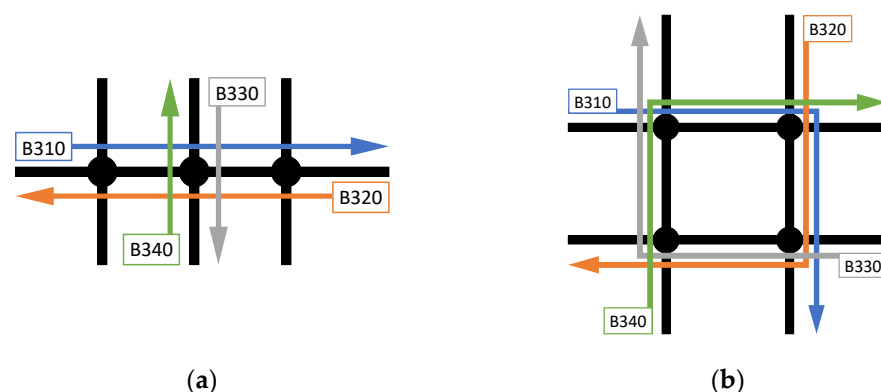


Figure 6. Bus routes: (a) Arterial_{1×3} bus routes.; (b) Grid_{2×2} bus routes.

5. Results

5.1. No Priority Control

Before assessing the efficacy of the proposed TSP model with the PF, it is essential to determine whether the control method can effectively manage typical traffic conditions without the PF. Table 2 reports experimental results of average travel times (ATTs) of buses

compared with MP and fixed time control on Arterial_{1×3} and Grid_{2×2} under no priority control. A shorter ATT signifies a reduction in the overall delay experienced by travelers at the intersection. The ATT is calculated as follows:

$$ATT = \frac{1}{n} \sum_{i=1}^n (T_i^{exit} - T_i^{enter}) \quad (21)$$

where T_i^{enter} and T_i^{exit} represent the times at which a vehicle i enters and exits the network, respectively.

The findings are as follows:

1. Light-to-Medium (LM) Volume

- In the Arterial_{1×3} network, the proposed model consistently exhibits the shortest ATT. Specifically, for routes B310 and B320, the ATT ranges from 133 to 134 s, while for routes B330 and B340, the ATT ranges from 45 to 46 s.
- In the Grid_{2×2} network, the proposed model outperforms both MP and fixed-time control strategies, with the ATT ranging from 154 to 162 s.
- Both in Arterial_{1×3} and Grid_{2×2} networks, MP control shows moderate performance, and fixed-time control results in the highest travel times highlighting its inadequacy in dynamic traffic situations.

2. Medium-to-Heavy (MH) Volume

- In the Arterial_{1×3} network, the proposed model with an ATT between 160 and 52 s maintains its superior performance, which is significantly lower than the other methods.
- In the Grid_{2×2} network, the proposed model continues to lead in performance with an ATT ranging from 195 to 220 s.
- MP control shows a decline in efficiency under heavier traffic conditions. Fixed-time control exhibits the highest travel times which further underscores its limitations in handling medium-to-heavy traffic volumes.

3. Summary

- The proposed model demonstrates superior performance relative to both fixed-time control and MP control. As traffic volume increases, the advantages of the proposed model become more pronounced.
- MP control, which is a greedy algorithm, achieves optimal solutions at each time step. However, it frequently switches signal phases, leading to increased lost time.
- Fixed-time control, which is predicated on overly simplistic assumptions or prior knowledge of traffic patterns, is prone to failure in dynamic traffic scenarios.

Table 2. ATTs (seconds) without priority control.

Bus Route	Arterial _{1×3}				Grid _{2×2}			
	B310	B320	B330	B340	B310	B320	B330	B340
LM volume								
Proposed	134	133	46	45	158	154	163	162
MP	175	176	46	46	175	176	174	171
Fixed time	184	181	50	53	167	165	209	184
MH volume								
Proposed	160	156	53	52	199	195	212	220
MP	216	219	56	56	259	260	239	227
Fixed time	318	367	132	145	347	358	410	451

5.2. Priority Control

5.2.1. Cases

This paper presents three distinct scenarios for each road network: the Base case, Case 1, and Case 2. The Base case operates without priority control, which corresponds to a PF value of 1. Case 1 involves specific bus routes that are granted priority control, while Case 2 applies priority control to all bus routes within the network. The PF settings for each scenario are summarized in Table 3, where “NP” denotes the absence of priority control and “PF” indicates the application of the priority factor.

Table 3. Test cases for this study.

Bus Route	Arterial _{1×3}				Grid _{2×2}			
	B310	B320	B330	B340	B310	B320	B330	B340
Base case	NP	NP	NP	NP	NP	NP	NP	NP
Case 1	PF	PF	NP	NP	PF	NP	NP	NP
Case 2	PF	PF	PF	PF	PF	PF	PF	PF

In the scenarios featuring PF, there are three defined priority levels: low, medium, and high. As the priority level increases, so does the corresponding PF value. The setting of the PF is determined through a trial-and-error approach. In the following experiments, the PF is assigned values of 2, 5, and 8, which effectively categorize the priority control levels into low, medium, and high. It is crucial to emphasize that an excessively high PF may result in unpredictable control behavior from the RL agent because the PF is determined following the training of the RL agent.

The Base case serves as a benchmark for evaluating the differences in performance before and after the implementation of priority control. Case 1 assumes the presence of a TSP arterial on the Arterial_{1×3} network or a designated TSP bus route on the Grid_{2×2} network. Case 2 is designed to simulate situations of conflicting priority requests. In a long arterial or large metropolitan region, multiple TSP requests often occur, especially for those intersections with multiple bus routes passing through.

5.2.2. ATT Analysis

This section presents the analysis of the ATT of buses and cars.

1. Bus ATTs

The reduction in the ATT for buses is a key indicator for evaluating the performance of TSP control systems [15]. Tables 4 and 5 present a comparative analysis of the ATT in Case 1 and Case 2 against the Base case for both the Arterial_{1×3} and Grid_{2×2} networks across various scenarios. The results are summarized as follows:

- Bus routes with a PF greater than 1 exhibit a lower ATT compared to the Base case with a PF of 1 under identical traffic conditions. The reduction in the ATT is notable, with the highest decrease observed at 17% in the Arterial_{1×3} network and 25% in the Grid_{2×2} network.
- In Case 1 of both the Arterial_{1×3} and Grid_{2×2} networks, agents are inclined to grant right-of-way to priority bus routes at intersections. Consequently, non-priority bus routes experience longer crossing times at these intersections. This represents a trade-off for allowing priority buses to traverse intersections more rapidly.
- In Case 2 for both Arterial_{1×3} and Grid_{2×2}, all bus routes demonstrate reduced ATTs compared to the Base case due to the application of PF across the board. However, the ATT for routes B310 and B320 in Arterial_{1×3}, as well as B310 in Grid_{2×2} in Case 2, is longer than in Case 1. This outcome indicates that agents do not prioritize a specific direction for bus routes at intersections uniformly; rather, all bus routes compete for the right-of-way. In the Arterial_{1×3} network, arterial buses compete with side-street buses for priority, while in the Grid_{2×2} network, bus routes from different directions

vie for precedence. These findings illustrate that the proposed TSP control with PF effectively addresses conflicting priority requests.

2. Passenger car ATTs

This section examines the travel patterns of passenger cars traveling from north to south on side streets within the Arterial_{1×3} network, as well as all passenger cars within the Grid_{2×2} network.

In the Arterial_{1×3} network, passenger cars traveling along the arterial (east-west direction) reduce ATTs due to the benefits of priority control. Consequently, this focus is especially pertinent considering the influence that priority buses have on passenger cars originating from side streets that intersect with the arterial roads. In contrast, the indistinguishable nature of main and minor roads in the Grid_{2×2} network necessitates equal consideration for priority buses and traffic flows in all directions.

Table 6 presents the results and findings, which are summarized as follows:

- Agents prioritize buses over passenger cars to enhance the efficiency of bus travel times. Consequently, the ATT for passenger cars can increase by as much as 12% in Arterial_{1×3} and 7% in Grid_{2×2}, while the ATT for buses may be reduced by up to 17% and 25%, respectively. This increase in passenger car travel time is considered acceptable due to its relatively minor impact.
- Generally, it is observed that the ATT for passenger cars tends to lengthen as the PF increases.
- Across both Arterial_{1×3} and Grid_{2×2} networks, passenger cars in Case 2 experience shorter ATTs compared to Case 1. This improvement can be attributed to the facilitation provided by all buses with a PF greater than 1, which enables passenger cars traveling in multiple directions to cross intersections more efficiently.

Table 4. Comparison of ATTs with Base case of Arterial_{1×3}.

Priority Level	Base Case		Case 1			Case 2		
	NP	Low PF = 2	Medium PF = 5	High PF = 8	Low PF = 2	Medium PF = 5	High PF = 8	
LM volume								
B310	134	127 (−5%)	118 (−12%)	110 (−17%)	131 (−2%)	120 (−10%)	115 (−14%)	
B320	133	127 (−5%)	118 (−12%)	111 (−17%)	127 (−4%)	120 (−9%)	115 (−13%)	
B330	46	44 (−3%)	49 (+7%)	50 (+9%)	45 (−2%)	44 (−3%)	41 (−10%)	
B340	45	46 (+3%)	47 (+5%)	50 (+13%)	44 (−2%)	42 (−6%)	41 (−7%)	
MH volume								
B310	160	153 (−4%)	141 (−12%)	135 (−16%)	145 (−10%)	146 (−9%)	133 (−17%)	
B320	156	151 (−3%)	135 (−14%)	134 (−14%)	149 (−4%)	137 (−12%)	140 (−11%)	
B330	53	52 (−1%)	54 (+3%)	56 (+6%)	49 (−7%)	49 (−7%)	45 (−15%)	
B340	52	54 (+5%)	58 (+12%)	57 (+11%)	49 (−6%)	49 (−5%)	48 (−8%)	

Table 5. Comparison of ATTs with Base case of Grid_{2×2}.

Priority Level	Base Case		Case 1		Case 2		
	Non PF = 1	Low PF = 2	Medium PF = 5	High PF = 8	Low PF = 2	Medium PF = 5	High PF = 8
LM volume							
B310	158	148 (−6%)	129 (−18%)	118 (−25%)	153 (−3%)	133 (−16%)	134 (−15%)
B320	154	155 (0%)	160 (+3%)	158 (+2%)	151 (−3%)	135 (−13%)	133 (−14%)
B330	163	163 (0%)	176 (+8%)	172 (+5%)	162 (−1%)	153 (−6%)	147 (−10%)
B340	162	159 (−2%)	172 (+6%)	173 (+7%)	161 (−1%)	145 (−11%)	148 (−9%)
MH volume							
B310	199	196 (−2%)	178 (−10%)	167 (−16%)	185 (−7%)	177 (−11%)	164 (−17%)
B320	195	215 (+10%)	202 (+3%)	210 (+8%)	192 (−2%)	167 (−15%)	164 (−16%)
B330	212	230 (+8%)	235 (+11%)	229 (+8%)	207 (−2%)	181 (−15%)	188 (−11%)
B340	220	254 (+16%)	255 (+16%)	266 (+21%)	205 (−7%)	205 (−7%)	185 (−16%)

Table 6. Comparison of ATTs of passenger cars with Base case.

Priority Level	Base Case		Case 1		Case 2		
	Non PF = 1	Low PF = 2	Medium PF = 5	High PF = 8	Low PF = 2	Medium PF = 5	High PF = 8
Side-street passenger cars in Arterial _{1×3}							
LM	45	45 (+1%)	47 (+6%)	50 (+12%)	45 (0%)	46 (+4%)	48 (+7%)
MH	51	52 (+2%)	54 (+5%)	57 (+11%)	50 (−3%)	52 (+2%)	54 (+6%)
All passenger cars in Grid _{2×2}							
LM	123	123 (0%)	126 (+3%)	126 (+3%)	124 (+1%)	128 (+4%)	129 (+5%)
MH	156	167 (+7%)	162 (+4%)	167 (+7%)	159 (+2%)	157 (0%)	158 (+1%)

5.2.3. Signal Timing Analysis

The traffic light at each intersection is managed by a single RL agent, referred to as a “controller” followed by a Latin letter. Arterial_{1×3} has three controllers, while Grid_{2×2} has four controllers. Figure 7 illustrates the positioning of each controller within the Arterial_{1×3} and Grid_{2×2} networks.

1. Average cycle length

Cycle length defines the time required for a complete sequence of indications. Tables 7 and 8 present the experimental results regarding the average cycle length for the Arterial_{1×3} and Grid_{1×3} networks, respectively.

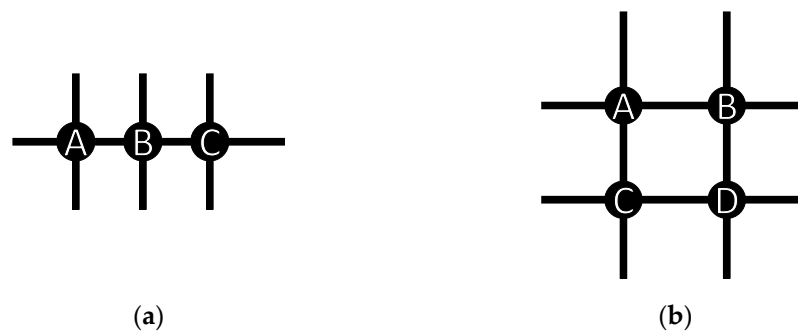


Figure 7. Positions of each agent: (a) Arterial_{1×3}; (b) Grid_{2×2}.

Table 7. Cycle length (seconds) in Arterial_{1×3}.

	Base Case		Case 1		Case 2		
	PF = 1	PF = 2	PF = 5	PF = 8	PF = 2	PF = 5	PF = 8
LM volume							
Controller A	49	47	51	54	48	50	55
Controller B	48	48	51	54	50	55	56
Controller C	48	47	50	54	49	51	54
MH volume							
Controller A	59	60	61	64	59	61	67
Controller B	60	61	64	66	60	65	71
Controller C	59	62	65	68	61	63	64

Table 8. Cycle length (seconds) in Grid_{2×2}.

	Base Case		Case 1		Case 2		
	PF = 1	PF = 2	PF = 5	PF = 8	PF = 2	PF = 5	PF = 8
LM volume							
Controller A	47	48	50	52	48	56	60
Controller B	47	48	50	52	50	55	57
Controller C	48	48	48	48	50	57	59
Controller D	47	47	49	49	48	55	60
MH volume							
Controller A	63	68	73	68	66	71	73
Controller B	63	66	68	69	61	67	65
Controller C	62	66	64	68	64	70	78
Controller D	59	69	66	70	60	68	74

The cycle length increases with higher traffic volumes, which is consistent with established principles of traffic signal design. Additionally, because the system allocates more time to prioritize buses approaching intersections with larger PF values, the cycle length tends to increase as the PF increases. This results in a reduced frequency of stops at those intersections. A comparison between Case 1 and Case 2 in both the Arterial_{1×3} and Grid_{1×3} networks shows that cycle lengths in Case 2 are generally longer than those in Case 1, indicating that cycle lengths are extended when a greater number of priority buses traverse the intersections.

2. Split

Within a traffic signal cycle, splits refer to the proportion of time allocated to each phase at an intersection. Tables 9 and 10 present the splits for controller B on the Arterial_{1×3} network and controller A on the Grid_{2×2} network, respectively. Controller B was selected

for analysis due to its central position on the Arterial_{1×3}. Meanwhile, all controllers in Grid_{2×2} are similar, making Controller A representative of the other controllers within the Grid_{2×2} network.

Table 9. Split of controller B in Arterial_{1×3}. Percentages marked in bold indicate the phase that serves the priority buses.

	Base Case		Case 1		Case 2		
	PF = 1	PF = 2	PF = 5	PF = 8	PF = 2	PF = 5	PF = 8
	LM volume						
Phase 1	22%	22%	21%	19%	21%	19%	19%
Phase 2	29%	29%	33%	34%	29%	32%	32%
Phase 3	21%	22%	20%	19%	21%	19%	19%
Phase 4	28%	27%	27%	27%	30%	30%	31%
	MH volume						
Phase 1	21%	20%	22%	19%	21%	20%	20%
Phase 2	30%	31%	31%	32%	31%	31%	31%
Phase 3	19%	19%	19%	21%	19%	18%	18%
Phase 4	30%	30%	28%	29%	29%	32%	31%

Table 10. Split of controller A in Grid_{2×2}. Percentages marked in bold indicate the phase that serves the priority buses.

	Base Case		Case 1		Case 2		
	PF = 1	PF = 2	PF = 5	PF = 8	PF = 2	PF = 5	PF = 8
	LM volume						
Phase 1	22%	21%	20%	20%	21%	19%	18%
Phase 2	28%	29%	30%	33%	28%	28%	30%
Phase 3	24%	24%	23%	22%	24%	27%	27%
Phase 4	27%	26%	26%	25%	27%	26%	25%
	MH volume						
Phase 1	20%	18%	19%	17%	19%	17%	17%
Phase 2	30%	29%	31%	32%	29%	29%	30%
Phase 3	24%	24%	22%	24%	23%	27%	25%
Phase 4	27%	29%	29%	27%	28%	28%	28%

The split allocated to movements on priority bus routes is greater than that for movements without priority bus routes, and this split increases as the PF rises. Moreover, there is no significant difference in the split between varying traffic volumes. Additionally, splits remain consistent for corresponding phases serving priority bus routes, even when different PF values are applied.

5.3. Dynamic PF

In the previous scenario analysis, the PF was maintained at a constant value, indicating that the priority control was fixed. However, it is essential to assign varying levels of priority control to individual buses based on specific circumstances, such as fluctuations in passenger numbers. The subsequent PF analysis employs the following formula:

$$PF = 1 + \frac{\text{number of passengers}}{\text{maximum number of passengers}} \times 7 \tag{22}$$

whereby the PF is determined according to varying passenger loads, resulting in different levels of priority control. As indicated by Equation (22), the PF is equal to 1 when there are no passengers on the bus, and it increases to 8 when the bus is at full capacity with passengers.

Assuming the maximum number of passengers is 80, Figure 8 displays the ATT for arterial bus routes (B310 and B320) in the Arterial $_{1\times 3}$ network, as well as for all bus routes in the Grid $_{2\times 2}$ network. The findings indicate that a dynamic PF can effectively reduce bus ATT as the number of passengers increases. In the Arterial $_{1\times 3}$ network, the ATT ranges from 113 s to 126 s under LM volume and from 137 s to 150 s under MH volume. In the Grid $_{2\times 2}$ network, the ATT ranges from 139 s to 161 s under LM volume and from 186 s to 214 s under MH volume. These results demonstrate that PF has broader applicability in dynamic situations such as aiming to enhance the operational regularity and punctuality of transit systems.

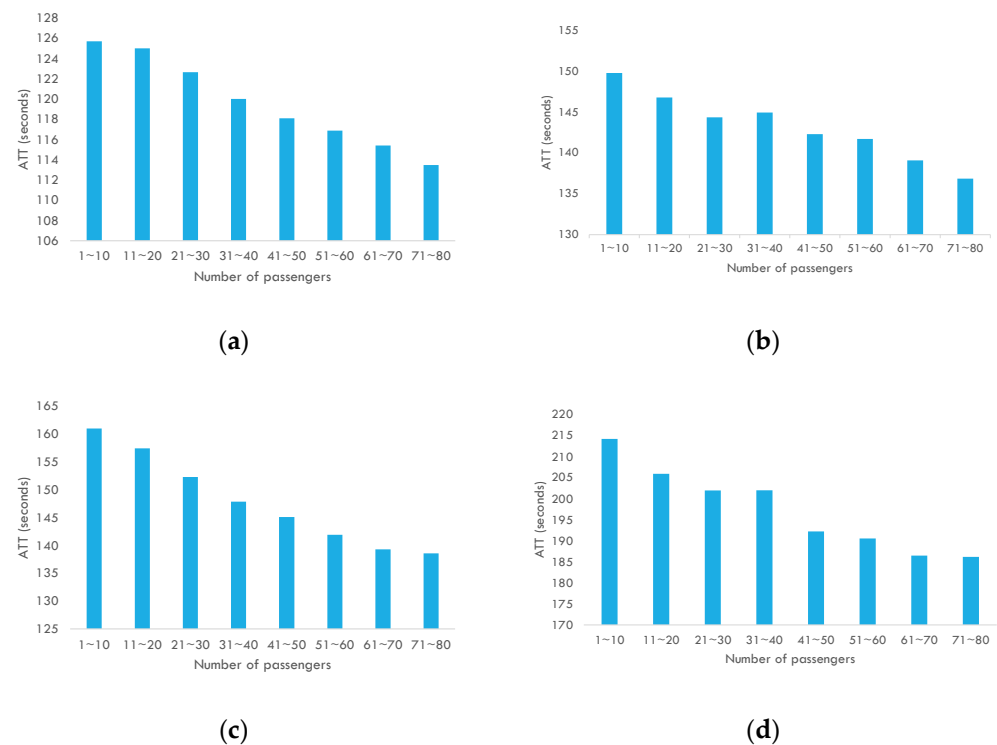


Figure 8. Bus ATT with dynamic PF: (a) B310 and B320 ATT on the Arterial $_{1\times 3}$ (Case 1)—LM volume; (b) B310 and B320 ATT on the Arterial $_{1\times 3}$ (Case 1)—MH volume; (c) Bus ATT on Grid $_{1\times 3}$ (Case 2)—LM volume; (d) Bus ATT on Grid $_{1\times 3}$ (Case 2)—MH volume.

6. Conclusions

This study introduces, for the first time, an innovative transit signal priority (TSP) control system that enhances reinforcement learning (RL)-based traffic signal control (TSC) by integrating a priority factor (PF) specifically designed for mixed-traffic road networks. This contribution not only improves the effectiveness of signal management for public transit but also promotes a more efficient flow of all traffic, representing a significant advancement in traffic signal control methodologies. Within the context of RL control construction, the paper thoroughly explains the rationale behind the design of states and actions, and it theoretically formulates the reward function using the principles of max-pressure (MP) control. To further prioritize buses over passenger cars, the study introduces the PF. The PF is designed to prioritize bus movements seeking to balance the competing objectives of enhancing bus operations while mitigating adverse impacts on non-transit users at signalized intersections.

To evaluate the effectiveness of the proposed RL-based TSP control method, experiments were conducted using VISSIM on two common types of road networks which are the arterial network identified as Arterial $_{1\times 3}$ and the grid network designated as Grid $_{2\times 2}$ within a dynamic traffic environment. The results indicated that the average travel time (ATT) for buses dropped 17% in Arterial $_{1\times 3}$ and 25% in Grid $_{2\times 2}$. This represents

a significant reduction. Furthermore, the proposed TSP with PF effectively addresses conflicting priority requests and mitigates negative externalities. When multiple buses with PF approach an intersection from different directions, the agent successfully resolves these conflicting requests.

An analysis of passenger car ATTs revealed an increase of up to 12% in Arterial_{1×3} and 7% in Grid_{2×2}, while the ATTs for buses decreased by as much as 17% and 25%, respectively. This outcome demonstrates that the agent optimizes overall intersection performance without significantly impacting passenger cars.

In addition to ATT, the paper analyzes signal timing under the proposed RL-based TSP control method. Examining signal timing enhances the understanding of the control behavior of RL agents. This analysis provides valuable insights into how RL agents manage traffic signals effectively. Regarding cycle length, it was observed that an increase in the number of buses with higher PF correlates with longer cycle lengths. In terms of signal splits, the phase duration is extended when servicing directions with buses that possess PF, as anticipated, since the agent prioritizes the rapid passage of buses through intersections.

The final section of the paper discusses the implementation of a dynamic PF based on passenger load rather than a constant PF. The results indicate that as passenger numbers increase, buses experience reduced travel times through intersections. This finding underscores the broader applicability of a PF under various dynamic conditions.

7. Future Research

Future research should concentrate on determining the PF in real-time control scenarios. Further studies are warranted to explore various factors, such as passenger waiting times at downstream bus stops and bus headways, to inform the PF calculations for operational regularity and punctuality of transit systems. Additionally, examining how to integrate these factors into future models is recommended. It is also essential to investigate how to establish the PF within a reasonable range.

To apply the proposed TSP approach in real-time control, one of the primary challenges is the need for extensive and accurate data such as traffic patterns, bus movements, passenger loads, and surrounding vehicles. Implementing a robust data collection framework, including the use of sensors, cameras, and GPS technology, can be resource-intensive and may require collaboration with local transit agencies and municipalities. Additionally, concerns related to data privacy and security must be addressed to ensure compliance with regulations.

Author Contributions: Conceptualization, H.-K.C., K.-P.K. and K.-I.W.; methodology, H.-K.C.; simulation, H.-K.C.; validation, H.-K.C., K.-P.K. and K.-I.W.; formal analysis, H.-K.C.; resources, K.-P.K. and K.-I.W.; data curation, H.-K.C.; writing—original draft preparation, H.-K.C.; writing—review and editing, H.-K.C., K.-P.K. and K.-I.W.; visualization, H.-K.C.; supervision, K.-P.K. and K.-I.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used in this research are presented in this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xu, M.; Ye, Z.; Sun, H.; Wang, W. Optimization model for transit signal priority under conflicting priority requests. *Transp. Res. Rec. J. Transp. Res. Board* **2016**, *2539*, 140–148. [[CrossRef](#)]
2. Zeng, X.; Zhang, Y.; Balke, K.N.; Yin, K. A real-time Transit Signal Priority Control Model considering stochastic bus arrival time. *IEEE Trans. Intell. Transp. Syst.* **2014**, *15*, 1657–1666. [[CrossRef](#)]
3. Wang, X.; Yin, Y.; Feng, Y.; Liu, H.X. Learning the max pressure control for urban traffic networks considering the phase switching loss. *Transp. Res. Part C Emerg. Technol.* **2022**, *140*, 103670. [[CrossRef](#)]

4. Abdulhai, B.; Pringle, R.; Karakoulas, G.J. Reinforcement Learning for True Adaptive Traffic Signal Control. *J. Transp. Eng.* **2003**, *129*, 278–285. [[CrossRef](#)]
5. Arel, I.; Liu, C.; Urbanik, T.; Kohls, A.G. Reinforcement learning-based multi-agent system for network traffic signal control. *IET Intell. Transp. Syst.* **2010**, *4*, 128–135. [[CrossRef](#)]
6. Chu, T.; Wang, J.; Codecà, L.; Li, Z. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 1086–1095. [[CrossRef](#)]
7. Levin, M.W. Max-pressure traffic signal timing: A summary of methodological and experimental results. *J. Transp. Eng. Part A Syst.* **2023**, *149*, 4. [[CrossRef](#)]
8. Ma, D.; Xiao, J.; Song, X.; Ma, X.; Jin, S. A back-pressure-based model with fixed phase sequences for traffic signal optimization under oversaturated networks. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 5577–5588. [[CrossRef](#)]
9. Levin, M.W.; Hu, J.; Odell, M. Max-pressure signal control with cyclical phase structure. *Transp. Res. Part C Emerg. Technol.* **2020**, *120*, 102828. [[CrossRef](#)]
10. Wei, H.; Chen, C.; Zheng, G.; Wu, K.; Gayah, V.; Xu, K. PressLight: Learning max pressure control to coordinate traffic signals in arterial network. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 1290–1298.
11. Wei, H.; Xu, N.; Zhang, H.; Zheng, G.; Zang, X.; Chen, C.; Zhang, W.; Zhu, Y.; Xu, K.; Li, Z. CoLight: Learning network-level cooperation for traffic signal control. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; pp. 1913–1922.
12. Wei, H.; Zheng, G.; Gayah, V.; Li, Z. A survey on traffic signal control methods. *arXiv* **2019**, arXiv:1904.08117. Available online: <http://arxiv.org/abs/1904.08117> (accessed on 1 September 2024).
13. Noaen, M.; Naik, A.; Goodman, L.; Crebo, J.; Abrar, T.; Abad, Z.S.H.; Bazzan, A.L.; Far, B. Reinforcement learning in urban network traffic signal control: A systematic literature review. *Expert Syst. Appl.* **2022**, *199*, 116830. [[CrossRef](#)]
14. Xu, T.; Barman, S.; Levin, M.W.; Chen, R.; Li, T. Integrating public transit signal priority into Max-Pressure Signal Control: Methodology and Simulation Study on a downtown network. *Transp. Res. Part C Emerg. Technol.* **2022**, *138*, 103614. [[CrossRef](#)]
15. Lin, Y.; Yang, X.; Zou, N.; Franz, M. Transit signal priority control at signalized intersections: A comprehensive review. *Transp. Lett.* **2014**, *7*, 168–180. [[CrossRef](#)]
16. Ngan, V.W.K. A comprehensive strategy for transit signal priority. Master's Thesis, University of British Columbia, Vancouver, BC, Canada, 2002.
17. Smith, H.R.; Hemily, P.B.; Ivanovic, M. *Transit Signal Priority (TSP): A Planning and Implementation Handbook*; ITS America: Washington, DC, USA, 2005.
18. Furth, P.G.; Muller, T.H. Conditional bus priority at signalized intersections: Better Service with less traffic disruption. *Transp. Res. Rec. J. Transp. Res. Board* **2000**, *1731*, 23–30. [[CrossRef](#)]
19. Mohammadi, R.; Roncoli, C.; Mladenovic, M.N. Transit Signal Priority in a connected vehicle environment: User throughput and schedule delay optimization approach. In Proceedings of the 2020 Forum on Integrated and Sustainable Transportation Systems (FISTS), Delft, The Netherlands, 3–5 November 2020. [[CrossRef](#)]
20. He, S.; Han, H.; Zhang, H.; Sun, S.; Qiu, T.Z. Connected Transit Bus Dynamic Priority Weight Modeling and conflicting request resolution control at the signalized intersection. *J. Adv. Transp.* **2022**, *2022*, 1–22. [[CrossRef](#)]
21. Ma, W.; Liu, Y.; Yang, X. A dynamic programming approach for optimal signal priority control upon multiple high-frequency bus requests. *J. Intell. Transp. Syst.* **2012**, *17*, 282–293. [[CrossRef](#)]
22. Xu, M.; An, K.; Ye, Z.; Wang, Y.; Feng, J.; Zhao, J.; Xu, M.; An, K.; Ye, Z.; Wang, Y.; et al. A bi-level model to resolve conflicting transit priority requests at Urban Arterials. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 1353–1364. [[CrossRef](#)]
23. Zeng, X.; Zhang, Y.; Jiao, J.; Yin, K. Route-based transit signal priority using connected vehicle technology to promote bus schedule adherence. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 1174–1184. [[CrossRef](#)]
24. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
25. Arulkumaran, K.; Deisenroth, M.P.; Brundage, M.; Bharath, A.A. Deep reinforcement learning: A brief survey. *IEEE Signal Process. Mag.* **2017**, *34*, 26–38. [[CrossRef](#)]
26. Chanloha, P.; Chinrungrueng, J.; Usaha, W.; Aswakul, C. Cell transmission model-based multiagent Q-learning for network-scale signal control with transit priority. *Comput. J.* **2013**, *57*, 451–468. [[CrossRef](#)]
27. Shabestray, S.M.A.; Abdulhai, B. Multimodal intelligent deep (MiND) traffic signal controller. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019; pp. 4532–4539. [[CrossRef](#)]
28. Cheng, H.K.; Kou, K.P.; Wong, K.I. Transit Signal Priority Control with Deep Reinforcement Learning. In Proceedings of the 2022 10th International Conference on Traffic and Logistic Engineering (ICTLE), Macau, China, 12–14 August 2022. [[CrossRef](#)]
29. Varaiya, P. Max pressure control of a network of signalized intersections. *Transp. Res. Part C Emerg. Technol.* **2013**, *36*, 177–195. [[CrossRef](#)]

-
30. Zheng, G.; Zang, X.; Xu, N.; Wei, H.; Yu, Z.; Gayah, V.; Xu, K.; Li, Z. Diagnosing reinforcement learning for traffic signal control. *arXiv* **2019**, arXiv:1905.04716.
 31. "PTV vissim", Traffic Simulation Software | PTV Vissim | PTV Group. Available online: <https://www.ptvgroup.com/> (accessed on 9 December 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.