

# Multi-hop Upstream Anticipatory Traffic Signal Control with Deep Reinforcement Learning

Xiaocan Li\*, Xiaoyu Wang<sup>†</sup>, Iliia Smirnov<sup>†</sup>, Scott Sanner\*, Baher Abdulhai<sup>†</sup>

<sup>1</sup>Department of Mechanical & Industrial Engineering, University of Toronto, Toronto, Canada

<sup>2</sup>Department of Civil & Mineral Engineering, University of Toronto, Toronto, Canada

CORRESPONDING AUTHOR: Xiaocan Li (e-mail: hsiatsan.li@mail.utoronto.ca).

**ABSTRACT** Coordination in traffic signal control is crucial for managing congestion in urban networks. Existing pressure-based control methods focus only on immediate upstream links, leading to suboptimal green time allocation and increased network delays. However, effective signal control inherently requires coordination across a broader spatial scope, as the effect of upstream traffic should influence signal control decisions at downstream intersections, impacting a large area in the traffic network. Although agent communication using neural network-based feature extraction can implicitly enhance spatial awareness, it significantly increases the learning complexity, adding an additional layer of difficulty to the challenging task of control in deep reinforcement learning. To address the issue of learning complexity and myopic traffic pressure definition, our work introduces a novel concept based on Markov chain theory, namely *multi-hop upstream pressure*, which generalizes the conventional pressure to account for traffic conditions beyond the immediate upstream links. This farsighted and compact metric informs the deep reinforcement learning agent to preemptively clear the multi-hop upstream queues, guiding the agent to optimize signal timings with a broader spatial awareness. Simulations on synthetic and realistic (Toronto) scenarios demonstrate controllers utilizing multi-hop upstream pressure significantly reduce overall network delay by prioritizing traffic movements based on a broader understanding of upstream congestion.

**INDEX TERMS** Traffic signal control, reinforcement learning, traffic pressure

## I. INTRODUCTION

**T**RAFFIC signal control (TSC) is a cornerstone of intelligent transportation systems, designed to optimize traffic flow at intersections, reduce congestion, and minimize delays. Traditional methods, such as pre-timed and actuated control, have been widely adopted [1]–[4], but they often struggle to adapt to dynamic and complex traffic conditions. To address these limitations, the concept of traffic pressure has emerged as a promising metric for adaptive signal control strategies. Traffic pressure, at the intersection level, quantifies the disparity in traffic statistics (e.g., vehicle count or density) between upstream and downstream links [5], [6], enabling more responsive control approaches [7]. For instance, PressLight [8] integrated traffic pressure into reinforcement learning (RL) agent’s reward design to improve network efficiency.

Despite these advancements, existing traffic pressure metrics remain limited in their spatial scope, focusing solely on immediate links at individual intersections while ignoring the

broader network context. As the minimal motivating example shown in Figure 1, at the right intersection, a myopic pressure-based controller would assign equal green time to eastbound and southbound flows because it perceives equal pressures from immediate upstream links. This approach neglects the longer queues and accumulating congestion further upstream on the eastbound route. This example is verified in numerical experiments in Section V. Such myopic decision-making exacerbates delays and reduces overall network efficiency, highlighting the need for a farsighted metric that accounts for multi-hop upstream conditions.

The goal of this work is to develop a generalized concept of traffic pressure that integrates *multi-hop upstream* conditions. This approach captures a more comprehensive view of traffic dynamics, allowing controllers to prioritize traffic movements that most effectively alleviate congestion. By integrating multi-hop upstream pressure into deep RL agent design, this work provides a farsighted and adaptive

framework that mitigates network delays and improves the overall performance of urban traffic networks.

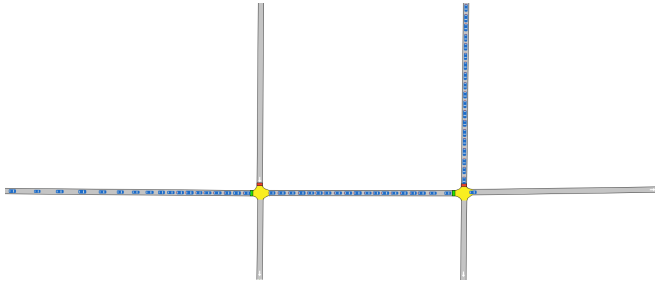


FIGURE 1: A motivating example to demonstrate the issue of existing myopic pressure definition and the need for farsighted multi-hop upstream pressure definition.

The contribution of this work are three folds:

- *Generalization of Traffic Pressure*: This paper introduces a novel concept of multi-hop upstream pressure grounded in Markov chain theory, which extends the conventional myopic traffic pressure to account for upstream conditions beyond immediate incoming links. This novel metric incorporates a broader spatial awareness than the traditional counterpart.
- *Integration into Deep Reinforcement Learning*: The multi-hop upstream pressure is integrated into the deep reinforcement learning framework, informing the agent’s observation and reward spaces. This encourages preemptive queue clearance and more effective signal timing optimization based on upstream traffic conditions, addressing the limitations of existing RL controllers based on traditional pressure definition.
- *Comprehensive Validation*: The effectiveness of the proposed approach is validated through extensive simulations on both synthetic and realistic traffic scenarios, including a Toronto-based case study. Results show significant reductions in overall network delays, demonstrating the practical advantages of using multi-hop upstream pressure for traffic signal control.

## II. LITERATURE REVIEW

In this section, we provide literature review on traffic signal control, and the variations of traffic pressures and their applications.

### A. Multi-intersection Traffic Signal Coordination

#### a: Traditional Traffic Signal Control

Traditional traffic signal control methods primarily focus on signal progression to optimize traffic flow. Pre-timed approaches, such as GreenWave [2] and Maxband [3], synchronize offsets across intersections to reduce vehicle stops in specific directions. While effective for stable traffic patterns, these methods lack adaptability to dynamic conditions.

Actuated and classical adaptive systems enhance real-time flexibility. Actuated control adjusts signals based on imme-

diately traffic detection, but its myopic nature limits network-level coordination [9]. Adaptive systems like SCATS [10] and SCOOT [4] expand coordination regionally or hierarchically but rely on pre-designed models, reducing their effectiveness in highly dynamic environments.

#### b: Reinforcement Learning based Adaptive Control

Reinforcement Learning (RL) has emerged as a promising approach for adaptive traffic signal control, leveraging data-driven techniques to optimize signal timings dynamically. RL-based methods are categorized into centralized, hierarchical, and decentralized structures.

- *Centralized Control*: A single agent observes and controls the entire network [11]. While this approach achieves network-level coordination, it struggles with scalability in large networks.
- *Hierarchical Control*: Multiple levels of agents are deployed, with upper-level agents providing macroscopic instructions and lower-level agents making finer decisions [12]–[14]. Hierarchical control balances scalability with coordination but requires careful design to ensure smooth interaction between agent levels.
- *Decentralized Control*: Each intersection is controlled by an independent agent [8], making the system highly scalable. However, the lack of inherent coordination can lead to suboptimal network performance. Strategies to enhance coordination include:
  - *Centralized Training with Decentralized Execution (CTDE)*: This approach trains agents jointly using shared information while maintaining decentralized execution during deployment [15].
  - *Agent Communication*: Communication frameworks allow agents to exchange local traffic states and coordinate actions, improving global performance. Neighbor RL [16] directly concatenates immediate neighbor intersections’ information. GC-NRL [17] uses Graph Convolutional Networks (GCN) to extract features across intersections. CoLight [18] leverages Graph Attentional Networks (GAT) to facilitate communication. eMARLIN [19], [20] embeds immediate neighbor intersections information into an embedding space. The reward designs of all these methods are only associated with local intersections, making these agents less farsighted.

Effective decentralized RL for TSC relies heavily on the design of agent observations and rewards, as the information available to each agent directly impacts its ability to make informed decisions. For example, PressLight [8] integrated traffic pressure into reward design. However, the vanilla traffic pressure is limited in its scope to immediate neighbors, while conditions beyond immediate links are critical for intersection coordination. In addition, feature extraction via neural networks for agent communication [18], [19] imposes

additional computational overhead and learning difficulties that further complicate the control task based on deep reinforcement learning. *This motivates the exploration of efficient and effective observation and reward designs that capture broader traffic conditions.* Our approach complements existing agent communication frameworks that extract shared information via neural networks.

### B. Traffic Pressure and Its Variations

The traffic pressure concept originates from resource reallocation strategies in wireless communication networks [21]. The primary application of traffic pressure is the MaxPressure control policy, which determines phase activation [5], [6], [22]–[24] or green time allocation [7], [25]–[27] in decentralized traffic control systems. MaxPressure has also been integrated into perimeter control strategies [7], [28] that prevent regional congestion by restricting the inflow to protected regions. While effective, some implementations of MaxPressure with phase activation-based action spaces have raised concerns about confusing phase sequences, which could frustrate drivers and increase safety risks. Solutions include fixed or variable cycle times and predefined phase orders, combined with stability guarantees [26], [29]. Further enhancements include integrating vehicle rerouting into MaxPressure for improved performance [23].

#### a: Variations in Traffic Pressure Definition

Numerous variations of traffic pressure have been developed, focusing on specific traffic statistics:

- **Queue Density:** Incorporating link lengths into pressure calculation ensures that shorter links with queues are prioritized over longer links with the same queue length [25]. This pressure definition is also used in reward design for RL-based signal control [8].
- **Phase Weights:** To prioritize specific phases, dynamic weights are introduced [30], [31], along with adaptive estimation of turning ratios and saturation flows.
- **Delay Time:** To improve fairness in waiting times, traffic delay is included in pressure definitions [24], [32].
- **Travel Time:** Recognizing the difficulty in measuring queues, travel times have been used as proxies for pressure definitions and tested in both simulation and real-world settings [27].
- **Platoon and Occupancy Prioritization:** C-MP incorporates space mean speed to prioritize large moving platoons [33], OCC-MP prioritizes high-occupancy vehicles to improve passenger-based efficiency [34], and PQ-MP integrates pedestrian queues to account for mixed traffic scenarios [35].

#### b: Multi-hop Extensions

Traffic pressure has been extended to multi-hop downstream applications for perimeter control. For instance, N-MP de-prioritizes phases when multi-hop downstream link densities

exceed a critical threshold [28], and [36] formalizes multi-hop downstream pressure grounded on Markov chain theory. However, these approaches primarily focus on downstream conditions, neglecting upstream traffic dynamics.

Capturing the potential of *upstream* traffic is crucial for preemptively clearing queues. To the best of our knowledge, existing pressure-based works only consider immediate upstream links, without extending the pressure's scope to further upstream conditions. To address the limitation of upstream scope, this work introduces a novel concept of *multi-hop upstream pressure*, grounded in Markov chain theory. The proposed metric is integrated into the observation space and reward function of deep reinforcement learning agents, enabling preemptive signal timing optimization and effective coordination across intersections.

## III. PROBLEM STATEMENT

### A. Traffic Signal Control as Decentralized Markov Decision Processes

In this work, traffic signal control is modeled as a Decentralized Markov Decision Process (DecMDP), which is defined by the tuple  $(n, \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$ , where  $\mathcal{S} = \cup_{i=1}^n \mathcal{O}_i$  is the system state space as joint observation spaces of  $n$  agents, and  $\mathcal{A} = \cup_{i=1}^n \mathcal{A}_i$  represents the joint action of  $n$  agents:

- **Observation space  $\mathcal{O}_i$ :** For each intersection  $i$  controlled by agent  $i$ , the observation space consists of the multi-hop pressure associated with each phase. For instance, if an intersection has an eastbound phase and a southbound phase, then the observation space is two-dimensional, representing the pressure for the eastbound and southbound phases, respectively. Mathematical definition for observation space can be found in Section IV-D.
- **Action space  $\mathcal{A}_i$ :** For each intersection  $i$  controlled by agent  $i$ , the action is defined as the cycle splits, representing the proportion of green time allocated to each phase at the intersection.
- **Transition probability  $\mathcal{T}$ :** The transition probability  $\mathcal{T}(s'|s, a)$  represents the probability of transitioning from the current global state  $s$  to a new global state  $s'$  after the joint action  $a = (a_1, a_2, \dots, a_n)$  is taken by the  $n$  agents. This probability models the dynamics of traffic flow in response to changes in signal timings. In our setting,  $\mathcal{T}$  is handled by the traffic simulator and is not exposed to agents.
- **Reward  $\mathcal{R}_i$ :** For each intersection  $i$  controlled by agent  $i$ , the reward is calculated as the negative sum of multi-hop potentials over phases. Mathematical definition for reward space can be found in Section IV-E. This reward design encourages each agent to clear upstream traffic as quickly as possible. It is important to note that a myopic pressure-based reward may lead to undesirable behavior, such as holding vehicles upstream to minimize the myopic pressure.

- **Discount factor  $\gamma$** : The discount factor  $\gamma \in [0, 1]$  determines the relative importance of future rewards compared to immediate rewards.

Reinforcement learning is employed to solve this MDP by training each agent  $i$  to learn an optimal policy  $\pi_i$ , which maximizes the expected discounted cumulative reward  $\mathbb{E}_{a_i \sim \pi_i} \left[ \sum_{k=0}^{\infty} \gamma^k R_i(o_i^{(t+k)}, a_i^{(t+k)}) \right]$ . Through repeated interactions with the environment, each agent observes the traffic conditions, selects actions, receives rewards, and updates its policy to improve long-term traffic efficiency.

In this study, we utilize the Proximal Policy Optimization (PPO) algorithm [37], a widely used RL algorithm known for its stability and efficiency.

#### IV. METHODOLOGY

This section outlines the framework for implementing a generalized multi-hop pressure model in traffic signal control. We first clearly define the mathematical notations that are used throughout this work in Table 1. Then, we model the traffic network structure with graph representations. Finally, we define and illustrate the multi-hop pressure and multi-hop potential metrics, both in its scalar and vectorized forms, and demonstrate its calculation through a simplified example.

##### A. Graph Representations of Traffic Networks

The traffic network is represented as a graph described in Definition 1. To simplify this representation, a supersink  $\Omega$  is introduced, consolidating all destinations into a single abstract node. Incorporating the supersink allows the adjacency matrix to exhibit the properties of a Markov chain transition matrix, enabling mathematical operations on the adjacency matrix to be interpreted through the Markov chain theory. The supersink is characterized by the following properties:

- **Zero Queue Density**: With infinite capacity, the supersink's queue density is always zero.
- **Absorption**: Links connected to the supersink are fully absorbed, and the supersink remains its own downstream neighbor. This property establishes vehicle movement on the graph as an Absorbing Markov Chain.
- **Binary Turning Ratio**: The turning ratio for any link connected to the supersink or for transitions within the supersink itself is 1, and 0 for all other cases.

**Definition 1** (Graph representation). *The graph representation  $G^e = (\mathcal{V}^e, \mathcal{E}^e)$ , where:*

- *The extended link set  $\mathcal{V}^e$  additionally includes a supersink vertex  $\Omega$ , i.e.,  $\mathcal{V}^e = \mathcal{V} \cup \{\Omega\}$ .*
- *The extended edge set  $\mathcal{E}^e$  additionally includes those edges reflecting connections to the supersink.*
- *Edge weight  $T_{uv}$  is the turning ratio from link  $u$  to link  $v$ . These weights are derived from real empirical data or traffic simulations, representing the probability of traffic flow transitions between links.*

To assist understanding of graph representations, an example is provided with a simplistic traffic network with 8 links in Figure 2a, and is mapped onto its graph representation depicted in Figure 2b, where the turning ratios are labeled on edges.

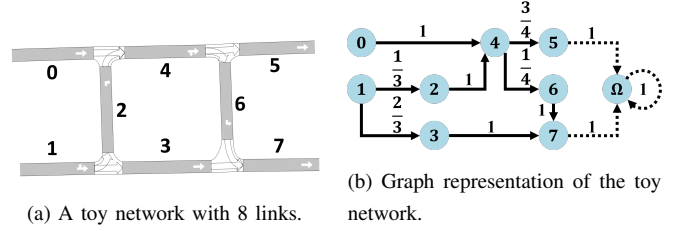


FIGURE 2: An example of graph representation for a toy traffic network. (a) A traffic network with 8 traffic links indexed from 0 to 7. (b) The weights shown on the edges are the fabricated turning ratios. The vertex  $\Omega$  is the supersink, and the edges in dashed lines represent graph  $G^e$  being extended from graph  $G$ .

##### B. Vehicle Movement as an Absorbing Markov Chain

The movement of a vehicle within a traffic network, guided by specific turning ratios, can be modeled as a time-homogeneous absorbing Markov chain. In this model, the presence of a vehicle on a link  $l$  is represented as a random variable with probability  $Pr(x = l)$ . The state space of the Markov chain is finite, comprising  $|\mathcal{V}^e|$  states, corresponding to the total number of links in the network. The transition matrix  $\mathbf{P}$  corresponds to the weighted adjacency matrix of the graph  $\mathcal{G}^e$ , defined as follows:

$$\mathbf{P} = \begin{bmatrix} T_{11} & \dots & T_{1|\mathcal{V}|} & T_{1\Omega} \\ \vdots & \ddots & \vdots & \vdots \\ T_{|\mathcal{V}|1} & \dots & T_{|\mathcal{V}||\mathcal{V}|} & T_{|\mathcal{V}|\Omega} \\ 0 & \dots & 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{T} & \mathbf{T}_{*\Omega} \\ \mathbf{0}^\top & 1 \end{bmatrix}, \quad (1)$$

##### C. Multi-hop Upstream Pressure: A Customizable Metric for Far-reaching Upstream Traffic Condition

Congestion from immediate upstream links has a more direct and significant impact than congestion several blocks further upstream. Therefore, the multi-hop pressure definition needs to capture the *diminishing influence* of distant congestion while still accounting for its *cumulative effect* on the current traffic link. To understand the upstream links at higher hops, we provide an example of upstream links for link 7. Mathematically, they are written as:

$$\mathcal{N}_u(7, 0) = \{7\} \quad (2)$$

$$\mathcal{N}_u(7, 1) = \{3, 6\} \quad (3)$$

$$\mathcal{N}_u(7, 2) = \{1, 4\} \quad (4)$$

$$\mathcal{N}_u(7, 3) = \{0\} \quad (5)$$

$$\mathcal{N}_u(7, h) = \{\}, \quad h \geq 4, h \in \mathbb{N}^+ \quad (6)$$



TABLE 1: Mathematical notations.

Symbol	Definition
<b>Graph Representation Related Notations</b>	
$\mathcal{V}$	The set of the whole traffic network's links. This is also the set of graph vertices.
$\mathcal{E}$	The set of graph edges. Each edge represents a permissible turning movement.
$\Omega$	Supersink, an abstract node that merges all destinations.
$\mathcal{V}^e$	The extended set of traffic links, including the supersink compared to $\mathcal{V}$ . See Definition 1.
$\mathcal{E}^e$	The extended set of edges. See Definition 1.
$G = (\mathcal{V}, \mathcal{E})$	The graph representation of the traffic network without supersink.
$G^e = (\mathcal{V}^e, \mathcal{E}^e)$	The graph representation of the traffic network with supersink. See Definition 1.
$\mathcal{N}_u(l, h)$	The set of $h$ -hop upstream links from link $l$ . 0-hop means the link itself, i.e., $\mathcal{N}_u(l, 0) = \{l\}$ .
$\mathcal{N}_d(l, h)$	The set of $h$ -hop downstream links from link $l$ . 0-hop means the link itself, i.e., $\mathcal{N}_d(l, 0) = \{l\}$ .
<b>Pressure &amp; Potential Related Notations</b>	
$T_{ij}$ [unitless]	Turning ratio from link $i$ to $j$ . The sum of turning ratios from link $i$ to all its 1-hop downstream links must be 1: $\sum_{j \in \mathcal{N}_d(i, 1)} T_{ij} = 1 \quad \forall i$ and $0 \leq T_{ij} \leq 1 \quad \forall i, j$
$\mathbf{T} \in \mathbb{R}^{ \mathcal{V}  \times  \mathcal{V} }$	The weighted adjacency matrix of graph $G$ where the $(i, j)$ -entry is $T_{ij}$ .
$\mathbf{P} \in \mathbb{R}^{ \mathcal{V}^e  \times  \mathcal{V}^e }$	The weighted adjacency matrix of graph $G^e$ , which is also a Markov transition matrix. See Eq. (1) for details.
$Q(l)$ [veh]	Queue length of link $l$ . Default speed threshold in simulator: Queue entering: $2m/s$ , Queue exit: $4m/s$ .
$\mathbf{Q} \in \mathbb{R}^{ \mathcal{V}^e }$	The concatenation of queue lengths for links in $\mathcal{V}^e$ . The order matches the rows and columns of $\mathbf{P}$ .
$p(l, h)$ [veh/km]	The pressure with $h$ -hop upstream for link $l$ .
$\mathbf{p}(h) \in \mathbb{R}^{ \mathcal{V}^e }$	The concatenation of $h$ -hop pressure for all links. The arranging order matches that of $\mathbf{Q}$ .
$\phi(l, h)$ [veh/km]	The potential with $h$ -hop upstream for link $l$ .
$\Phi^{\text{up}}(h)$ [veh/km]	The concatenation of $h$ -hop potential for all links. The arranging order matches that of $\mathbf{Q}$ .
$\Phi^{\text{down}}$ [veh/km]	The concatenation of immediate downstream traffic potential for all links. The arranging order matches that of $\mathbf{Q}$ .
$L_{\text{in}}(i)$	The set of incoming links for intersection $i$ .
$L_{\text{in}}(i, \theta)$	The set of incoming links for phase $\theta$ in intersection $i$ .
$\Theta(i)$	The set phases in intersection $i$ controlled by RL.
$p(\theta)$ [veh/km]	Phase pressure for phase $\theta$ . See Definition 2.

The scalar formulation of multi-hop upstream pressure, designed to calculate the pressure for a single link, is less compact and is thus presented in the Appendix. In contrast, the vectorized formulation enables simultaneous computation of pressures for *all* links in the traffic network. This vectorized approach significantly improves computational efficiency compared to processing each link individually.

#### 1) Vectorized Version: Multi-hop Pressure for All Links

Unlike the scalar version could only compute pressure for one link at a time, the vectorized version can compute pressures for *all* links in the traffic network simultaneously, which accelerates the computation upon implementation compared to iterating over each link in the traffic network:

Recursive Form:

$$\mathbf{p}(0) = \mathbf{Q} - \mathbf{P}\mathbf{Q} \quad (7)$$

$$\mathbf{p}(h) = \mathbf{p}(h-1) + (\mathbf{P}^h)^\top \mathbf{Q}, \quad h \in \mathbb{N}^+ \quad (8)$$

Unrolled Form:

$$\mathbf{p}(h) = \sum_{h'=0}^h (\mathbf{P}^{h'})^\top \mathbf{Q} - \mathbf{P}\mathbf{Q}, \quad h \in \mathbb{N} \quad (9)$$

#### a: Interpretations of $\mathbf{P}^h$

The term  $\mathbf{P}^h$  in Eq. (8) deserves meticulous interpretation. In Markov chain theory, the entry  $(i, j)$  in the  $h$ -th power of the transition matrix  $\mathbf{P}^h$ , denoted as  $(\mathbf{P}^h)_{ij}$ , indicates the probability of transitioning from vertex  $i$  to vertex  $j$  in exactly  $h$  steps. In the context of a traffic network, where entries correspond to turning ratios,  $(\mathbf{P}^h)_{ij}$  represents the probability of a vehicle traveling from link  $i$  to link  $j$  through any possible sequence of  $h$  links. This implies that link  $i$  is one of the  $h$ -hop upstream links of link  $j$ , i.e.,  $i \in \mathcal{N}_u(j, h)$ .

- *h-hop influence*: The entry  $(\mathbf{P}^h)_{ij}$  quantifies the influence of link  $i$  on link  $j$  after  $h$  transitions. It reflects

the notion that traffic pressure propagates across the network, extending beyond local effects to distant links.

- *Decay of influence over hop*: As  $\mathbf{P}^h$  involves repeated multiplication of  $\mathbf{P}$ , the influence decreases with increasing  $h$  due to turning ratios being bounded within  $[0, 1]$ . This captures the natural attenuation of congestion effects over distance in a traffic network.
- *Pressure contribution*: Multiplying  $[(\mathbf{P}^h)^\top]_{:,j}$  by  $\mathbf{Q}$  takes weighted sums on the traffic condition (e.g., queue density) over all  $h$ -hop upstream links exerted on link  $j$ . High congestion at link  $i$ , combined with a significant  $[(\mathbf{P}^h)^\top]_{ij}$ , results in a substantial contribution to the pressure at link  $j$ . The cumulative contribution over all 0 to  $h$  hops upstream links, formalized as  *$h$ -hop upstream potential*, is discussed in Section IV-E.
- *Independent contribution*: The term  $(\mathbf{P}^h)^\top \mathbf{Q}$  captures the additional pressure exerted on a link by queues at  $h$ -hop upstream links, not included in  $(h-1)$ -hop upstream links. This isolates the unique contribution of the  $h$ -hop and highlights how congestion propagates spatially and temporally. Understanding this distinction is critical for designing controllers that mitigate congestion effectively using multi-hop pressure information.

#### D. Observation Space Design: Multi-hop Upstream Pressure for Phases

**Definition 2** (Phase Pressure). *Given a control plan for an intersection  $i$ , the phase pressure is defined as the summation of link pressure over all incoming links in phase  $\theta$ :*

$$p(\theta) = \sum_{l \in L_{in}(i, \theta)} p(l, h) \quad (10)$$

The observation  $o_i \in \mathcal{O}_i$  for agent  $i$  is the concatenation of phase pressure in intersection  $i$ :

$$o_i = \parallel_{\theta \in \Theta(i)} p(\theta) \quad (11)$$

#### E. Reward Design: Multi-hop Upstream Potentials

Adopted from physics, another perspective for the pressure in Eq. (9) is the difference of upstream potential and downstream potential:

$$\mathbf{p}(h) = \Phi^{\text{up}}(h) - \Phi^{\text{down}} \quad (12)$$

where  $\Phi^{\text{up}}(h)$  is the  $h$ -hop upstream traffic potential and  $\Phi^{\text{down}}$  is the immediate downstream traffic potential:

$$\Phi^{\text{up}}(h) = \sum_{h'=0}^h (\mathbf{P}^{h'})^\top \mathbf{Q} \quad (13)$$

$$\Phi^{\text{down}} = \mathbf{P}\mathbf{Q} \quad (14)$$

To encourage RL agents to clear upstream queues, the reward for agent  $i$  is defined as the negation of the  $h$ -hop upstream potential across all incoming links at intersection  $i$ :

$$r_i = - \sum_{l \in L_{in}(i)} \phi^{\text{up}}(l, h) \quad (15)$$

A higher number of upstream hops in the reward calculation encourages the agent to preemptively allocate longer green times to clear upstream queues, facilitating coordination across intersections. Notably, the number of hops used for observation is the same as that used in the reward.

*Difference to pressure-based reward*: For comparison purpose, we also provide pressure-based reward formalization as the negation of the  $h$ -hop upstream potential across all incoming links at intersection  $i$ :

$$r_i = - \sum_{l \in L_{in}(i)} p(l, h) \quad (16)$$

When  $h = 0$ , it is a special case of myopic pressure reward used in PressLight [8].

## V. EXPERIMENTAL SETUP

The proposed traffic signal control scheme is evaluated using the traffic simulator Aimsun [38]. This section provides detailed description of the experimental setup in traffic network architecture and the traffic demand.

### A. Tested Scenarios

Both synthetic and realistic scenarios are tested. When designing scenarios, we provide a wide range of complexity from the simplest scenario to a complicated one, to verify that our approach works on diverse scenarios.

#### 1) Synthetic Scenarios

The scenario design breaks down into two parts: traffic network and traffic demand. We designed two traffic networks and three traffic demand saturation levels, resulting in  $2 \times 3 = 6$  synthetic scenarios in total.

##### a: Traffic Networks

Two simplified traffic networks are synthesized, as shown in Figure 3:

- *Network 1x2*: A minimal network with two intersections, designed to validate the effectiveness of multi-hop upstream pressure, adhering to the philosophy of minimal viable product in scientific research.
- *Network 1x3*: An extension of Network 1x2, featuring three intersections along an arterial road.

*Link channelization and phasing scheme*: Both synthetic networks share the following settings: The distance between adjacent intersections is 100 meters. Each link is single-lane and restricted to through movements, with no turning lanes. All intersections are signalized, operating with two phases: eastbound movement and southbound movement. The cycle length is 90 seconds, including two 5-second interphases.

##### b: Traffic demands

The demand saturation level can be categorized into three levels in ascending order:

- *Undersaturated*: 50% of heavily saturated demand.
- *Slightly saturated*: 75% of heavily saturated demand.
- *Heavily saturated*: The demand profile is tabulated in Table 2, where the maximal traffic flow is 2700vph at the first 30 minutes, greatly exceeding the capacity of the intersection (approximately 1800vph). The southbound flow is at the rightmost intersection only.

TABLE 2: Heavily saturated demand profile for synthetic networks. Flow unit: vph. Time unit: minute.

Network	Direction	0 - 30	30 - 60	60 - 90	90 - 120
Network 1x2	EB	1800	0	0	0
	SB	900	0	0	0
Network 1x3	EB	1800	0	1000	0
	SB	900	900	900	0

One might question why a constant flow demand is not used in Network 1x3. The reason is that constant demand can be effectively managed by pre-timed constant control, thus a constant demand is insufficient to demonstrate the advantages of multi-hop upstream pressure. Instead, we design a dynamic demand that no constant controllers is optimal and highlighting the need for a more adaptive control.

For any of these 6 synthetic scenarios, the optimal controller is expected to allocate longer green times to the eastbound phase to accommodate greater EB demands at the rightmost intersection, while the other intersection(s) should consistently assign the maximum allowed green time to the eastbound phase, given the absence of southbound flow.

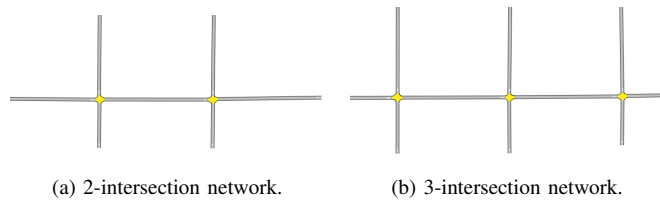


FIGURE 3: Tested synthetic arterial networks. Link channelization and phasing scheme are described in Section V-A.

## 2) Realistic Scenario

**Toronto testbed:** This testbed simulates a neighborhood around the intersection of Sheppard Avenue and Highway 404 in Toronto, Ontario, Canada, implemented in the Aimsun simulator (Figure 4). This neighborhood comprises 12 signalized intersections, including the pivotal Sheppard Avenue intersection near a major bus station and the on/off ramps of Highway 404. Out of these 12 signalized intersections, 4 congestion-prone intersections on the Sheppard Avenue are controlled by our method, while the rest 8 signalized intersections experiencing light demand are controlled by the

City Plan, as labeled in Figure 4. The area also includes the Fairview Mall, which features large parking lots adjacent to the arterial road. Distances between consecutive signalized intersections range from 150 to 300 meters. The demand profile spans the morning period from 7:30 to 10:00 AM, reaching its peak around 9 AM. The demand consists of three types of vehicles: cars, trucks, and buses. The buses are simulated following the schedules of two public transit service providers: Toronto Transit Commission and York Region Transit. The demand is calibrated using publicly available turning movement counts from the City of Toronto. The linear regression coefficient ( $R^2 = 0.9119$ ) indicates a good fit of our demand profile to real-world traffic.

### a: Baselines

We compare our approach against established traffic control baselines. These baselines represent different commonly used traffic signal control strategies, ranging from non-adaptive pre-timed control to advanced learning-based adaptive methods. The following outlines the key baseline methods used in our evaluation:

- *Pre-timed (non-adaptive) Control: Webster method (Synthetic Scenarios Only):* The cycle length and cycle splits are pre-defined according to historical flows.
- *Learning-based Adaptive Control: PressLight (Both Synthetic Scenarios and Toronto Testbed):* As a deep RL approach, PressLight leverages *myopic* pressure in reward design, where the *immediate* upstream and downstream traffic statistics are incorporated in pressure calculation.
- *Semi-Actuated Control: City Plan (Toronto Testbed Only):* A standard dual-ring NEMA phasing scheme with semi-actuated control [39]. This control plan is replicated based on the actual implementation. One may request the traffic signal timing information from the City of Toronto <sup>1</sup>.

### b: Evaluation Metrics

To comprehensively evaluate the performance of our proposed approach, we use three evaluation metrics that reflect different aspects of traffic statistics:

- *Total Time Spent (TTS) (hour):* Summation of each vehicle's travel time spent starting from vehicle generation to exit, including time in virtual queue.
- *Total Queue Time (Include Virtual) (hour):* Summation of each vehicle's queue time from vehicle generation to exit. Therefore, time in the virtual queue is included.
- *Total Virtual Queue Time (hour):* Summation of each vehicle's time spent in the virtual queue.

<sup>1</sup>Request Signal Timing Information: <https://www.toronto.ca/services-payments/streets-parking-transportation/traffic-management/traffic-signals-street-signs/request-signal-timing-information/>

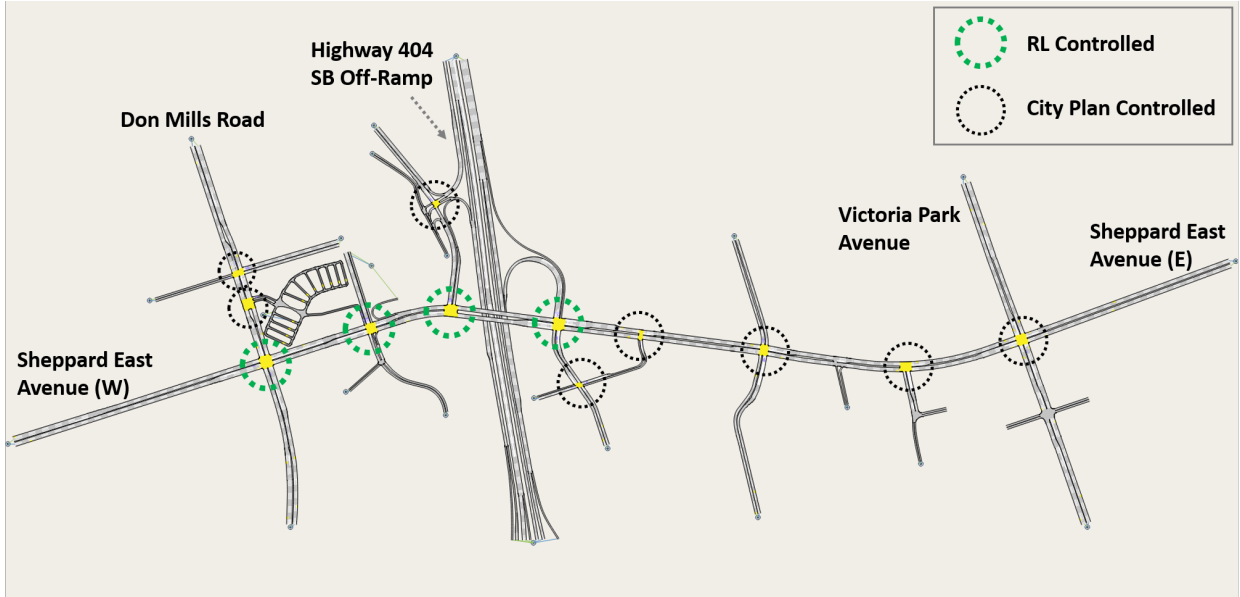


FIGURE 4: The Toronto network testbed. Four consecutive intersections on the Sheppard Avenue corridor are controlled by our methods as they encounter large flows. The other eight intersections not experiencing heavy congestion are less critical, therefore are controlled by the city plan.

## VI. RESULTS & DISCUSSION

### A. Synthetic Scenario: Network 1x2

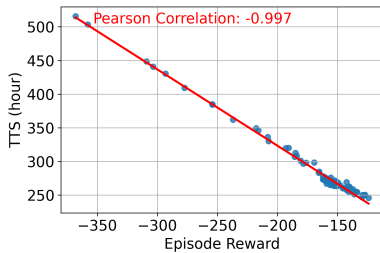


FIGURE 5: TTS vs Episode Reward.

*Learning Diagnosis:* To investigate the relationship between the original objective (TTS) and surrogate objective (episode rewards), Figure 5 demonstrates a near-perfect negative linear relationship (Pearson correlation: -0.997) between the episode reward defined in Eq. (15), and TTS in the network. This strong negative correlation indicates that maximizing the episode reward effectively leads to substantial reductions in TTS, thereby validating the use of multi-hop potentials as a reliable surrogate for traffic efficiency. Minimization of the multi-hop upstream potential indicates less congested upstream traffic.

The performance comparison between the proposed method and baselines are shown in Table 3. The proposed farsighted (1-hop) agent beats the pre-timed Webster method and the myopic RL method PressLight, under all demand levels. Webster method has the worst performance as it is not adaptive to the dynamic demand. The performance of

PressLight is similar but slightly worse compared to our method with 0-hop upstream in undersaturated and heavily saturated demand levels, because PressLight – with pressure-based reward – would deliberately hold vehicles upstream to achieve the minimization of pressure, whereas our potential-based reward encourages vehicles to move downstream.

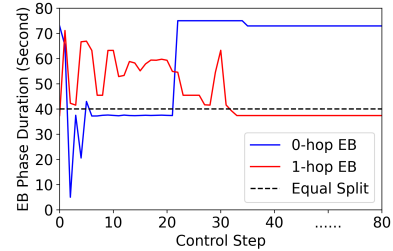


FIGURE 6: The allocated green time for eastbound phase for the right intersection in Network 1x2. The farsighted agent assigns longer green time for eastbound phase when queues exist before time step 30.

Figure 6 illustrates the control action (eastbound phase duration) over time for the right intersection in the Network 1x2 scenario. The southbound green time is automatically determined based on the remaining cycle time, as the total cycle length is fixed. The agent informed by myopic upstream pressures (0 hop) learned near equal splits for the eastbound and southbound phases, as the agent only observed equal queue lengths on immediate eastbound and southbound links, resulting in suboptimal green time allocation. In contrast, agents informed by farsighted pressures



TABLE 3: Performance comparison of all methods on synthetic networks

Network	Demand Level	Method	TTS	Total Queue Time (Include Virtual)	Total Virtual Queue Time
Network 1x2	Undersaturated	Webster	21.6	6.0	0.0
		PressLight	21.0	5.5	0.0
		Ours: 0-hop	20.8	5.3	0.0
		<b>Ours: 1-hop</b>	<b>20.7</b>	<b>5.2</b>	<b>0.0</b>
	Slightly Saturated	Webster	114.8	68.4	23.5
		PressLight	91.5	37.3	13.6
		Ours: 0-hop	88.6	34.8	15.1
		<b>Ours: 1-hop</b>	<b>76.7</b>	<b>36.7</b>	<b>1.7</b>
	Heavily Saturated	Webster	270.1	196.8	130.9
		PressLight	249.9	197.9	145.5
		Ours: 0-hop	242.6	187.6	138.0
		<b>Ours: 1-hop</b>	<b>221.5</b>	<b>155.8</b>	<b>64.4</b>
Network 1x3	Undersaturated	Webster	33.6	7.86	0.0
		PressLight	30.8	5.82	0.0
		Ours: 0-hop	30.1	5.17	0.0
		Ours: 1-hop	30.07	5.15	0.0
	<b>Ours: 2-hop</b>	<b>30.04</b>	<b>5.12</b>	<b>0.0</b>	
	Slightly Saturated	Webster	134.4	62.6	24.4
		PressLight	99.7	30.6	10.6
		Ours: 0-hop	104.0	32.5	12.6
		Ours: 1-hop	93.4	24.7	7.4
	<b>Ours: 2-hop</b>	<b>85.0</b>	<b>20.5</b>	<b>2.7</b>	
	Heavily Saturated	Webster	325.2	210.1	132.7
		PressLight	311.1	219.1	144.5
Ours: 0-hop		309.3	222.9	140.4	
Ours: 1-hop		293.7	199.9	134.4	
<b>Ours: 2-hop</b>	<b>272.9</b>	<b>173.3</b>	<b>78.9</b>		

(1 hop) as observations and rewards learn better control policies, consistently assigning greater green time to the eastbound phase, with an average of 2.7:1 green time splits for the eastbound and southbound phases, which is crucial for managing queues efficiently in this scenario.

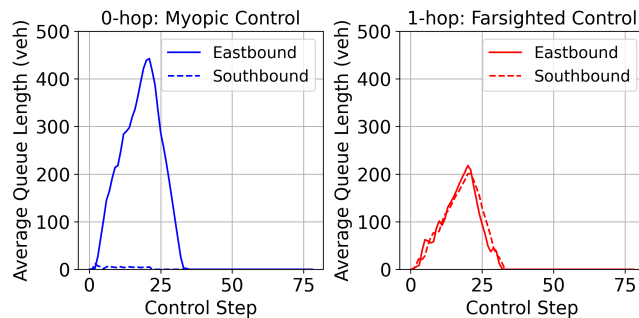


FIGURE 7: Comparison of average queue lengths over time between myopic (0-hop) and farsighted (1-hop) control.

Consistent with the control action in Figure 6, the queue lengths under the myopic and farsighted agents are compared in Figure 7. The myopic agent results in a sharp increase in

the queue length on the eastbound. In contrast, the farsighted agent has lower peak queue lengths and more balanced eastbound and southbound queue lengths. The sudden increase of eastbound phase duration for the myopic agent is because no new vehicles are generated after timestep 20, and southbound queues are cleared at timestep 22, thus the myopic agent can finally assign maximal allowed green time to clear eastbound queues. This comparison highlights the advantage of multi-hop upstream pressures and potentials into the RL agent design that significantly reduce congestion in tested scenarios.

### B. Synthetic Scenario: Network 1x3

Figure 8 presents the impact of upstream hop information on total time spent (TTS) under three demand levels: (a) undersaturated, (b) slightly saturated, and (c) heavily saturated:

- In undersaturated conditions (Figure 8a), TTS remains nearly constant across 0-hop, 1-hop, and 2-hop scenarios, with minimal variation. This indicates that in low-demand conditions, the inclusion of additional upstream hop information has no adverse impact on performance. Since the network is not congested, simpler control us-

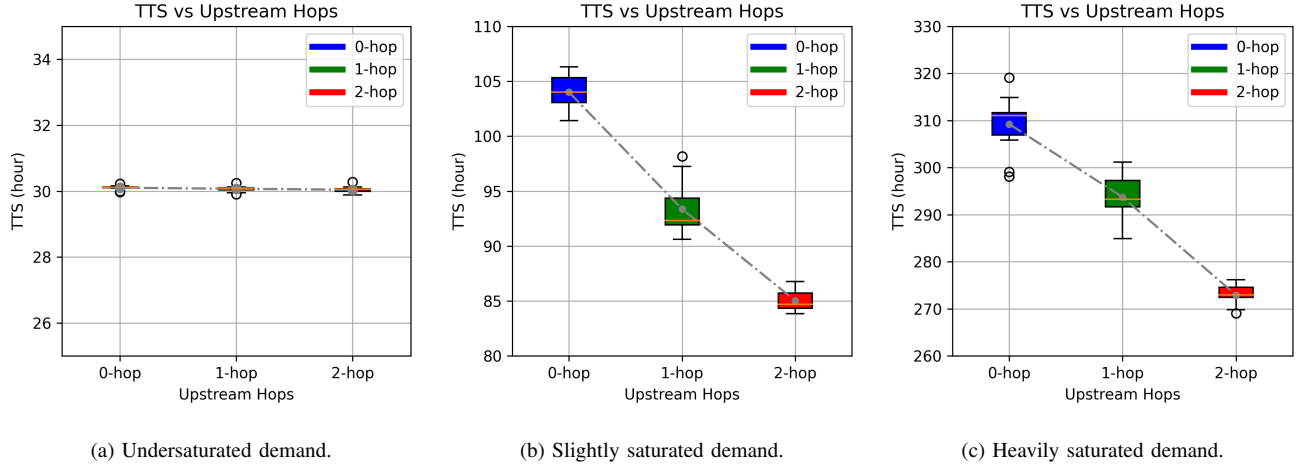


FIGURE 8: TTS vs upstream hops for (a) *undersaturated* demand, (b) *slightly saturated* demand, (c) *heavily saturated* demand. Farsighted pressures contribute to an improved performance when the demand is saturated, and does no harm to undersaturated cases.

ing immediate pressures (0-hop) is sufficient to achieve optimal performance.

- In slightly saturated conditions (Figure 8b), TTS begins to show variation across the different hop levels. The 1-hop and 2-hop configurations achieve lower TTS than the 0-hop scenario, with the 2-hop configuration yielding the lowest TTS. This suggests that under moderate congestion, farsighted setups (1-hop and 2-hop) enable the agents to coordinate more effectively, reducing delays by considering upstream traffic conditions.
- In heavily saturated conditions (Figure 8c), the benefits of using farsighted pressure still exist. The TTS decreases progressively from 0-hop to 2-hop. This demonstrates that when the network is heavily congested, incorporating additional upstream information better manages the traffic.

The results indicate that farsighted pressure (1-hop and 2-hop) provides substantial benefits in saturated and heavily saturated scenarios by enabling more effective green time allocation based on upstream congestion levels. In contrast, in undersaturated scenarios, the additional upstream information does not affect performance, as immediate pressures alone suffice to maintain optimal flow. Therefore, multi-hop pressure control improves traffic efficiency for saturated conditions without detriment to undersaturated cases.

### C. Realistic Scenario: Toronto Testbed

In this section, we discuss the results for the Toronto testbed.

Figure 9 compares the performance of potential-based reward and the pressure-based reward, previously defined in Eq. (15) and Eq. (16). Although both PressLight (0-hop pressure-based reward) and our 0-hop potential-based RL method are myopic, PressLight exhibits higher variability (less robustness) in performance. This may be attributed to

its pressure-based reward design, which can unintentionally encourage the agent to hold vehicles upstream to minimize local pressure, leading to higher TTS and instability. The higher variability of RL agents using pressure-based rewards has also been observed in empirical evaluations [40], [41]. In contrast, our potential-based reward consistently incentivizes the agent to release vehicles to downstream links, promoting smoother traffic flow and reducing variability. *In later results and discussions, we only focus on potential-based reward.*

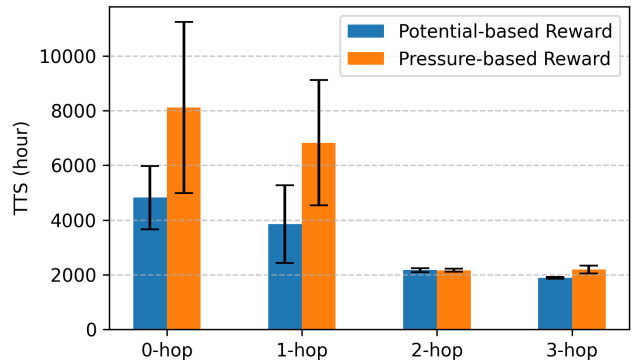


FIGURE 9: Potential-based vs pressure-based reward design.

Figure 10 compares the TTS among our methods and the City Plan baseline on the Toronto testbed, cross validated by queue time comparison in Table 4. Compared to the City Plan (2303 hours), Our RL models using 0-hop and 1-hop upstream setups show higher TTS (4383 and 3214 hours, respectively). The increased total and virtual queue times indicate that these models struggle due to their myopic observation and reward. In contrast, the 2-hop and 3-hop RL models outperform the City Plan, with 6% and 19%

improvement on TTS, respectively. The results demonstrated the benefit of incorporating multi-hop upstream traffic, enabling a more coordinated signal control.

TABLE 4: Average performance on the Toronto testbed

Method	TTS	Total Queue Time (Include Virtual)	Total Virtual Queue Time
City Plan	2303	1312	367
PressLight	8116	7501	2809
Ours: 0-hop	4383	2640	1110
Ours: 1-hop	3214	2032	83
Ours: 2-hop	2170	1104	89
<b>Ours: 3-hop</b>	<b>1878</b>	<b>880</b>	<b>3</b>

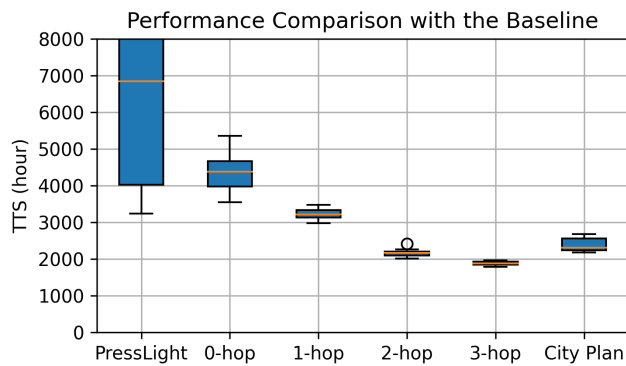


FIGURE 10: Performance comparison between our methods and the baseline. The variability comes from 10 different replications by setting 10 unique random seeds.

The heatmap in Figure 11 highlights the differences in queue lengths between the City Plan and our RL approach using 3-hop upstream setup. The compared queue lengths are average values throughout the whole simulation, better reflecting the congestion than just comparing on a time slice. Green roads indicate reduced queue lengths with the RL approach, while red roads show increased queues.

TABLE 5: Top 4 largest number of trips origins.

Origin	Trips (%)	Trips (veh)
Highway 404 N	15.5%	4241
Don Mills Road N	12.5%	3415
Don Mills Road S	9.4%	2579
Sheppard Avenue W	9.3%	2555

A key improvement is observed at the Highway 404 southbound off-ramp. The RL approach focuses on optimizing the Highway 404 off-ramp, which is supported by the trip distribution data in Table 5 that 15.5% of local trips originate from the Highway 404 North origin, which is also the largest trip numbers, making it a critical point for signal control.

Overall, most roads benefit from our proposed approach, indicating improved traffic flow. While the southbound link connected to the Fairview Mall (second left intersection) and the westbound link (third left intersection) experience increased queues as a result of competition among different signal phases, these increased queues are compensated by the significant improvement in prioritizing the Highway 404 SB off-ramp. This farsighted prioritization obtains a substantial gain rather than merely a trade-off.

Our RL approach also learns intersection coordination: traffic from the Highway 404 off-ramp, which feeds into Sheppard Avenue, requires adjacent intersections (the second and fourth RL-controlled intersections counting from the left) to coordinate by allocating longer green times along the arterial road to manage the incoming flow effectively. Our RL approach achieved this coordination, as the two adjacent RL-controlled intersection’s incoming links are greatly improved on queues, rendered in green.

## VII. CONCLUSION

This paper introduces a novel concept of multi-hop upstream pressure and integrates it to RL agent design to address the limitations of myopic pressure-based control methods that focus solely on immediate upstream links. The proposed multi-hop upstream pressure accounts for an abstracted view over a greater upstream area beyond the immediate upstream link, providing a broader spatial awareness for optimizing signal timings and achieving coordination.

Our experiments, conducted in both the synthetic scenario and the realistic Toronto testbed scenario, demonstrate that the RL agents utilizing the multi-hop upstream metric perform better in reducing network delays compared to myopic approaches. Notably, the approach performs exceptionally well in oversaturated scenarios and remains effective in undersaturated scenarios, benefiting from multi-hop information from further upstream links.

Future work could refine this approach by incorporating dynamic turning ratio estimation, expanding the scope to more complex networks, and applying multi-hop pressure to other traffic control problems, such as congestion pricing and dynamic perimeter identification.

## APPENDIX

### Scalar Version: Multi-hop Upstream Pressure for a Single Link

To gently guide the reader step-by-step, we first review the vanilla version of traffic pressure, then demonstrate how to extend it to higher-hop upstream versions.

#### a: Pressure with 0-hop Upstream:

Adopted from physics, the existing standard traffic pressure [5] is defined as the difference between immediate upstream queue length and the summation of immediate downstream queue lengths weighted by turning ratios. Mathematically,



FIGURE 11: The heatmap of queue difference between the City Plan and our RL approach with 3-hop upstream setup. The road-wise comparison is based on the average queue length over the whole simulation time. Roads labeled in green indicate improved queue length in our approach compared to the City Plan, while roads in red indicate longer queues.

for a link  $l$ , its pressure with 0-hop upstream is:

$$p(l, 0) = Q(l) - \sum_{j \in \mathcal{N}_u(l, 1)} T_{lj} Q(j) \quad (17)$$

Apparently, 0-hop upstream is myopic in terms of knowing the traffic conditions beyond the link of interest. When considering further neighborhoods, the concept of 0-hop upstream can be extended to multi-hop upstream to account for a wider range of traffic networks, capturing the cumulative effect of traffic congestion in neighboring areas.

#### b: Pressure with 1-hop Upstream:

Compared to pressure with 0-hop upstream, extra traffic information from 1-hop upstream links is integrated. The influence of 1-hop upstream links on the current link  $l$  is naturally weighted by the turning ratio from 1-hop upstream links to link  $l$ :

$$p(l, 1) = p(l, 0) + \sum_{i_1 \in \mathcal{N}_u(l, 1)} T_{i_1 l} Q(i_1) \quad (18)$$

$$= p(l, 0) + (\mathbf{P}^1)_{:,l}^\top \mathbf{Q} \quad (19)$$

#### c: Pressure with 2-hop Upstream:

Similarly, the impact of 2-hop upstream links is added upon pressure with 1-hop upstream. The congestion at 2-hop upstream links has less influence than 1-hop upstream links on the current link  $l$ , and is naturally discounted by the

turning ratio from 2-hop upstream links to link  $l$ :

$$p(l, 2) = p(l, 1) + \sum_{i_1 \in \mathcal{N}_u(l, 1)} \sum_{i_2 \in \mathcal{N}_u(i_1, 1)} T_{i_2 i_1} T_{i_1 l} Q(i_2) \quad (20)$$

$$= p(l, 1) + [(\mathbf{P}^2)_{:,l}]^\top \mathbf{Q} \quad (21)$$

#### d: Pressure with $h$ -hop Upstream:

To generalize, the impact of  $h$ -hop upstream links is added upon pressure with  $(h - 1)$ -hop upstream. The decay of influence from  $h$ -hop upstream links to link  $l$  is captured by the turning ratio from  $h$ -hop upstream links to link  $l$ :

$$p(h, 1) = p(l, h - 1) + \sum_{i_1 \in \mathcal{N}_u(l, 1)} \sum_{i_2 \in \mathcal{N}_u(i_1, 1)} \dots \sum_{i_h \in \mathcal{N}_u(i_{h-1}, 1)} T_{i_h i_{h-1}} \dots T_{i_2 i_1} T_{i_1 l} Q(i_h) \quad (22)$$

$$= p(l, h - 1) + [(\mathbf{P}^h)_{:,l}]^\top \mathbf{Q} \quad (23)$$

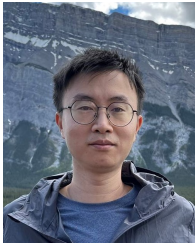
## ACKNOWLEDGMENT

### REFERENCES

- [1] F. WEBSTER, "Traffic signal settings," *Road Research Technical Paper*, 1958.
- [2] A. Warberg, J. Larsen, and R. M. Jørgesen, *Green wave traffic optimization-a survey*. Informatics and Mathematical Modeling, Technical University of Denmark, 2008.
- [3] J. D. Little, M. D. Kelson, and N. H. Gartner, "Maxband: A versatile program for setting signals on arteries and triangular networks," 1981.
- [4] P. Hunt, D. Robertson, R. Bretherton, and M. C. Royle, "The scoot on-line traffic signal optimisation technique," *Traffic Engineering & Control*, vol. 23, no. 4, 1982.
- [5] P. Varaiya, "Max pressure control of a network of signalized intersections," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 177–195, 2013.

- [6] —, “The max-pressure controller for arbitrary networks of signalized intersections,” in *Advances in Dynamic Network Modeling in Complex Transportation Systems*. Springer, 2013, pp. 27–66.
- [7] D. Tsitsokas, A. Kouvelas, and N. Geroliminis, “Two-layer adaptive signal control framework for large-scale dynamically-congested networks: Combining efficient max pressure with perimeter control,” *Transportation Research Part C: Emerging Technologies*, vol. 152, p. 104128, 2023.
- [8] H. Wei, C. Chen, G. Zheng, K. Wu, V. Gayah, K. Xu, and Z. Li, “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*, 2019, pp. 1290–1298.
- [9] D. Ni, *Actuated Control*. Cham: Springer International Publishing, 2020, pp. 211–224.
- [10] P. Lowrie, “Scats-a traffic responsive method of controlling urban traffic,” *Sales information brochure published by Roads & Traffic Authority, Sydney, Australia*, 1990.
- [11] L. Kuyer, S. Whiteson, B. Bakker, and N. Vlassis, “Multiagent reinforcement learning for urban traffic control using coordination graphs,” in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part I 19*. Springer, 2008, pp. 656–671.
- [12] M. Geng, S. Pateria, B. Subagdja, and A.-H. Tan, “Hisoma: A hierarchical multi-agent model integrating self-organizing neural networks with multi-agent deep reinforcement learning,” *Expert Systems with Applications*, vol. 252, p. 124117, 2024.
- [13] B. Zhou, Q. Zhou, S. Hu, D. Ma, S. Jin, and D.-H. Lee, “Cooperative traffic signal control using a distributed agent-based deep reinforcement learning with incentive communication,” *IEEE Transactions on Intelligent Transportation Systems*, 2024.
- [14] B. Xu, Y. Wang, Z. Wang, H. Jia, and Z. Lu, “Hierarchically and cooperatively learning traffic signal control,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 1, 2021, pp. 669–677.
- [15] S. El-Tantawy, B. Abdulhai, and H. Abdelgawad, “Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (marlin-atse): methodology and large-scale application on downtown toronto,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1140–1150, 2013.
- [16] I. Arel, C. Liu, T. Urbanik, and A. G. Kohls, “Reinforcement learning-based multi-agent system for network traffic signal control,” *IET Intelligent Transport Systems*, vol. 4, no. 2, pp. 128–135, 2010.
- [17] T. Nishi, K. Otaki, K. Hayakawa, and T. Yoshimura, “Traffic signal control based on reinforcement learning with graph convolutional neural nets,” in *2018 21st International conference on intelligent transportation systems (ITSC)*. IEEE, 2018, pp. 877–883.
- [18] H. Wei, N. Xu, H. Zhang, G. Zheng, X. Zang, C. Chen, W. Zhang, Y. Zhu, K. Xu, and Z. Li, “Colight: Learning network-level cooperation for traffic signal control,” in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 1913–1922.
- [19] X. Wang, A. Taitler, I. Smirnov, S. Sanner, and B. Abdulhai, “emarlin: distributed coordinated adaptive traffic signal control with topology-embedding propagation,” *Transportation Research Record*, vol. 2678, no. 4, pp. 189–202, 2024.
- [20] —, “emarlin+: Overcoming partial-observability caused by sensor limitations and short detection ranges in traffic signal control,” in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, 2023, pp. 2337–2342.
- [21] L. Tassiulas and A. Ephremides, “Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks,” in *29th IEEE Conference on Decision and Control*. IEEE, 1990, pp. 2130–2132.
- [22] T. Wongpiromsarn, T. Uthacharoenpong, Y. Wang, E. Frazzoli, and D. Wang, “Distributed traffic signal control for maximum network throughput,” in *2012 15th international IEEE conference on intelligent transportation systems*. IEEE, 2012, pp. 588–595.
- [23] A. A. Zaidi, B. Kulcsár, and H. Wymeersch, “Traffic-adaptive signal control and vehicle routing using a decentralized back-pressure method,” in *2015 European Control Conference (ECC)*. IEEE, 2015, pp. 3029–3034.
- [24] J. Wu, D. Ghosal, M. Zhang, and C.-N. Chuah, “Delay-based traffic signal control for throughput optimality and fairness at an isolated intersection,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 2, pp. 896–909, 2017.
- [25] A. Kouvelas, J. Lioris, S. A. Fayazi, and P. Varaiya, “Maximum pressure controller for stabilizing queues in signalized arterial networks,” *Transportation Research Record*, vol. 2421, no. 1, pp. 133–141, 2014.
- [26] T. Le, P. Kovács, N. Walton, H. L. Vu, L. L. Andrew, and S. S. Hoogendoorn, “Decentralized signal control for urban road networks,” *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 431–450, 2015.
- [27] P. Mercader, W. Uwayid, and J. Haddad, “Max-pressure traffic controller based on travel times: An experimental analysis,” *Transportation Research Part C: Emerging Technologies*, vol. 110, pp. 275–290, 2020.
- [28] H. Liu and V. V. Gayah, “N-mp: A network-state-based max pressure algorithm incorporating regional perimeter control,” *Transportation Research Part C: Emerging Technologies*, p. 104725, 2024.
- [29] M. W. Levin, J. Hu, and M. Odell, “Max-pressure signal control with cyclical phase structure,” *Transportation Research Part C: Emerging Technologies*, vol. 120, p. 102828, 2020.
- [30] N. Xiao, E. Frazzoli, Y. Li, Y. Luo, Y. Wang, and D. Wang, “Further study on extended back-pressure traffic signal control algorithm,” in *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE, 2015, pp. 2169–2174.
- [31] N. Xiao, E. Frazzoli, Y. Luo, Y. Li, Y. Wang, and D. Wang, “Throughput optimality of extended back-pressure traffic signal control algorithm,” in *2015 23rd Mediterranean Conference on Control and Automation (MED)*. IEEE, 2015, pp. 1059–1064.
- [32] H. Liu and V. V. Gayah, “A novel max pressure algorithm based on traffic delay,” *Transportation Research Part C: Emerging Technologies*, vol. 143, p. 103803, 2022.
- [33] T. Ahmed, H. Liu, and V. V. Gayah, “C-mp: A decentralized adaptive-coordinated traffic signal control using the max pressure framework,” *arXiv preprint arXiv:2407.01421*, 2024.
- [34] —, “Occ-mp: A max-pressure framework to prioritize transit and high occupancy vehicles,” *Transportation Research Part C: Emerging Technologies*, vol. 166, p. 104795, 2024.
- [35] T. Xu, Y. Bika, and M. W. Levin, “Ped-mp: A pedestrian-friendly max-pressure signal control policy for city networks,” *Journal of Transportation Engineering, Part A: Systems*, vol. 150, no. 7, p. 04024028, 2024.
- [36] X. Li, X. Wang, I. Smirnov, S. Sanner, and B. Abdulhai, “Generalized multi-hop traffic pressure for heterogeneous traffic perimeter control,” *arXiv preprint arXiv:2409.00753*, 2024.
- [37] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
- [38] Aimsun, *Aimsun Next 22 User’s Manual*, Aimsun Next 22.0.1 ed., Barcelona, Spain, Accessed on: May 20, 2023 2022. [Online]. Available: <https://docs.aimsun.com/next/22.0.1/>
- [39] FHWA, *Traffic Analysis Toolbox Volume IV: Guidelines for Applying CORSIM Microsimulation Modeling Software*. U.S. Department of Transportation Federal Highway Administration, 2007, ch. Appendix F: Actuated Signal Control.
- [40] V. Jayawardana, C. Tang, S. Li, D. Suo, and C. Wu, “The impact of task underspecification in evaluating deep reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 881–23 893, 2022.
- [41] Z. Zhang, M. Quinones-Grueiro, W. Barbour, Y. Zhang, G. Biswas, and D. Work, “Evaluation of traffic signal control at varying demand levels: A comparative study,” in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023, pp. 3215–3221.





**Xiaocan Li** received the B.Eng. degree from Beihang University, Beijing, China, in 2017, and the M.Sc. degree in Control Theory and Engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2020. He is currently a Ph.D. candidate with the Department of Mechanical & Industrial Engineering at the University of Toronto, ON, Canada. His research interests include deep reinforcement learning, spatiotemporal prediction, and their application to intelligent transportation systems.

(GO) Transit Authority from 2004 to 2006. He served as a Canada Research Chair (CRC) in ITS from 2005 to 2010. His research team won international awards including the International Transportation Forum innovation award in 2010 (Hossam Abdelgawad), IEEE ITS 2013 (Samah El-Tantawy) and INFORMS 2013 (Samah El-Tantawy). In 2015 he has been inducted as a Fellow of the Engineering Institute of Canada (EIC). In 2018, he won the prestigious CSCE Sandford Fleming (Career Achievement) Award for his contribution to transportation in Canada. He has been elected Fellow of the Canadian Academy of Engineering in 2020. In 2021, he won the Ontario Professional Engineers Awards (OPEA) Engineering Medal for career Engineering Excellence.



**Xiaoyu Wang** received the B.Eng. degree in automation from Tianjin University, Tianjin, China, in 2016, and the M.Sc. degree in control science and technology from Shanghai Jiao Tong University, Shanghai, China, in 2019. He is currently a Ph.D. candidate with the Department of Civil and Mineral Engineering at the University of Toronto, ON, Canada. His research interests include control, reinforcement learning, and their application to intelligent transportation and multi-agent systems.



**Iliia Smirnov** is a Research Associate at the Department of Civil and Mineral Engineering at the University of Toronto. He completed a Ph.D. in Pure Mathematics (Algebraic Geometry) at Queen's University, ON, Canada in 2020. His research interests include Planning, Reinforcement Learning, and Optimization in Intelligent Transportation Systems, as well as Enumerative Algebraic Geometry / Intersection Theory.



**Scott Sanner** received the B.Sc. degree in computer science from the Carnegie Mellon University, Pittsburgh, PA, USA, in 1999, the M.Sc. degree in computer science from Stanford University, Stanford, CA, USA, in 2002, and the Ph.D. degree in computer science from the University of Toronto, ON, Canada, in 2008. He is an Associate Professor in Industrial Engineering and Cross-appointed in Computer Science at the University of Toronto. Scott's research focuses on a broad range of AI topics spanning sequential decision-making and

applications of machine/deep learning to Smart Cities. Scott is currently an Associate Editor for the Machine Learning Journal (MLJ) and the Journal of Artificial Intelligence Research (JAIR). Scott was a co-recipient of paper awards from the AI Journal (2014), Transport Research Board (2016), CPAIOR (2018) and a recipient of a Google Faculty Research Award in 2020.



**Baher Abdulhai** received the Ph.D. degree in engineering from the University of California, Irvine, CA, USA, in 1996. He is a Professor in Civi Engineering at the University of Toronto, ON, Canada. He has 35 years of experience in transportation systems engineering and Intelligent Transportation Systems (ITS). He is the founder and Director of the Toronto Intelligent Transportation System Center, and the founder and co-Director of the i-City Center for Automated and Transformative Transportation Systems (iCity-

CATTS). He received several awards including IEEE Outstanding Service Award, Teaching Excellence award, and research awards from Canada Foundation for Innovation, Ontario Research Fund, and Ontario Innovation Trust. He served on the Board of Directors of the Government of Ontario