A Novel Self-Learning Intelligent Traffic Signal Control System for Congested Urban Areas

Samah El-Tantawy is a Postdoctoral Fellow in Civil Engineering Department, University of Toronto, ON, M5S 1A4, Canada (phone: 416-978-5049; fax: 416-978-5054; e-mail: samah.el.tantawy@utoronto.ca)

Samah El-Tantawy is a Professor in Civil Engineering Department, University of Toronto, ON, M5S 1A4, Canada (phone: 416-978-5049; fax: 416-978-5054; e-mail: baher.abdulhai@utoronto.ca)

Hossam Abdelgawad is a Postdoctoral Fellow in Civil Engineering Department, University of Toronto, ON, M5S 1A4, Canada (phone: 416-978-5049; fax: 416-978-5054; e-mail: hossam.abdelgawad@alumni.utoronto.ca). .

ABSTRACT

Population is steadily increasing worldwide resulting in intractable traffic congestion in urban dense areas. Adaptive Traffic Signal Control (ATSC) has shown strong potential to effectively alleviate urban traffic congestion by adjusting the signal timing plans in real-time in response to traffic fluctuations to achieve desirable objectives (e.g., minimize delay).The problem of coordinated ATSC is challenging due to the exponential growth in the number of joint timing plans to be explored as the network size grows. A novel Multi-Agent Reinforcement Learning for Integrated Network of Adaptive Traffic Signal Controllers (MARLIN-ATSC) system is designed, developed and evaluated at a large-scale urban network of 59 intersections in the lower downtown core of the City of Toronto for the morning rush hour. MARLIN-ATSC offers two possible modes: (1) independent mode, i.e. each intersection controller is independently working of other agents; and (2) integrated mode, where each controller coordinates the signal control actions with the neighbouring intersections. The MARLIN-ATSC large-scale application was conducted to examine two cases: 1) corridor-specific agent coordination, and 2) network-wide agent coordination. In the network-wide coordination experiments three scenarios were investigated: uniform demand profile, variable demand profile, and unfamiliar drivers (using a low percentage of familiar drivers). The results show unprecedented reduction in the average intersection delay ranging from 27% in mode 1 to 39% in mode 2 at the network level; and travel time savings of 15% in mode 1 and 26% in mode 2, along the busiest routes in downtown Toronto.

Keywords: Reinforcement Learning, Adaptive Traffic Signal Control, Multi-Agent Reinforcement Learning

INTRODUCTION

Population is steadily increasing worldwide and the GTHA is no exception. Consequently the demand for mobility is rapidly increasing and congestion is turning into a household daily chore, hampering not only our quality of life but also our economic competitiveness. When the growth in social and economic activities outpace the cash-strapped growth of transportation infrastructure, congestion is inevitable. Among the myriad of demand and supply management possibilities to combat congestion, Adaptive Traffic Signal Control (ATSC) target enhancing infrastructure efficiency by adjusting the traffic signal timings in real time in response to traffic fluctuations to achieve a chosen objective (e.g., minimize delay). ATSC, in general, has a great potential to outperform both pre-timed and actuated control (McShane *et al.* 1998). Existing ATSC systems, however, have non-trivial limitations that make them relatively inefficient, expensive, and difficult to upkeep, ultimately limiting their potential benefits. For instance, treating intersections as isolated nodes that are independent of neighboring intersections limits the efficiency gains of such technology. Therefore, optimally controlling the operation of multiple intersections simultaneously can be synergetic and beneficial. Such integration certainly adds more complexity to the system that science has not been able to resolve until very recently. Coordination has been typically approached in a centralized way (e.g., SCOOT (Hunt *et al.* 1981), TUC(Diakaki *et al.* 2002)) which is only feasible if communication channels amongst all intersections and the central control location are available, which is resource demanding. SCATS (Sims and Dobinson 1979) is another example of an adaptive signal control system that is hierarchical and distributed system in which an area is divided into smaller subsystems (in the range of 1–10 intersections) that perform independently. PRODYN (Farges *et al.* 1983), OPAC (Gartner 1983), RHODES (Head *et al.* 1992) are also examples of adaptive systems that are decentralized but their relatively complex computation schemes make their implementation costly (Bazzan 2009).

The coordination mechanism in the systems above is employed along an arterial (where the major demand is). Although it is important to efficiently operate traffic signals along arterials where the major demand is (e.g., progression), it is also important to consider the network-wide effect of such operation. In a signalized urban network setting, considering a network-wide objective has the potential to improve overall network performance, mobility, and reduce emissions.

Based on recent scientific advances in the last decade, coordination can be plausibly achieved using reinforcement learning and game-theoretic approaches (Bazzan 2009). Reinforcement Learning (RL), from Artificial Intelligence (AI), has shown good potential for self-learning closed-loop optimal traffic signal control in the stochastic traffic environment((Abdulhai and Kattan 2003),(El-Tantawy and Abdulhai 2010)). RL has the added advantage of being able to perpetually learn and improve service over time. In RL, a traffic signal represents a control agent that interacts with the traffic environment in a closed-loop system to achieve the optimal mapping between the environment's traffic state and the corresponding optimal control action, offering an optimal control law (Sutton and Barto 1998). The mapping from states to actions is also referred to as the control policy. The agent iteratively receives a feedback reward for the actions taken and adjusts the policy until it converges to the optimal control policy. Applying RL to a transportation network of multiple signalized intersections is associated with some challenges. Agents typically react to changes in the environment at the individual level but the overall behavior of all agents may not be optimal. Each agent is faced with a moving-target learning problem in which the agent's optimal policy changes as the other agents' policies change over time(Bazzan 2009). Game Theory provides the tools to model the multi-agent systems as a multiplayer game and provide the rational strategy to each player in a game. Multi-Agent Reinforcement Learning (MARL) is an extension of RL to multiple agents in a stochastic game (SG) (i.e. multiple players in a stochastic environment). The decentralized traffic control problem is an excellent testbed for MARL due to the inherited dynamics and stochastic nature of the traffic system ((Bazzan 2009), (El-Tantawy and Abdulhai 2010)), which is our focus in this paper.

A novel Multi-Agent Reinforcement Learning for Integrated Network of Adaptive Traffic Signal Controllers (MARLIN-ATSC) system is designed and developed to directly address the challenges faced by MARL-based systems. First is the exponential growth in the state-action space with the increase in the number of agents. Second is that the majority of the MARL-based ATSC in the literature assume that agents learn independently in which case each agent acts individually in its local environment without explicit coordination1 with other agents in the environment. Although this simplifies the problem, it limits their usefulness in case of a network of agents. For example in over-saturated traffic conditions, queues could easily propagate from a downstream intersection (agent) and spills back to the upstream intersections (agents) in a network-wide cascading fashion; such cases require a network-wide multi-agent coordination as discussed earlier. Thus, flexible and computationally efficient approaches are becoming instrumental in controlling a network of agents; plausibly by employing heuristics and approximate approaches based on modifying the existing MARL techniques (Bazzan 2009).

To address these limitations, we present MARLIN-ATSC that offers the following features and characteristics: 1) decentralized design and operation- typically less expensive compared to the centralized system; 2) scalable to accommodate any network size; 3) robust – with no single point of failure, 4) model-free - does not require a model of the traffic system that is challenging to obtain; 5) self-learning - reduces human intervention in the operation phase after deployment (the most costly component of operating existing ATSCs) ; 6) coordinated – by implementing mode 2 (integrated mode), which coordinates the operation of intersections in two-dimensional road networks (e.g. grid network), a new feature that is unprecedented in ATSC state-of-the-art and practice. In addition, MARLIN-ATSC is tested on a large-scale simulated network of 59 intersections in downtown Toronto using the input data (e.g. traffic counts, signal timings, etc.) provided by City of Toronto.

RL FOR ADAPTIVE TRAFFIC SIGNAL CONTROL SYSTEMS

Thorpe (1997) applied the SARSA RL algorithm to a simulated traffic light control problem. The results showed that the SARSA RL algorithm outperformed the fixed timing plans by reducing the average vehicle waiting time by 29%. Wiering (2000) utilized model-based RL (with state transition models and state transition probabilities) to control traffic-light agents to minimize the waiting time of vehicles in a small grid network. The experimental results showed that RL systems outperform the non-adaptive systems by 22% in waiting time. Abdulhai et al. (2003) applied a model-free Q-Learning technique to a simple two-phase isolated traffic signal in a two-dimensional road network. Q-Learning for the isolated traffic-light controller outperformed the pre-timed control scheme for the variable traffic flow case by around 44%. Camponogara and Kraus Jr (2003) formulated the traffic signal control problem as a

[1] It is important to not confuse the coordination that is concerned about creating green wave along a certain corridor by adjusting the offset timing (defined as progression hereafter) with the mechanism between agents (signalized intersections) to coordinate their policies such that a certain objective is achieved for the entire traffic network (defined as coordination hereafter). In this paper, coordination refers to the latter one.

distributed stochastic game in which agents employed a distributed Q-Learning algorithm. When testing policy 3 (i.e. both agents run Q-learning), a 43% reduction in waiting time was achieved compared to policy 1(assigns the same probability to all actions available to an agent). Oliveira et al. (2006) extended the RL to multiple isolated traffic lights. They proposed an RL method called Reinforcement Learning with Context Detection (RL-CD), which can handle stochastic traffic patterns that occur due to the traffic dynamics. Richter et al. (2007) applied the Natural Actor Critic (NAC) algorithm to a 10 × 10 junction grid simulation network. NAC outperformed SAT (adaptive controller inspired by SCATS) by 20% reduction in average network travel time. Another example can be found in Arel et al. (2010) where RL is used to control the central intersection in a network of 5 intersections while the other 4 intersections use the longest-queue-first heuristic. Li et al. (2008) proposed an RL-based approach in which each agent considered the weighted sum of its local delay and its neighbors' delays as the outcome of its action. Salkham et al. (2008) proposed a similar algorithm to provide adaptive and efficient urban traffic control. Medina et. al. (2012) used Q-learning, and an approximate DP algorithm to control the traffic signals in which the learning agent considered its local state in addition to information about the congestion levels of neighbouring intersections.

In most of the previous studies the algorithms have been applied to simplified scenarios and under strong assumptions in terms of traffic behavior by considering simplified simulation environment ((Abdulhai, Pringle et al. 2003; Camponogara and Kraus Jr 2003; De Oliveira, Bazzan et al. 2006; Richter, Aberdeen et al. 2007; Arel, Liu et al. 2010)) , and/or assuming a hypothetical traffic flows ((Thorpe 1997; Wiering 2000; Abdulhai, Pringle et al. 2003; Camponogara and Kraus Jr 2003; De Oliveira, Bazzan et al. 2006; Richter, Aberdeen et al. 2007; Shoufeng et al. 2008; Arel, Liu et al. 2010)) which does not necessarily mimic the reality in traffic networks. Moreover, the previous studies considered independent learning agents and did not consider any explicit mechanism for coordination.

On the other hand, Kuyer et al. (2008) was found the only algorithm, to the best of authors' knowledge, that considered explicit coordination mechanism between the learning agents extending the work of Wiering (2000) using the Max-plus algorithm. Max-plus algorithm was used to estimate the optimal joint action by sending locally optimized messages among connected agents. However, Max-plus algorithm was found computationally demanding as it requires negotiations between the agents to coordinate their actions. Due to real-time nature of the ATSC problem, this forces the agents to report their current best action at any time even if the action found so were sub-optimal. Also, the use of a model-based RL approach adds unnecessary complexities compared to using a model free approach like Q-Learning.

In conclusion, there are two major challenges associated with applying RL (MARL) to ATSC problem; the need for coordination, and the cure of dimensionality as discussed below:

Need for Coordination:

The need for coordination stems from the fact that the effect of any agent's action on the environment depends also on the actions taken by the other agents. Hence, the agents' choices of actions must be mutually consistent in order to achieve their intended effect(Busoniu, Babuska et al. 2008) .

It can be concluded from the reviewed literature that the majority of the previous studies consider independent learning agents Oliveira et al. (De Oliveira, Bazzan et al. 2006), Camponogara and Kraus Jr (Camponogara and Kraus Jr 2003) , Bazzan (Bazzan 2005), Richter et al. (Richter, Aberdeen et al. 2007) , Arel et al. (Arel, Liu et al. 2010) , Wiering (Wiering 2000), Li et al. (Li, Zhao et al. 2008) , Salkham et al. (Salkham, Cunningham et al. 2008). Although Kuyer et al. (Kuyer, Whiteson et al. 2008) considered the two-level coordination, it suffers from the above mentioned limitations.

Curse of Dimensionality:

Although the existence of few coordination-based MARL methods (e.g., OAL(Claus and Boutilier 1998), and NSCP(Weinberg and Rosenschein 2004)), they suffer from the curse of dimensionality issue that arises because the state space is growing exponentially with the number of agents. Even in SG-based MARL approaches that are proven to optimally converge to the joint policy, each agent has to keep a set of tables whose size is exponential in the number of agents: |S1|×…×|SN|×|A1|×…×|AN | where Si and Ai represent the state and action spaces for agent i, respectively. In addition to the dimensionality issue, these methods require each agent to observe the state of the whole system which is infeasible in case of transportation networks.

In the next section we introduce a new algorithm that maintains a coordination mechanism between agents without compromising the dimensionality of the problem.

MULTI-AGENT REINFORCEMENT LEARNING FOR INTEGRATED NETWORK OF ADAPTIVE TRAFFIC SIGNAL CONTROLLERS (MARLIN-ATSC) PLATFORM

The MARLIN-ATSC platform is illustrated in Figure 1. The platform consists of two main layers; the first layer is an input configuration layer that is responsible for configuring and providing the necessary input to the second layer.

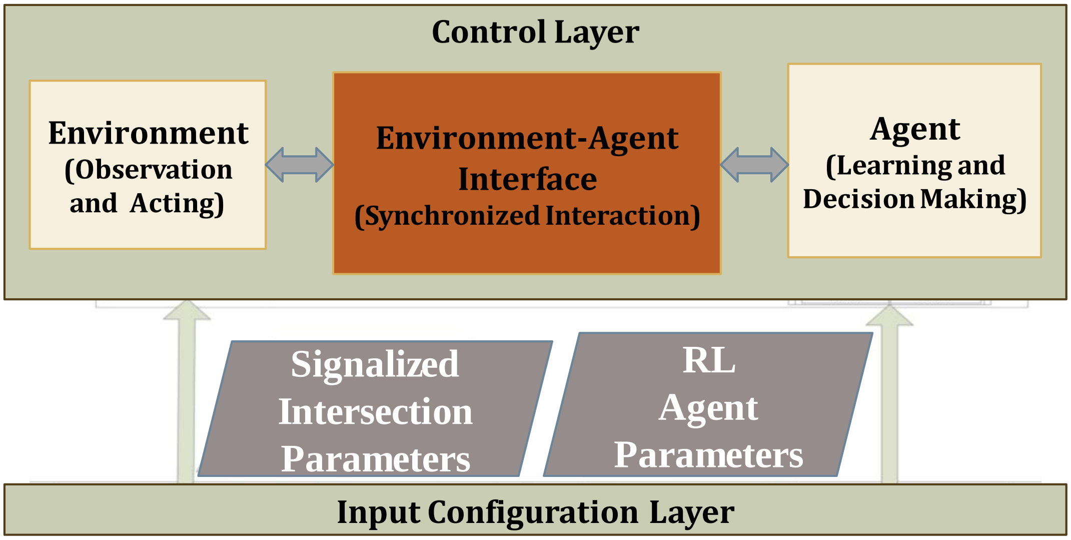

Control Layer
Environment (Observation and Acting)
Environment-Agent Interface (Synchronized Interaction)
Agent (Learning and Decision Making)
Signalized Intersection Parameters
RL Agent Parameters
Input Configuration Layer

Figure 1- MARLIN -ATSC Platform

The configuration layer has two main roles; 1) configures the simulation-based learning environment (model) such that the simulated environment closely matches the real-world environment, 2) configures the RL-design parameters.

The second layer is a control layer that includes three interacting components (as shown in Figure 1):

Agent

The Agent component implements the control algorithm; The agent is the learner and the decision-maker that interacts with the environment by first receiving the system's state and fthe reward and then selecting an action accordingly. A Generic agent model is developed using Java Programming Language such that different levels of coordination, learning methods, state representations, phasing sequence, reward definition, and action selection strategies can be tested for any control task. In MARLIN-ATSC, agents can implement one of the following two control modes:

- **Independent Mode:** In this mode, each controller has an RL agent working independently of other agents using Multi-Agent Reinforcement Learning for Independent controllers (MARL-I) in which each agent implements a Q-Learning algorithm (Watkins and Dayan 1992).
- **Integrated Mode:** In this mode, each controller coordinates the signal control actions with the neighbouring controllers by implementing MARLIN learning algorithm.

In typical traffic networks, signalized intersections (agents) are physically connected through the roadway network which represents a set of neighbouring agents. Although in traffic networks, the agent is incapable of observing the conditions of the entire network, it is possible to observe the conditions of the neighbouring agents. In MARLIN, each agent plays a game with all its adjacent intersections in its neighbourhood. The agent has a number of learning modules; each corresponds to one game. The state-space and the action-space are distributed such that the agent learns the joint policy with one of the neighbours.

The following are the steps for the learning approach designed in MARLIN that is formally described in a pseudo code in Algorithm 1:

- If there are $|\mathrm{NB_i}|$ neighbours for agent i, there are $|\mathrm{NB_i}|$ partial state and action spaces for agent i. Each partial state space and action space consists of agent i and one of the neighbours $\mathrm{NB_i[j]}$, s.t. $\mathrm{j} \in \mathrm{NB_i}$ ($\mathrm{S_i}, \mathrm{S_{NB_i[j]}}, \mathrm{A_i}, \mathrm{A_{NB_i[j]}}$).
- Each agent i is builds a model that estimates the policy for each of its neighbours and represented by a matrix $M_{i,NB_i[j]}$, s.t. $\mathrm{j} \in \mathrm{NB_i}$ where the rows are the joint states $\mathrm{S_i} \times \mathrm{S_{NB_i[j]}}$ and the columns are the neighbour’s actions $\mathrm{A_{NB_i[j]}}$. Each cell $M_{i,NB_i[j]}([\mathrm{s_i}, \mathrm{s_{NB_i[j]}}], \mathrm{a_{NB_i[j]}})$ represents the probability that agent $NB_i[j]$ takes action $\mathrm{a_{NB_i[j]}}$ at the joint state $[\mathrm{s_i}, \mathrm{s_{NB_i[j]}}]$ using the count of visits to the state-action $v([s_i^k, s_{NB_i[j]}^k], a_{NB_i[j]}^k)$ for the state-action pair$([s_i^k, s_{NB_i[j]}^k], a_{NB_i[j]}^k)$ (Eqn. 1).
- Each agent i learns the optimal joint policy for agents i and $\mathrm{NB_i[j]}\ \forall\ \mathrm{j} \in \{1, \ldots, |\mathrm{NB_i}|\}$ by updating the Q-values that are represented by a matix of $|\mathrm{S_i} \times \mathrm{S_{NB_i[j]}}|$ rows and $|\mathrm{A_i} \times \mathrm{A_{NB_i[j]}}|$ columns where each cell $\mathrm{Q_{i,NB_i[j]}}([\mathrm{s_i}, \mathrm{s_{NB_i[j]}}], [\mathrm{a_i}, \mathrm{a_{NB_i[j]}}])$ represents the Q-value for a state-action pair in the partial spaces corresponding to the pair of connected agents $(\mathrm{i}, \mathrm{NB_i[j]})$;
- Each agent updates Q-values $\mathrm{Q_{i,NB_i[j]}}([\mathrm{s_i}, \mathrm{s_{NB_i[j]}}], [\mathrm{a_i}, \mathrm{a_{NB_i[j]}}])$ using the value of the best-response action taken in the next state. The best-response value ($\mathrm{br_i^k}$) is the maximum expected Q-value at the next state which is calculated using the models for other agents (Eqn. 2).
- Each agent decides its action without direct interaction with the neighbors. Instead, the agent uses the estimated models for the other agents and act accordingly. Agent i chooses the next action using a simple heuristic decision procedure which bias action selection toward actions that has the maximum expected Q-value over its neighbours $\mathrm{NB_i}$. The likelihood of Q-values is evaluated using the models of the other agents, $\mathrm{M_{i,NB_i[j]}}$, estimated in the learning process (Eqn. 4).

Algorithm 1: MARLIN Learning

```
Initialization at time $k = 0$:
For each agent $i$, $i \epsilon \{1,2,\ldots,N\}$:
For each neighbor $j \epsilon \{1,2,\ldots\ |NB_i|\}$
    Initialize $s_i^0, a_i^0, a_{NB_i[j]}^0$
    $M_{i,NB_i[j]}^{\ 0}([s_i, s_{NB_i[j]}], a_{NB_i[j]}) = 1/|A_{NB_i[j]}|$,   $Q_{i,NB_i[j]}^k([s_i, s_{NB_i[j]}], [a_i, a_{NB_i[j]}]) = 0$
End for
End for
For each time step $k$ , do:
  For each agent $i$, $i \epsilon \{1,2,\ldots,N\}$, do:
  For each neighbor $NB_i[j]$, $j \epsilon \{1,2,\ldots\ |NB_i|\}$ do:

    a. Observe $a_{NB_i[j]}^k$, $s_i^{k+1}$ $s_{NB_i[j]}^{k+1}$, and $r_i^k$
    b. Update $M_{i,NB_i[j]}$
       $M_{i,NB_i[j]}([s_i^k, s_{NB_i[j]}^k], a_{NB_i[j]}^k) = \frac{v([s_i^k, s_{NB_i[j]}^k], a_{NB_i[j]}^k)}{\sum_{a_{NB_i[j]} \in A_{NB_i[j]}} v([s_i^k, s_{NB_i[j]}^k], a_{NB_i[j]})}$   (1)
    c. Choose the maximum expected Q-value at state $s_{NB_i[j]}^{k+1}$
$br_i^k = max_{a_i \in A_i} \left[\sum_{a_{NB_i[j]} \in A_{NB_i[j]}} Q_{i,NB_i[j]}^k([s_i^k, s_{NB_i[j]}^k], [a_i, a_{NB_i[j]}] \cdot M_{i,NB_i[j]}([s_i^k, s_{NB_i[j]}^k], a_{NB_i[j]})\right]$   (2)
    d. Update $Q_{i,NB_i[j]}$
$Q_{i,NB_i[j]}^k([s_i^k, s_{NB_i[j]}^k], [a_i^k, a_{NB_i[j]}^k]) = (1-\alpha) Q_{i,NB_i[j]}^{k-1}([s_i^k, s_{NB_i[j]}^k], [a_i^k, a_{NB_i[j]}^k]) + \alpha[r_i^k + \gamma\ br_i^k]$   (3)
Decide
$a_i^{k+1} = \underset{a_i \in A_i}{\mathrm{argmax}} \left[\sum_{j \epsilon \{1,2,\ldots,|NB_i|\}} \sum_{a_{NB_i[j]} \in A_{NB_i[j]}} Q_{i,NB_i[j]}^k([s_i^k, s_{NB_i[j]}^k], [a_i, a_{NB_i[j]}]) \cdot M_{i,NB_i[j]}([s_i^k, s_{NB_i[j]}^k], a_{NB_i[j]})\right]$   (4)
End For
End For
End For
```

Simulation Environment

The simulation environment component models the traffic environment. In this paper, Paramics, a microscopic traffic simulator, is used to model traffic environment (Quadstone Paramics 2012). Paramics models stochastic vehicle flow by employing speed regulations, car-following, gap acceptance, and overtaking rules. Paramics provides three methods of traffic assignment that could be employed at different levels: “all-or-nothing” assignment, stochastic assignment, and dynamic feedback assignment. In this application a dynamic stochastic traffic assignment was used where: 1) a random noise was added to the travel cost to account for the heterogeneity among drivers’ perception of travel cost, 2) a dynamic feedback interval was used to update route travel times for familiar drivers in the simulation. Paramics API functions were used to construct the state, execute the action, and calculate the reward for each signalized intersection.

Some of the main challenges in deigning any RL system are the design of the state definition, action definition, and reward definition. In (El-Tantawy and Abdulhai 2011), a comprehensive investigation of these key issues in RL-based signal control for isolated intersections is conducted. The state definition, action definition, and reward definition recommended in (El-Tantawy and Abdulhai 2011) and (El-Tantawy and Abdulhai 2012)are adopted in this paper as follows (for more details on the definitions, please refer to (El-Tantawy and Abdulhai 2011));

State Definition : Queue length

The agent’s state is represented by a vector of 2+P components, where P is the number of phases. The first two components are: 1) index of the current green phase, and 2) elapsed time of the current phase. The remaining P components are the maximum queue lengths associated with each phase.

Action Definition: Variable Phasing Sequence

The agent is designed to account for variable phasing sequence in which the control action is either to extend the current phase or to switch to any other phase according to the fluctuations in traffic, possibly skipping unnecessary phases. Therefore, this algorithm is an acyclic timing scheme with variable phasing sequence in which not only the cycle length is variable but also the phasing sequence is not predetermined. Hence, the action is the phase that should be in effect next.

Reward Definition: the reduction in the Total Cumulative Delay

The immediate reward for certain agent is defined as the reduction (saving) in the total cumulative delay associated with that agent, i.e., the difference between the total cumulative delays of two successive decision points. The total cumulative delay at time k is the summation of the cumulative delay, up to time k, of all the vehicles that are currently in the intersections’ upstreams. If the reward has a positive value, this means that the delay is reduced by this value after executing the selected action. However, a negative reward value indicates that the action results in an increase in the total cumulative delay.

Interface

Interface component manages the interactions between the agent and the simulation environment by exchanging the state, reward, and action. The interaction between the agent and the environment is associated with the following design elements:

- A synchronized interaction between the agent and the environment was designed to ensure that the simulation environment is held while the agent is performing the learning and the decision making processes and finally produces the action that should be executed by simulation environment. At the same time, the agent should be on hold until the action is executed in the environment and the resultant state and the reward are measured.

- The system was designed such that the interaction frequency is variable for each agent. The interaction occurs each specified time interval (1 sec in this research) as long as the current green for a signalized intersection that is associated with an agent i exceeded the minimum green time. Otherwise, the interaction starts after the minimum green.

The agent was designed to learn off-line through a simulation environment (such as the micro-simulation model employed in the experiments) before field implementation. After convergence to the optimal policy, the agent can either be deployed in the field – by mapping the measured state of the system to optimal control actions directly using the learned policy; or it can continue learning in the field by starting from the learned policy.

EXPERIMENTAL RESULTS

Testbed Network:

MARLIN-ATSC is tested on a simulated network of the Lower Downtown Toronto network. The lower downtown of Toronto is the core of the City of Toronto. The lower downtown of Toronto in this study is bounded to the South by the Queens Quay corridor, to the West by the Bathurst St, to the East by the Don Valley Parkway (DVP) and to the North by Front St. Toronto is the oldest, densest, most diverse area in the region and its downtown core contains one of the highest concentrations of economic activity in the country. This paper demonstrates large-scale application of MARLIN-ATSC on a simulated replica of the lower downtown core. A base-case simulation model for the lower downtown core was originally developed using Paramics, a microscopic traffic simulator, in the ITS Centre and Testbed at the U of T for year 2006. In this application, the model is further refined to reflect the signal timing sheets provided by the City of Toronto2. The analysis period considered in this application is the AM peak hour, which has around 25,000 vehicular trips.

Benchmarks

It is typically difficult to find a benchmark for large-scale traffic signal control problems given that the operational details of most traffic control systems are not easily available due to obvious commercial reasons. The performance of MARLIN-ATSC approach is compared to the Base Case (BC) scenario in which traffic signals, as defined and operated by the City of Toronto, are mix of fixed-time control, semi-actuated control, and SCOOT control as shown in Figure 2. It is worth noting that due to the limited technical details about the operation of SCOOT, it is approximated in this thesis as an enhanced fully-actuated control in which loop detectors are placed on all approaches and the extension times are conducted second-by-second.

[2] Contact person is Rajnath Bissessar, City of Toronto – Transportation Services, Manager of the UTCS

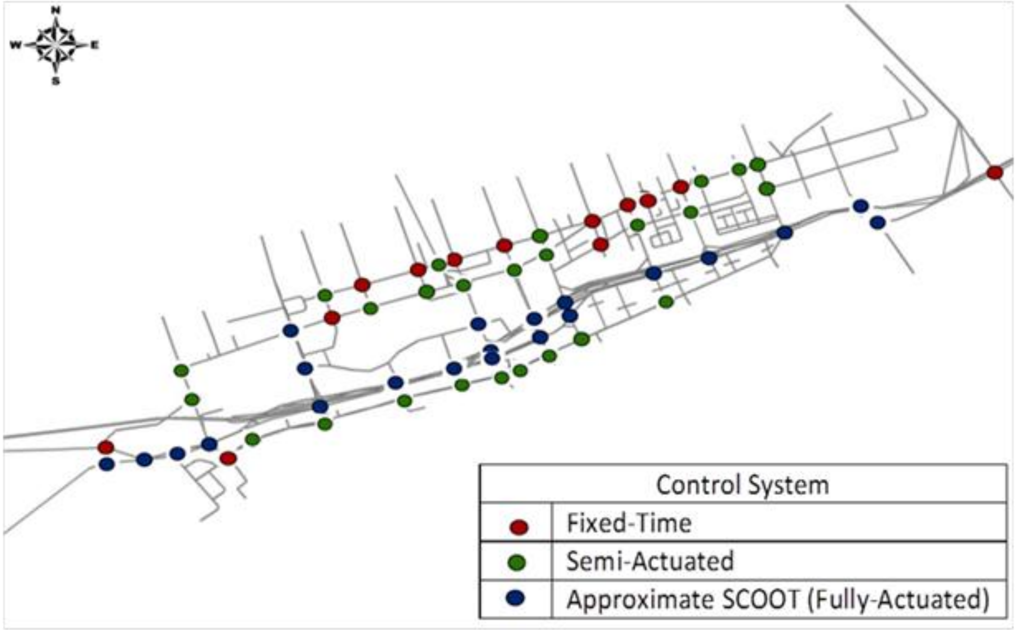


Figure 2- Currently Implemented Signal Control Systems

Experimental Setup

Two types of coordination are examined: 1) arterial (corridor)-specific coordination, and 2) network-wide coordination.
In the arterial control case MARL-I and MARLIN are tested on the Lakeshore Arterial, an arterial where considerable variations in traffic are anticipated because it runs parallel to the Gardner Expressway. MARLIN is used to test two cases: 1) arterial-specific coordination with agents coordinated only with their neighbours on the specific arterial (called MARLIN-A), and 2) network-wide agents coordinated with all neighbours (called MARLIN-C (Network Control)).
In the network control experiments three scenarios are investigated; uniform demand profile, variable demand profile, and unfamiliar drivers.
- Normal: in this scenario uniform demand profile is used. The percentage of familiar drivers is 60% and the feedback interval is 2 min;
- Variable: in this scenario variable demand profile is used (as shown in Figure 3) to replicate any expected highly variable traffic conditions;
- Unfamiliar Drivers: in this scenario 30% familiar drivers and a 4 min feedback interval are considered to investigate the case when drivers are less aware of other routing options.
In each of the three above scenarios, results are reported for BC control systems (existing conditions), MARL-I (represents MARLIN-ATSC Independent Mode with no communication between agents), and MARLIN (represents MARLIN-ATSC Integrated Mode with coordination between agents).

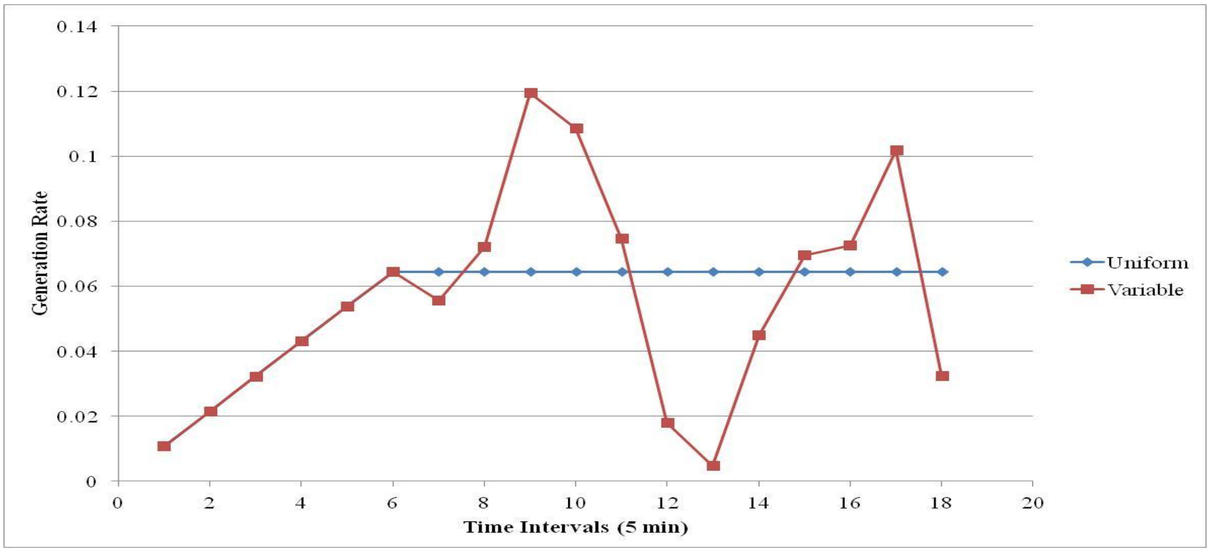


Figure 3- Demand Profiles

Results and Discussion

Arterial Control Experiments

In these experiments, Lake Shore Boulevard (LS Blvd) is selected as the testbed arterial. LS is a major East/West arterial in Toronto’s waterfront. It is an alternative route for the Gardiner Expressway, and therefore it is a candidate route for many trips destined to downtown Toronto. Consequently it is one of the busiest and key corridors in the network, where substantial variations in traffic flows are anticipated. The modelled section of the arterial is around 7 km long and contains 20 intersections.

MARL-I and MARLIN-A are only applied to the signalised intersections along the arterial, while the remaining intersections in the network are operated with the same control as the BC. It is also important to note that MARLIN-C is applied over the entire network in order to study the effect of two-dimensional (network) coordination vs one-dimensional (arterial) coordination.

Figure 4 shows the average delay per intersection for major and minor streets for the 20 intersections along the LS corridor. It is shown in Figure 4 that MARL-I, MARLIN-A and MARLIN-C outperform the BC by around 10%, 29%, 55% in terms of average delay/intersections along the major street approaches (i.e. EB and WB) and by around 49%, 63%, 69% in terms of average delay/intersections along the minor street approaches (i.e. NB and SB). This indicates that the system not only minimises the delay along major street approaches, but also along those of the minor streets.

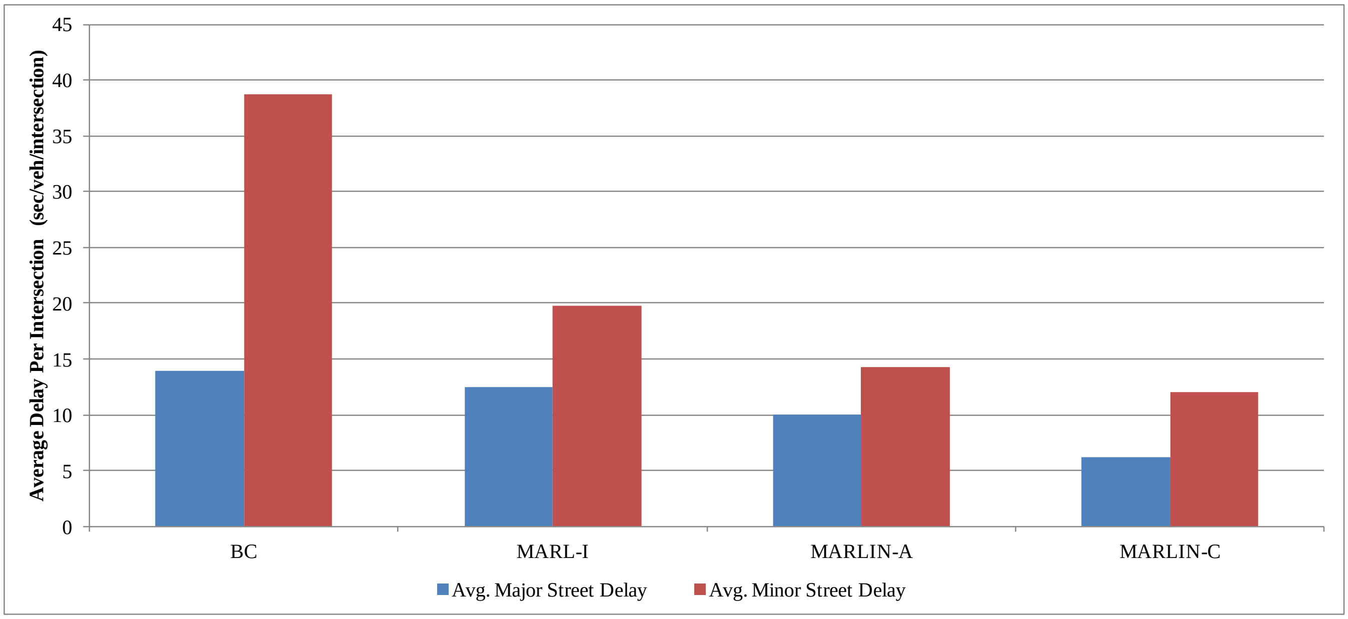


Figure 4- Comparison of Major and Minor Street Average Delay Per Intersection

The route travel time variability in the morning rush hour is illustrated in Figure 5 and Figure 6 have been produced. MARLIN-C outperforms the other systems with respect to travel time variability, which is expected because in the MARLIN-C all the signalised intersections in the network are operating using MARLIN. MARLIN-A on the other hand still exhibits less travel time variability compared to the BC, but not to the level achieved by MARLIN-C. Overall MARL-I and MARLIN systems exhibit less travel time variability, which is indicative of travel time reliability. Reliable travel times reflects a robust system and less cost due to congestion.

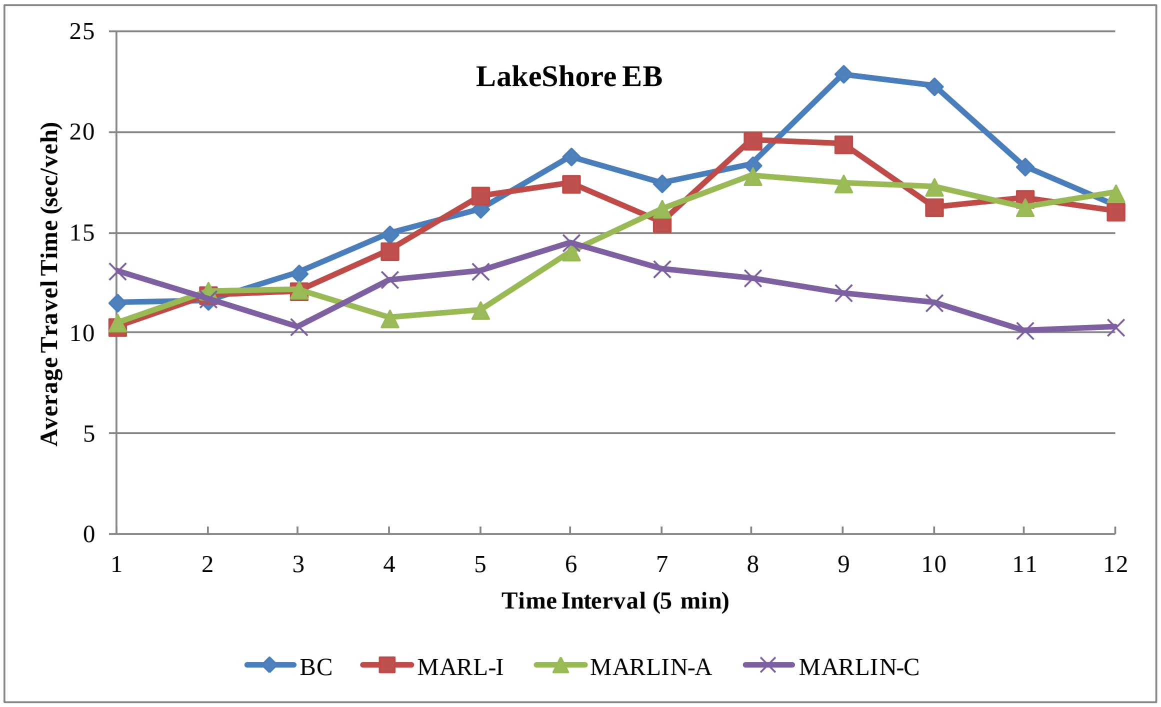


Figure 5- Lake Shore EB Travel Time Variations for: Base Case, MARL-I, MARLIN

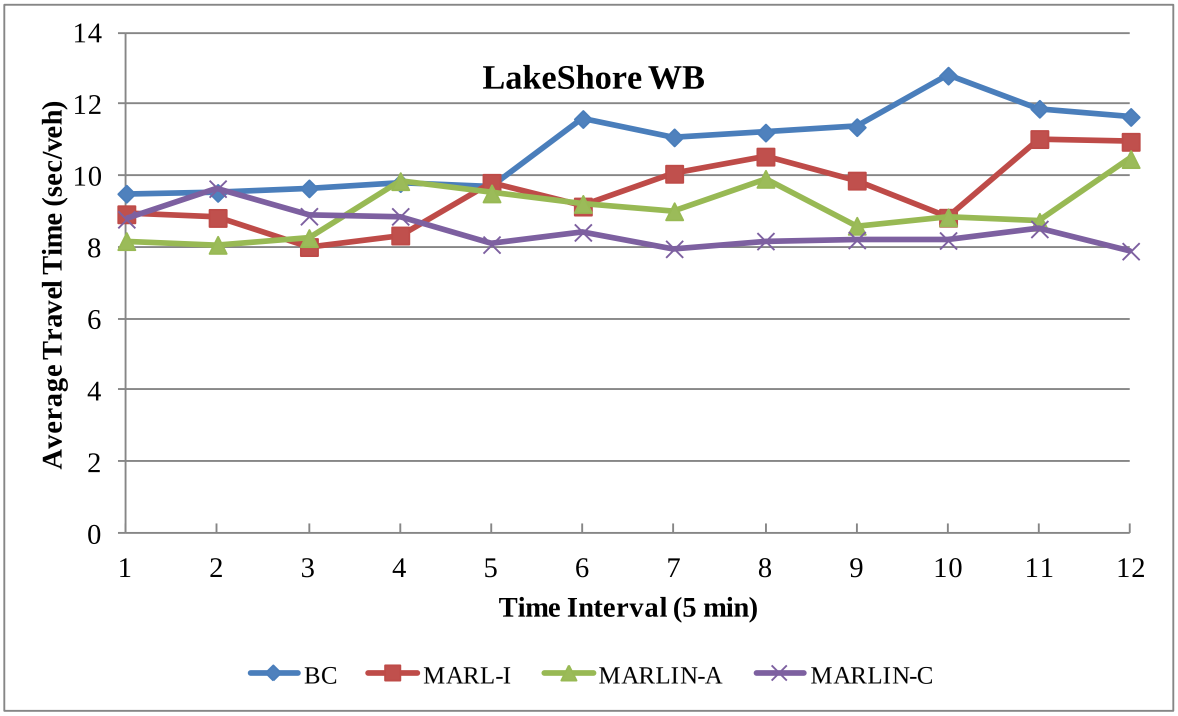


Figure 6- Lake Shore WB Travel Time Variations for: Base Case, MARL-I, MARLIN

Network Control Experiments

- **Normal Scenario**

The performance of each control system is evaluated based on the following measures of effectiveness:

- Average Delay Per Vehicle (sec/veh)
- Average Max. Queue Length Per Intersection (veh)
- Average Standard Deviation of Queue Lengths Across Approaches (veh)
- Number of Completed trips
- Average CO2 emissions factors (gm/km)
- Average Travel Time For Selected Routes (min)

Table 1 compares the performance of the BC against the MARLIN-ATSC system with and without communication among agents, i.e, MARLIN and MARL-I, respectively.

The analysis of the results shown in Table 1 leads to the following findings:

The two MARLIN-ATSC algorithms result in lower average delay, higher throughput, shorter queue length, and stop time compared to those from the base case. The most notable improvements are the average delay (38% MARLIN vs BC), standard deviation of average queue length (31% MARLIN vs BC), CO2 emission factors (30% MARLIN vs BC).

These substantial improvements are not only due to the intelligence of the RL algorithm, but also as a result of the coordination mechanism between the agents to reach a network-wide set of actions that minimize the long term delay. This coordination results in the so-called “metering” effect from the upstream intersection to the downstream intersection while accounting for the queues and delays at the downstream intersection. In fact, the tangible savings in the standard deviation in the queue length is interesting because this means balanced queue among all intersection approaches.

Table 1- Network-Wide MOE in the Normal Scenario

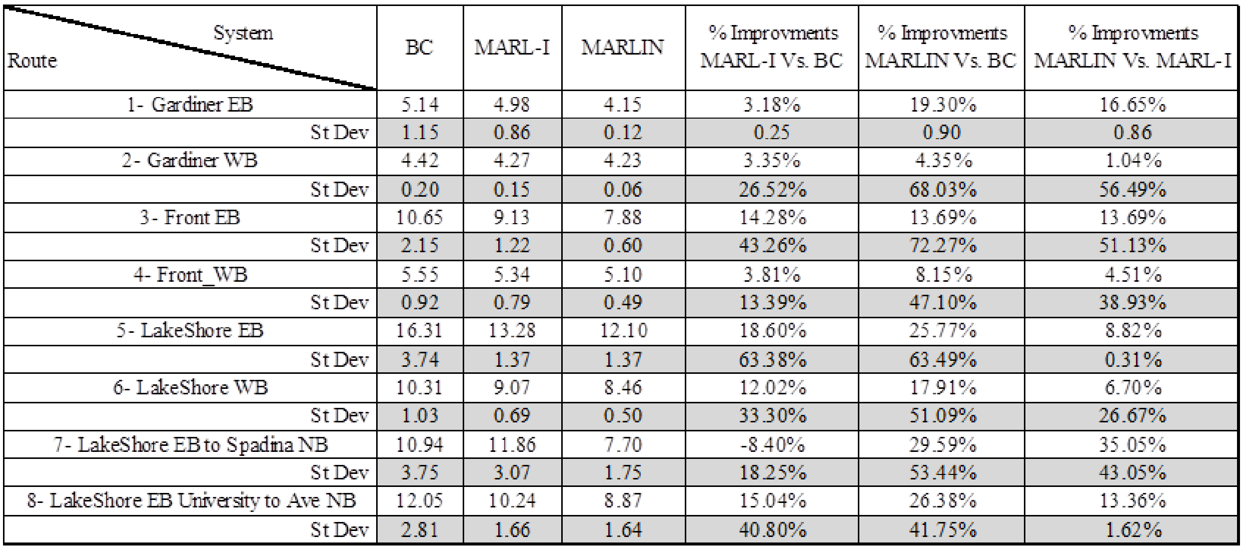
System MOE	BC	MARL-I	MARLIN	% Improvments MARL-I Vs. BC	% Improvments MARLIN Vs. BC	% Improvments MARLIN Vs. MARL-I
Average Intersection Delay (sec/veh)	35.27	25.72	22.02	27.06%	37.57%	14.41%
Throughput (veh)	23084	23732	24482	2.81%	6.06%	3.16%
Avg Queue Length (veh)	8.66	6.60	5.88	23.77%	32.07%	10.88%
Std. Avg. Queue Length (veh)	2.12	1.62	1.47	23.37%	30.74%	9.61%
Avg. Link Delay (sec)	9.45	8.50	5.04	10.07%	46.73%	40.76%
Avg. Link Stop Time (sec)	2.74	2.57	2.02	5.95%	26.06%	21.38%
Avg. Link Travel Time (sec)	16.81	15.81	12.32	5.97%	26.70%	22.05%
CO2 Emission Factor (gm/km)	587.28	421.34	412.21	28.26%	29.81%	2.17%

- MARL-I outperforms the BC in all the MOEs, most notably are the average intersection delay (27%) and the C02 emission factor (28%). However, comparing MARLIN to MARL-I, it is found that the latter experience relatively higher delays because in MARL-I the actions are based only on locally collected data and thereby results in more vehicles retained in the network at the end of the simulation (6% throughput improvement in MARLIN vs 2.8% throughput improvement in MARL-I).

It is important to study the effect of various control systems on the travel time and travel time variability for selected key routes in the lower downtown core of Toronto. Route travel times and standard deviation in travel time for the BC, MARL-I, and MARLIN scenarios for eight key routes in the network are presented in Table 2. To further study the route travel times within the simulation hour, the travel time for the eight routes are plotted in . The analysis of Table 2 and Figure 7 leads to the following conclusions:

- It is clear that MARLIN outperforms MARL-I and BC in all routes. The % improvements range from 4 % in route 2 to 30% in route 7. MARL-I outperforms BC in almost all cases; the % improvements ranges from 3% in route 1 to 15% in route 5 with the exception of route 7, the BC scenario performs better than MARL-I.

Table 2 - Route Travel Times for BC, MARL-I, and MARLIN



System Route	BC	MARL-I	MARLIN	% Improvments MARL-I Vs. BC	% Improvments MARLIN Vs. BC	% Improvments MARLIN Vs. MARL-I
1- Gardiner EB	5.14	4.98	4.15	3.18%	19.30%	16.65%
St Dev	1.15	0.86	0.12	0.25	0.90	0.86
2- Gardiner WB	4.42	4.27	4.23	3.35%	4.35%	1.04%
St Dev	0.20	0.15	0.06	26.52%	68.03%	56.49%
3- Front EB	10.65	9.13	7.88	14.28%	13.69%	13.69%
St Dev	2.15	1.22	0.60	43.26%	72.27%	51.13%
4- Front_WB	5.55	5.34	5.10	3.81%	8.15%	4.51%
St Dev	0.92	0.79	0.49	13.39%	47.10%	38.93%
5- LakeShore EB	16.31	13.28	12.10	18.60%	25.77%	8.82%
St Dev	3.74	1.37	1.37	63.38%	63.49%	0.31%
6- LakeShore WB	10.31	9.07	8.46	12.02%	17.91%	6.70%
St Dev	1.03	0.69	0.50	33.30%	51.09%	26.67%
7- LakeShore EB to Spadina NB	10.94	11.86	7.70	-8.40%	29.59%	35.05%
St Dev	3.75	3.07	1.75	18.25%	53.44%	43.05%
8- LakeShore EB University to Ave NB	12.05	10.24	8.87	15.04%	26.38%	13.36%
St Dev	2.81	1.66	1.64	40.80%	41.75%	1.62%

- It is interesting to find that the Gardiner Expressway EB traffic (inbound) travel time improves by 19% in the MARLIN scenario. Alleviating the congestion on Spadina St and York St off-ramps contributes the most to these savings. This clearly shows the effect of the downstream capacity on the freeway performance. For the Gardiner WB direction traffic was not as congested as the EB but MARLIN still attains 4% improvement in average route travel times.
- The most congested routes appear to be routes 7 and 8; through which traffic originated at the west end of the study area and destined in the downtown core (Spadina St and University Ave). MARLIN achieves 30% and 26% improvements in route 7 and 8, respectively, which reflects the superior effect of the two-dimensional coordination between agents.
- From observing the temporal distribution of route travel time across the simulation hour, it is generally found that MARLIN is stable and exhibits less variation compared to the BC and MARL-I scenarios. While the BC scenario exhibits the highest variability in travel time (as shown in the standard deviation values in Table 2), MARL-I still shows some variations, most notably in the most two congested routes (route 7, 8). MARLIN shows stable route travel times in all routes.

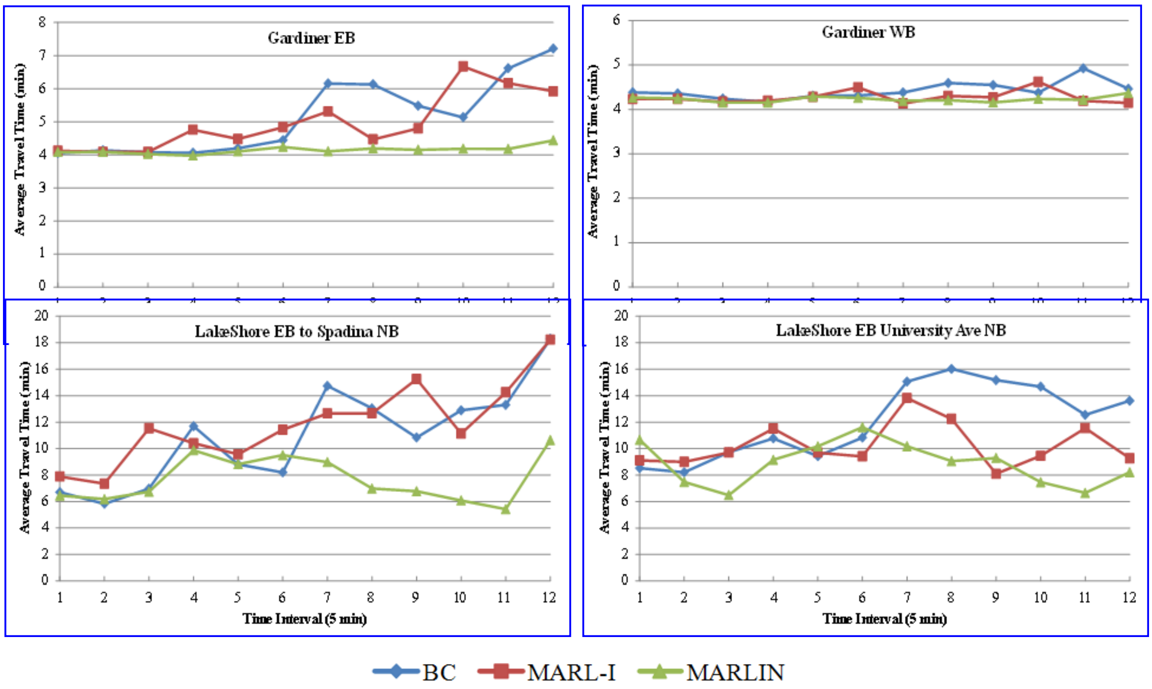


Figure 7- Average Route Travel Time for Selected Routes

- ## Variable Profile Scenario

Due to the large-scale size of the network and for the sake of conciseness, only average delay, throughput, and average and Std of queue length are presented for the variable profile test case in Table 3.

Table 3 - Network-Wide MOE for Variable Profile Scenario

System MOE	BC	MARL-I	MARLIN	% Improvments MARL-I Vs. BC	% Improvments MARLIN Vs. BC	% Improvments MARLIN Vs. MARL-I
Average Intersection Delay (sec/veh)	46.26	29.41	24.51	36.41%	47.01%	16.67%
Throughput (veh)	18452	21600	24518	17.06%	32.87%	13.51%
Avg Queue Length (m)	14.20	9.97	6.26	29.84%	55.91%	37.16%
Std. Avg. Queue Length (m)	3.19	2.22	1.55	30.45%	51.45%	30.20%

The analysis of the results shown in Table 3 leads to the following findings:

- Compared to the uniform profile case, considerable deterioration in the performance of BC is observed (24% increase in average delay, 25% decrease in throughput, 39% increase in average queue length, and 34% increase in Std of average queue lengths);
- The two MARLIN-ATSC algorithms considerably outperform the BC in all the MOEs. The most notable improvements are in average delay (47% MARLIN vs BC), throughput (33% MARLIN vs BC), average queue length (56% MARLIN vs BC), and Std of average queue length (51% MARLIN vs BC);
- MARL-I outperforms the BC in all the MOEs, most notably the average intersection delay (36%). Additionally, comparing MARLIN to MARL-I it is found that the latter experience higher delays because in MARL-I the actions are based on local traffic states with no coordination between other agents, therefore MARL-I results in more vehicles retained in the network at the end of the simulation (14% throughout improvement in MARLIN vs MARL-I). It is found that MARL-I exhibits higher average queue length values than MARLIN (37%) and less balancing for the queue lengths across the approaches (30%);

- This deterioration in the performance of the BC in the variable profile case is due to the fact that high variations in the traffic stream are causing long queues to form and blockages at some intersections. These cases result in substantial performance degradation in a cascading fashion (especially in this grid-like urban network) as the traffic control system in the BC could not respond to the change in traffic patterns in an adequately adaptive manner. This observation is a clear indication that the BC is not robust, as its performance degrades considerably with slight changes in the traffic conditions. The key to the good performance achieved by MARLIN-ATSC, on the other hand, is that agents could promptly adapt the signal timings, by skipping unnecessary phases and by providing optimised timing per phase to any variations in the traffic stream.

- **Drivers Unfamiliarity**

The main focus in this section is to investigate the effect of Travellers Information Systems (TIS) and the unfamiliarity of drivers on the performance of different control systems as shown in Table 4.

Table 4 - Network-Wide MOE for the Driver Familiarity Scenario

System MOE	BC	MARL-I	MARLIN	% Improvments MARL-I Vs. BC	% Improvments MARLIN Vs. BC	% Improvments MARLIN Vs. MARL-I
Average Intersection Delay (sec/veh)	42.37	39.20	23.47	7.48%	44.60%	40.12%
Throughput (veh)	11601	13739	22664	18.43%	95.36%	64.96%
Avg Queue Length (m)	13.13	10.43	6.82	20.53%	48.06%	34.64%
Std. Avg. Queue Length (m)	2.96	2.43	1.81	17.93%	38.71%	25.32%

The analysis of the results shown in Table 4 leads to the following findings:

- Compared to the uniform profile case, a substantial deterioration in the performance of BC is observed using the low percentage of familiar drivers (16% increase in average delay, 34% increase in average queue length, and 28% increase in Std of queue lengths);
- The two MARLIN-ATSC algorithms considerably outperform the BC in all the MOEs. MARL-I outperforms the BC in all the MOEs, most notably the average queue length (20%). However, comparing MARLIN to MARL-I it is found that the latter experiences higher delays and congestion (64% throughout improvement in MARLIN vs MARL-I);
- This deterioration in the performance of the BC in the low driver familiarity is attributed to the fact that a fewer number of travellers are receiving updated travel times (and costs in general) to their destination, therefore they are less likely to switch routes in the case where their route travel time to destination increases. This effect creates oversaturation conditions in some areas of the network, where MARL-I agents fail to converge to the optimal policy.

CONCLUSIONS AND FUTURE WORK

In this paper, previous studies that tackled the ATSC problem using MARL approaches were reviewed and the gaps in literature were highlighted. The major challenges for using MARL-based signal control system was the need for coordination and the curse of dimensionality. In order to attain the compromise of achieving coordination-based decentralized adaptive real-time control without suffering from the curse of dimensionality challenge that is associated with MARL techniques, a Multi-Agent Reinforcement Learning for Integrated Network of Adaptive traffic Signal Controllers (MARLIN-ATSC) system was presented. In this system, each agent plays a game with its immediate neighbours. Each agent learns and converges to the best response policy to all neighbours' policies. The paper demonstrated the essence of MARLIN-ATSC on a large-scale urban network of 59 intersections in Downtown Toronto. A MARLIN-ATSC large-scale application was conducted to examine two cases: 1) corridor-specific agent coordination, and 2) network-wide agent coordination. In the corridor experiments, the LS Blvd was selected as the testbed arterial as it is one of the most important corridors in downtown Toronto. In the network-wide coordination experiments, three scenarios were investigated: uniform demand profile, variable demand profile, and unfamiliar drivers (using a low percentage of familiar drivers). In each of the three scenarios above, the results were reported for BC control systems from the field (simulated using signal timing sheets provided by the City of Toronto, MARL-I (represents MARLIN-ATSC Independent Mode with no communication between agents), and MARLIN (represents MARLIN-ATSC Integrated Mode with coordination between agents). In the arterial experiment, MARLIN generally outperformed the BC in terms of average travel time, and Std of travel time. In the normal scenario of network-wide coordination experiments, MARL-I outperformed the BC in all the MOEs. However, comparing MARLIN to MARL-I it was found that the latter experienced relatively higher delays. Compared to the uniform profile case, substantial deterioration in the performance of BC in the variable profile scenario was observed. Although MARL-I also outperformed the BC in all the MOEs in the variable profile scenario, comparing MARLIN to MARL-I it is found that the latter produced higher delays because in MARL-I the actions are based on locally sensed states with no coordination with other agents. Thereby MARL-I resulted in more vehicles being retained in the network at the end of the simulation. In the unfamiliar drivers scenario a large deterioration in the performance of BC was observed. The two MARLIN-ATSC algorithms substantially outperformed the BC in all the MOEs. However, comparing MARLIN to MARL-I it was found that the latter produced higher delays and congestion. It was found that the deterioration in the performance of the BC in the low drivers familiarity case is attributed to the fact the unfamiliar drivers are less likely to switch routes in cases where their routes to the destination became congested, which created oversaturation conditions in some areas of the network and hence also deteriorated the performance of MARL-I.

To quantify the benefits of MARLIN-ATSC relative to existing ATSC systems such as SCOOT, without approximation, the following approaches could be used in the future: 1) compare the simulation-based measures of MARLIN with the real life observations and benefits of SCOOT for SCOOT- controlled intersections; 2) use hardware-in-the-loop simulation (HILS) methodologies to replicate the logic of SCOOT within the simulation software such as Paramics.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support of Connaught and OGS Scholarships from University of Toronto. The authors also would like to acknowledge the support of City of Toronto staff for the data they provided to this research.

REFERENCES

Abdulhai, B. and L. Kattan (2003). "Reinforcement learning: Introduction to theory and potential for transport applications." Canadian Journal of Civil Engineering 30(6): 981-991.

Abdulhai, B., R. Pringle, et al. (2003). "Reinforcement Learning for True Adaptive Traffic Signal Control." Journal of Transportation Engineering 129(3): 278-285.

Arel, I., C. Liu, et al. (2010). "Reinforcement learning-based multi-agent system for network traffic signal control." IET Intelligent Transport Systems 4(2): 128–135.

Basar, T. and G. J. Olsder (1999). Dynamic Noncooperative Game Theory. London, U.K, Classics in Applied Mathematics.

Bazzan, A. L. C. (2005). "A distributed approach for coordination of traffic signal agents." Autonomous Agents and Multi-Agent Systems 10(1): 131-164.

Bazzan, A. L. C. (2009). "Opportunities for multiagent systems and multiagent reinforcement learning in traffic control." Autonomous Agents and Multi-Agent Systems 3(18): 342–375.

Busoniu, L., R. Babuska, et al. (2008). "A comprehensive survey of multiagent reinforcement learning." IEEE Transactions on Systems, Man and Cybernetics 38(2): 156–172.

Camponogara, E. and W. Kraus Jr (2003). Distributed learning agents in urban traffic control. The 11th Portuguese Conference on Artificial Intelligence.

Claus, C. and C. Boutilier (1998). The dynamics of reinforcement learning in cooperative multiagent systems. The 15th National Conference on Artificial Intelligence and 10th Conference on Innovative Applications of Artificial Intelligence, Madison, US.

De Oliveira, D., A. L. C. Bazzan, et al. (2006). Reinforcement Learning-based Control of Traffic Lights in Non-stationary Environments: A Case Study in a Microscopic Simulator. EUMAS06, Citeseer.

Diakaki, C., M. Papageorgiou, et al. (2002). "A multivariable regulator approach to trafficresponsive network-wide signal control." Control Engineering Practice 10(2): 183–195.

El-Tantawy, S. and B. Abdulhai (2010). An agent-based learning towards decentralized and coordinated traffic signal control. 13th International IEEE Conference on Intelligent Transportation Systems (ITSC).

El-Tantawy, S. and B. Abdulhai (2010). "Towards multi-agent reinforcement learning for integrated network of optimal traffic controllers (MARLIN-OTC)." Transportation Letters: The International Journal of Transportation Research 2(2): 89-110.

El-Tantawy, S. and B. Abdulhai (2011). Comprehensive Analysis of Reinforcement Learning Methods and Parameters for Adaptive Traffic Signal Control. Transportation Research Board, Washington D.C.

El-Tantawy, S. and B. Abdulhai (2012). Neighborhood Coordination-based Multi-Agent Reinforcement Learning for Coordinated Adaptive Traffic Signal Control. Transportation Research Board Washington D.C.

Farges, J. L., J. J. Henry, et al. (1983). The PRODYN real-time traffic algorithm. The 4th IFAC/IFIP/IFORS Symposium on Control in Transportation Systems, Baden-Baden, Germany.

Gartner, N. H. (1983). "OPAC: A demand-responsive strategy for traffic signal control." Transportation Research Record: Journal of the Transportation Research Board 906: 75-81.

Head, K. L., P. B. Mirchandani, et al. (1992). "Hierarchical framework for real-time traffic control." Transportation Research Record 1360: 82-88.

Hunt, P. B., D. I. Robertson, et al. (1981). "SCOOT-a traffic responsive method of coordinating signals." Technical Report, Transport and Road Research Laboratory, Crowthorne, England.

Kuyer, L., S. Whiteson, et al. (2008). Multiagent reinforcement learning for urban traffic control using coordination graph. The 19th European Conference on Machine Learning.

Li, T., D. B. Zhao, et al. (2008). Adaptive Dynamic Programming for Multi-intersections Traffic Signal Intelligent Control. 11th International IEEE Conference on Intelligent Transportation Systems.

McShane, W. R., R. P. Roess, et al. (1998). Traffic engineering, Prentice Hall.

Medina, J. C. and R. F. Benekohal (2012). Q-learning and Approximate Dynamic Programming for Traffic Control – A Case Study for an Oversaturated Network. Transportation Research Board Annual Meeting.

Quadstone Paramics (2012). Paramics Microscopic Traffic Simulation Software http://www.paramics-online.com.

Richter, S., D. Aberdeen, et al. (2007). Natural actor-critic for road traffic optimisation. Advances in Neural Information Processing Systems. Cambridge, MIT Press. 19.

Salkham, A., R. Cunningham, et al. (2008). A collaborative reinforcement learning approach to urban traffic control optimization. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

Shoufeng, L., L. Ximin, et al. (2008). Q-Learning for Adaptive Traffic Signal Control Based on Delay Minimization Strategy. IEEE International Conference on Networking, Sensing and Control.

Sims, A. G. and K. W. Dobinson (1979). SCAT-The Sydney Co-ordinated Adaptive Traffic System–Philosophy and Benefits. International Symposium on Traffic Control Systems.

Sutton, R. S. and A. G. Barto (1998). Introduction to reinforcement learning. Cambridge Mass., MIT Press.

Thorpe, T. (1997). "Vehicle traffic light control using sarsa." Master's Project Report, Computer Science Department, Colorado State University, Fort Collins, Colorado.

Watkins, C. and P. Dayan (1992). "Q-learning." Machine learning 8(3): 279-292.

Weinberg, M. and J. S. Rosenschein (2004). Best-response multiagent learning in non-stationary environments. The 3rd International Joint Conference on Autonomous Agents and Multiagent Systems.

Wiering, M. (2000). Multi-agent reinforcement learning for traffic light control. The 17th International Conference on Machine Learning