Throughput-Optimal Scheduling for Multi-Hop Networked Transportation Systems With Switch-Over Delay

Ping-Chun Hsieh, Xi Liu, Jian Jiao, I-Hong Hou, Yunlong Zhang, and P. R. Kumar
Texas A&M University
{pingchun.hsieh,xiliu,jiaojian,ihou,yz61,prk}@tamu.edu

ABSTRACT

The emerging connected-vehicle technology provides a new dimension for developing more intelligent traffic control algorithms for signalized intersections. An important challenge for scheduling in networked transportation systems is the switch-over delay caused by the guard time before any traffic signal change. The switch-over delay can result in significant loss of system capacity and hence needs to be accommodated in the scheduling design. To tackle this challenge, we propose a distributed online scheduling policy that extends the well-known Max-Pressure policy to address switch-over delay by introducing a bias factor favoring the current schedule. We prove that the proposed policy is throughput-optimal with switch-over delay. Furthermore, the proposed policy remains optimal when there are both connected signalized intersections and conventional fixed-time ones in the system. With connected-vehicle technology, the proposed policy can be easily incorporated into the current transportation systems without additional infrastructure. Through extensive simulation in VISSIM, we show that our policy indeed outperforms the existing popular policies.

CCS CONCEPTS

•Networks →Network resources allocation; •Theory of computation →*Scheduling algorithms;*

KEYWORDS

Networked transportation system; Throughput-optimality; Switch-over delay; Scheduling

ACM Reference format:
Ping-Chun Hsieh, Xi Liu, Jian Jiao, I-Hong Hou, Yunlong Zhang, and P. R. Kumar. 2017. Throughput-Optimal Scheduling for Multi-Hop Networked Transportation Systems With Switch-Over Delay. In *Proceedings of MobiHoc'17, Chennai, India, July 10-14, 2017,* 10 pages.
DOI: http://dx.doi.org/10.1145/3084041.3084065

MobiHoc'17, Chennai, India

DOI: http://dx.doi.org/10.1145/3084041.3084065

1 INTRODUCTION

Traffic congestion in urban area has been an increasingly severe problem in all cities of different sizes. According to a recent study [19], every driving commuter in the U.S. spends on average 30 to 60 hours of extra time on the road each year. Furthermore, about two thirds of the extra time is from road congestion. For an urban transportation network which consists of intersections as nodes, and roads between intersections as edges, intersections are often the source of road congestion as well as being accident-prone areas [14].

Recently, there has been considerable work in exploring novel scheduling strategies for intersections from the perspective of *networked transportation systems*, which incorporate emerging connected-vehicle technologies such as vehicle-to-vehicle (V2V) communication and vehicle-to-infrastructure (V2I) communication. With connected-vehicle technologies, roadside infrastructure can obtain accurate and real-time information about the number of vehicles waiting in each lane [23]. The scheduling problem in networked transportation systems then becomes very similar to that in computer networks. Each intersection corresponds to a router, each lane corresponds to a queue, and each vehicle corresponds to a packet. Indeed, there have been efforts to apply the well-known Max-Pressure policy of computer networks [22] to networked transportation systems [25].

Currently, most scheduling algorithms manage traffic flows at intersections via traffic signals, whose color switches periodically between red and green. Transition from the green to red phase is not instantaneous, but requires a guard time for safety, usually of about 3-8 seconds [12]. The throughput during this transition phase is nearly zero. In addition, there is also throughput loss when a new green phase starts or ends because of acceleration or deceleration of vehicles. We capture such capacity loss by introducing *switch-over delay* in this paper. The switch-over delay needs to be explicitly addressed in designing scheduling policies for intersections. Unfortunately, most of the existing literature on scheduling intersections via traffic signals ignores the effect of switch-over delay. In fact, Ghavami *et al.* [7] demonstrate that, while dynamic signal control policies such as the Max-Pressure policy outperforms conventional fixed-time policies in general, the performance of the dynamic signal control policies can be seriously affected by capacity loss when switch-over delay is considered.

Furthermore, during the transition from a traditional transportation system to a fully connected system, only some of the intersections are equipped with sensors and V2I/V2V communication [13], while the rest relying on conventional fixed-time control policies. In such *partially-connected systems*, any new

proposed policies will need to coexist well with conventional ones.

This paper aims to address all the above challenges. We propose a distributed scheduling policy for networked transportation systems and formally prove that the proposed policy is throughput-optimal even when there is switch-over delay. The proposed policy accommodates the switch-over delay by adding a bias factor favoring the current schedule. Moreover, we introduce a superframe structure which achieves synchronization among connected intersections and serves as a natural structure for stability analysis. Our main contribution can be summarized as follows:

- Switch-over delay is considered and a throughput-optimal policy is proposed.
- The proposed policy is distributed with low implementation complexity, and therefore scales well with network size.
- The proposed policy does not require any knowledge of traffic demands.
- The policy continues to be throughput-optimal when there are both connected-technology and fixed-time intersections in the system. It therefore performs well even in partially-connected transportation systems.
- The proposed policy is evaluated via realistic microscopic simulation on a standard simulator for transportation research.

While the paper focuses on networked transportation systems, the theoretical results are also applicable to many other applications with switch-over delay, such as optical networks [16], wireless networks with directional antennas [17], and multi-threaded operating systems [6]. The rest of the paper is organized as follows. Section 3 describes the model of intersections and multi-hop transportation systems. The proposed scheduling policy is illustrated and the proof of optimality is provided in Section 5. Section 7 presents the simulation results. Section 8 concludes the paper.

2 RELATED WORK

In current transportation systems, traffic signals are often adaptively controlled by proprietary traffic control suites, such as SCATS [15] and SCOOT [20]. Following the fixed-time control paradigm, these software suites require real-time traffic statistics to optimize cycle splits and offsets in the timing plan for given objective functions. However, traffic demand can change rapidly with time, and it is difficult and costly to collect the required statistics in a timely manner.

Differing from the fixed-time approach, scheduling design based on real-time queue length information is attracting increasing attention due to recent progress in connected-vehicle technology. For example, adaptive control based on queue length is proposed in [23], where the queue length is estimated via probe vehicles with V2I and V2V communication. On the other hand, inspired by results in computer networks [22], Varaiya [25] and Wongpiromsarn *et al.* [26] propose a Max-Pressure policy for signal control and formally prove that it is throughput-optimal when the queue capacity is infinite and the routing rates are known. To relax the assumption of infinite queue capacity, Xiao *et al.* [27] present a variation of the Max-Pressure policy that is throughput-optimal within a reduced capacity region when the queue capacity is finite but large enough. To relax the assumption on routing rates, Gregoire *et al.* [8] also propose a back-pressure-based signal control policy and prove that it is throughput-optimal with unknown routing rates. Despite the above progress, none of these policies takes the switch-over delay into account.

In the existing literature on the scheduling design for systems with switch-over delay, [1, 2, 4, 11] are the most relevant to the scope of this paper. Armony and Bambos [1] study a system of parallel queues with switch-over delay and propose a family of dynamic cone policies and batch policies to achieve optimal throughput. Subsequently, Hung and Chang [11] present a generalized version of the dynamic cone policy to reduce the complexity of the original cone policy. Chan [4] also presents a Max-Weight type policy with hysteresis and prove that it is throughput-optimal for a system of parallel queues with deterministic service processes. Celik *et al.* [2] propose a family of generalized Max-Weight policies and prove that any policy satisfying the proposed criteria is throughput-optimal. As an example in [2], the Variable Frame-Based Max-Weight (VFMW) policy introduces a frame structure to avoid excessive capacity loss due to switch-over delay. However, all the above policies are designed specifically only for single-hop systems and hence the optimality results may not carry over multi-hop systems. In this paper, we regard VFMW as the reference policy for comparison in the simulations. In Section 7, we show that the VFMW policy, which is throughput-optimal for single-hop systems, can actually perform poorly in multi-hop systems.

3 SYSTEM MODEL

We model a multi-hop transportation system by a directed graph $(\mathcal{V}, \mathcal{L})$, where $\mathcal{V}$ denotes the set of intersections and $\mathcal{L}$ is the set of directional links connecting the intersections. Each link has a start node and an end node. In this paper, we use the terms *node* and *intersection* interchangeably. For convenience, we also include a common virtual source node v_{s} as well as a common virtual destination node v_{d} in the directed graph. We assume time is slotted. The links can be further divided into three categories: internal links $\mathcal{L}_{\mathrm{int}}$, entry links $\mathcal{L}_{\mathrm{entry}}$, and exit links $\mathcal{L}_{\mathrm{exit}}$. Each entry link has the same start node v_{s} and an end node $v \in \mathcal{V}$ where $v \neq v_{\mathrm{d}}$. Similarly, each exit link has the same end node v_{d} and a start node $v \in \mathcal{V}$ where $v \neq v_{\mathrm{s}}$. Therefore, entry links and exit links together characterize the boundary of a system. This model can also take garages into account by modeling each garage as an entry link plus an exit link.

Given two links $i, j \in \mathcal{L}$ incident to the same intersection, link i is called a downstream link of j (or equivalently, i is an upstream link of j) if the end node of link i is the same as the start node of link j. We use $\mathcal{D}(i)$ and $\mathcal{U}(i)$ to denote the set of all the downstream and upstream links of each link i, respectively. Without loss of generality, we suppose that each link has at most $U_{\max}$ upstream links. Moreover, the link pair (i, j) forms a *movement* of vehicles. We denote

$\mathcal{M}_v$ to be the set of movements of each intersection $v \in \mathcal{V}$ and define $\mathcal{M} := \cup_{v\in\mathcal{V}} \mathcal{M}_v$. A collection of non-conflicting movements is called an *admissible phase* of an intersection. Figure 1 shows a standard intersection with eight movements and four admissible phases.

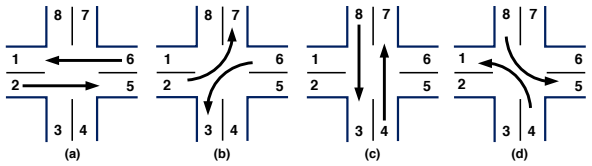


Figure 1: A typical intersection with eight movements and 4 admissible phases.

In this typical intersection, each link has two upstream links and two downstream links. For ease of explanation, we assume that vehicles can only go straight or turn left, but cannot turn right, in this example. Each movement (i, j) has an associated queue $Q_{i,j}$ holding incoming vehicles. In other words, we assume that there exists a separate queue for each left-turn and through movement. We assume that each queue has infinite capacity such that there is no overflow or blockage at each intersection. Throughout this paper, we use the three-tuple $\mathcal{G} = (\mathcal{V}, \mathcal{L}, \mathcal{M})$ to denote a transportation system.

External vehicles enter the system only via entry links. For any entry link i and its downstream link $j \in \mathcal{D}(i)$, let $\{A_{i,j}(t)\}_{t\geq 0}$ be an i.i.d. sequence of external arrivals at $Q_{i,j}$ with average external arrival rate $\lambda_{i,j} > 0$, and $A_{i,j}(t) \leq A_{\max}$ at any time t. For any non-entry link i and $j \in \mathcal{D}(i)$, we simply let $A_{i,j}(t) = 0$ for all t and hence $\lambda_{i,j} = 0$. For ease of later discussion, we also define $\lambda_i := \sum_{j\in\mathcal{D}(i)} \lambda_{i,j}$ to be the total external arrival rate through each link i. Similarly, let $\{S_{i,j}(t)\}_{t\geq 0}$ be an i.i.d. sequence of potential service rates of the movement (i, j), with average service rate $\mu_{i,j}$, for each movement $(i, j) \in \mathcal{M}$. We assume that $S_{i,j}(t) \leq S_{\max}$, for any movement (i, j) and any time t. $S_{i,j}(t)$ captures the variation in the passage time required by different vehicles. Since $S_{i,j}(t)$ depends on instantaneous conditions such as vehicle speed and driver behavior, it is difficult for the traffic scheduler to obtain information about potential service rates. Therefore, we presume that the traffic scheduler only has the information of *average service rate*, which is often called *saturation flow* in the transportation community. The average service rate of a movement is roughly proportional to the number of lanes of that movement [24].

In the multi-hop model, vehicles are routed in a probabilistic manner. When a vehicle enters a link i, it joins a downstream link $j \in \mathcal{D}(i)$ independently with probability $r_{i,j}$, with $\sum_{j\in\mathcal{D}(i)} r_{i,j} = 1$. We assume that $r_{i,j} > 0$, for all movements $(i, j) \in \mathcal{M}$. Let $R_{i,j}(t)$ denote the proportion of vehicles that join $Q_{i,j}$ from among the vehicles entering link i at time t, with $0 \leq R_{i,j}(t) \leq 1$. Since each vehicle chooses its route independently, $\mathbb{E}[R_{i,j}(t)] = r_{i,j}$ for any time t by the basic properties of multinomial random variables. Note that the above model of arrivals, service, and routing is similar to that of the classic open Jackson network.

For each intersection, based on its scheduling policy, at each time slot exactly one of the admissible phases is chosen to have the right of way. Let $I_{i,j}(t)$ be the indicator function denoting whether $Q_{i,j}$ is scheduled at the corresponding intersection at time t. Therefore, for each intersection $v \in \mathcal{V}$, we can use a $|\mathcal{M}_v|$-dimensional binary vector to represent the scheduled phase of the intersection. Let $\mathcal{I}_v$ be the collection of the schedule vectors of all the admissible phases at the intersection v. Then, under a scheduling policy, each intersection v determines $\mathbf{I}_v(t) \in \mathcal{I}_v$ at each time t.

In order to guarantee absolute safety, non-zero time delay is inserted for an intersection to switch the right of way from the current admissible phase to the next. Such lost service time during traffic signal change is modeled as a *switch-over delay*, during which all the movements at the intersection are prohibited and hence the throughput is zero. For simplicity, we assume that the switch-over delay is T_S slot(s) for all the intersections. An intersection is said to be *active* if it is not in switch-over. Let $X_{i,j}(t)$ be the indicator function that the movement (i, j) is active in time slot t. For each intersection, the time between two switch-over events is called a *frame*.

In this paper, each intersection is either a *fixed-time intersection* or a *connected intersection*. A fixed-time intersection simply follows the weighted round-robin policy with the weights determined a priori according to long-term average traffic demands. In contrast, a connected intersection dynamically makes scheduling decisions based on real-time information obtained via connected-vehicle technology, such as queue length. We use $\mathcal{V}_F$ and $\mathcal{V}_C$ to denote the set of fixed-time intersections and connected intersections, respectively.

For simplicity of notation, we use boldface fonts for vectors and matrices throughout the paper. For example, $\boldsymbol{\lambda} = (\lambda_i)_{i\in\mathcal{L}}$ denotes the per-link external arrival rate vector, and $\mathbf{Q}(t) = (Q_{i,j}(t))_{(i,j)\in\mathcal{M}}$ the queue length vector of all the queues in the system.

4 CAPACITY REGION

To study throughput-optimality, we first need to characterize the capacity region of a multi-hop transportation system.

Definition 4.1. A multi-hop transportation system is strongly stable under a scheduling policy π if

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{\tau=0}^{T-1} \sum_{(i,j)\in\mathcal{M}} \mathbb{E}\big[Q_{i,j}(\tau)\big] < \infty. \tag{1}$$

When so, we say that the policy π stabilizes the system.

Next, we define the feasible external arrival rate vectors:

Definition 4.2. Given a multi-hop transportation system $\mathcal{G}$, an external arrival rate vector $\boldsymbol{\lambda} = (\lambda_i)_{i\in\mathcal{L}}$ is *feasible* if there exists a scheduling policy under which the system is strongly stable with $\boldsymbol{\lambda}$.

We can define the capacity region as follows:

Definition 4.3. The capacity region is defined as the closure of the set of all feasible external arrival rate vectors $\boldsymbol{\lambda}$.

To explicitly characterize the capacity region, we first obtain the *effective arrival rate*, which includes both external arrivals as well as arrivals from upstream links of each link, and then characterize the capacity region. Let λ_i^* be the effective arrival rate of link i. According to our model, we have $\lambda_i^* = \lambda_i$ for all $i \in \mathcal{L}_{\text{entry}}$. For any link $j \in \mathcal{L} \setminus \mathcal{L}_{\text{entry}}$, the effective arrival rate is determined by $\lambda_j^* = \sum_{i:j\in\mathcal{D}(i)} \lambda_i^* r_{i,j}$. Let $\boldsymbol{\lambda}^* = (\lambda_i^*)_{i\in\mathcal{L}}$ be the effective arrival rate vector, and $\mathbf{R} = (r_{i,j})_{i,j\in\mathcal{L}}$ the routing probability matrix. Then, we can write the system of traffic equations in matrix form:

$$\boldsymbol{\lambda}^* = \boldsymbol{\lambda} + \mathbf{R}^\intercal \boldsymbol{\lambda}^*, \tag{2}$$

where $\mathbf{R}^\intercal$ is the transpose of the routing probability matrix. Note that (2) is similar to the system of traffic equations of an open Jackson network. Let $\mathbf{1}$ be an $|\mathcal{L}|\times|\mathcal{L}|$ identity matrix. It is easy to verify that (2) has as a unique solution $\boldsymbol{\lambda}^* = (\mathbf{1}-\mathbf{R}^\intercal)^{-1}\boldsymbol{\lambda}$, where $(\mathbf{1} - \mathbf{R}^\intercal)$ is invertible (Section 2.1 in [5]).

For each fixed-time intersection v, let $\xi_v \in (0, 1)$ be the average fraction of time in which the intersection v is in switch-over. Let Λ be the set of all the external arrival rate vectors $\boldsymbol{\lambda}$ for which the following conditions hold: (i) For each fixed-time intersection v, there exists $\epsilon > 0$ and a vector $\boldsymbol{\Sigma}_v = (\Sigma_{i,j})_{(i,j)\in\mathcal{M}_v}$ in the convex hull of $\mathcal{I}_v$ such that the effective arrival rates satisfy

$$\xi_v \mu_{i,j} \Sigma_{i,j} > \lambda_i^* r_{i,j} + \epsilon, \quad \forall (i,j) \in \mathcal{M}_v, \tag{3}$$

i.e. there is at least a small service margin for every movement at v. (ii) For each connected intersection $v \in \mathcal{V}_C$ there exists $\epsilon > 0$ and a vector $\boldsymbol{\Sigma_v} = (\Sigma_{i,j})_{(i,j)\in\mathcal{M}_v}$ in the convex hull of $\mathcal{I}_v$ such that

$$\mu_{i,j} \Sigma_{i,j} > \lambda_i^* r_{i,j} + \epsilon, \quad \forall (i,j) \in \mathcal{M}_v\,. \tag{4}$$

Let $\bar{\Lambda}$ denote the closure of Λ. The following provides a sufficient condition for feasibility of an arrival rate vector.

THEOREM 1. *For a multi-hop transportation system with switch-over delay, an external arrival rate vector $\boldsymbol{\lambda} = (\lambda_i)_{i\in\mathcal{L}}$ is feasible if $\boldsymbol{\lambda} \in \Lambda$.*

PROOF. This can be proved by finding an appropriate fixed-time policy for each connected intersection. By Theorem 1 in [25], we know that given any $\boldsymbol{\lambda} \in \Lambda$, there exists a fixed-time policy for each connected intersection such that the whole system is strongly stable. Hence, $\boldsymbol{\lambda}$ is feasible if $\boldsymbol{\lambda} \in \Lambda$. □

Next, we provide a necessary condition for feasibility.

THEOREM 2. *For a multi-hop transportation system with switch-over delay, if $\boldsymbol{\lambda} \notin \bar{\Lambda}$, then there exists no policy under which the system is strongly stable.*

PROOF. This is a direct result of Theorem 1 in [25]. □

The capacity region can be characterized as follows:

THEOREM 3. *Given a multi-hop transportation system $\mathcal{G}$ with switch-over delay, the capacity region of $\mathcal{G}$ is $\bar{\Lambda}$.*

In this paper, we focus on the interior of the capacity region and define throughput-optimality as follows:

Definition 4.4. Given a multi-hop transportation system $\mathcal{G}$, a scheduling policy π is said to be *throughput-optimal* if the system is strongly stable under π for any external arrival rate vector $\boldsymbol{\lambda} \in \Lambda$.

5 SCHEDULING FOR THROUGHPUT OPTIMALITY

In this section, we introduce our scheduling policy for connected intersections and prove that it is throughput-optimal under switch-over delay.

5.1 A Throughput-Optimal Scheduling Policy

To begin with, we define *pressure* as follows:

Definition 5.1. For any time t, the pressure of a movement $(i, j) \in \mathcal{M}$ is defined as the difference between the queue length of (i, j) and the weighted average of the queue lengths of (j, k) for every $k \in \mathcal{D}(j)$, i.e.

$$W_{i,j}(t) := Q_{i,j}(t) - \sum_{k:k\in\mathcal{D}(j)} r_{j,k} Q_{j,k}(t). \tag{5}$$

In addition, for any intersection v, the pressure of any admissible phase $\mathbf{I}_v = (I_{i,j}) \in \mathcal{I}_v$ is defined as $\sum_{i,j\in\mathcal{M}_v} \mu_{i,j} I_{i,j} W_{i,j}(t)$.

We also introduce a useful definition:

Definition 5.2. A scheduling policy π is said to be *max-pressure-at-switch-over* if π always schedules the phase with the maximum pressure at each switch-over event.

Now, we formally present the *Biased Max-Pressure (B-MP) scheduling policy* in Algorithm 1. In B-MP, time is divided into consecutive *superframes*. At the beginning of a superframe, the duration of the superframe is calculated by (6). Whenever a connected intersection switches, it always switches to a phase with the maximum pressure, and therefore B-MP is max-pressure-at-switch-over. Under B-MP, a connected intersection will only switch under two conditions: (i) at the beginning of each superframe, or (ii) when conditions (7) and (8) specified below are satisfied. From conditions (7)-(8), B-MP only makes a switch when the maximum pressure is larger than the pressure of the current phase by a certain portion. Condition (7) can be interpreted as adding a bias factor favoring the pressure of the current phase, and hence the name B-MP. This bias for the current phase is to prevent the traffic signal from significant capacity loss due to frequent switch-overs.

Moreover, within a superframe, each connected intersection under B-MP can make scheduling decisions independently based on only the local queue length information. Therefore, B-MP is fully distributed within each superframe and the coordination among the connected intersections is minimal. We use t_k to denote the beginning of the k-th superframe, with $t_0 := 0$. Let $T_k := t_{k+1} - t_k$ be the length of the k-th superframe. Let M_k^v be the number of switch-over events in the k-th superframe, for each connected intersection v. Since each superframe may contain a different number of frames at different connected intersections, we use $t_{k,l}^v$ to denote the time

of the l-th switch-over at intersection v in the k-th superframe, setting $t^v_{k,0} := t_k$.

Algorithm 1 Biased Max-Pressure Policy (B-MP)

1: Fix $\beta \in (0,1)$. At time $t = t_k$, set the length of the k-th superframe as:

$$T_k := \Big(\sum_{(i,j)\in\mathcal{M}} Q_{i,j}(t_k) \Big)^{\beta}, \tag{6}$$

and begin the next superframe at $t_{k+1} := t_k + T_k$.

2: Find the phase with the largest pressure at current time t,

$$\mathbf{I}^*_v(t) \in \arg\max_{\mathbf{I}\in\mathcal{I}_v} \sum_{(i,j)\in\mathcal{M}_v} \mu_{i,j} I_{i,j} W_{i,j}(t).$$

Ties are broken arbitrarily.

3: If $\mathbf{I}^*_v(t) \neq \mathbf{I}^*_v(t-1)$, initiate switch-over over the next T_S slots, and then apply the new schedule $\mathbf{I}^*_v(t)$ for one slot. Else, directly apply $\mathbf{I}^*_v(t)$ for one slot.

4: For any $t \in [t^v_{k,l}, t^v_{k,l+1})$ in the rest of the k-th superframe, find the phase $\mathbf{I}^*_v(t)$ that has the largest pressure. If the intersection is not in switch-over at time t, the intersection makes a switch if the following condition is satisfied:

$$\Big(1 + B_v(t^v_{k,l})\Big)\Big(\sum_{(i,j)\in\mathcal{M}_v} \mu_{i,j} I^*_{i,j}(t-1) W_{i,j}(t) \Big)^+ \tag{7}$$

$$< \Big(\sum_{(i,j)\in\mathcal{M}_v} \mu_{i,j} I^*_{i,j}(t) W_{i,j}(t) \Big)^+, \tag{8}$$

where x^+ is a shorthand for $\max\{x, 0\}$, and $B_v(\cdot)$ is the "bias function" defined as

$$B_v(t) := \zeta T_S \min\left\{1, \left(\left[\sum_{(i,j)\in\mathcal{M}_v} W_{i,j}(t) \right]^+ \right)^{-\alpha} \right\} \tag{9}$$

with $\alpha \in (0,1)$ and $\zeta > 0$. Else, stay at the current phase.

5: Repeat Step 3 and 4 until the end of the k-th superframe.

6: At $t = t_{k+1}$, go back to step 1 and repeat the above procedure for the next superframe.

5.2 Proof of Throughput-Optimality

To study system stability, we consider the queue length update over one superframe. Define $\Delta Q_{i,j}(t_k) := Q_{i,j}(t_{k+1}) - Q_{i,j}(t_k)$. For any movement (i,j) with link $i \in \mathcal{L}_{\text{entry}}$, we have

$$\Delta Q_{i,j}(t_k) \tag{10}$$

$$= -\sum_{t=t_k}^{t_{k+1}-1} \Big(S_{i,j}(t) I_{i,j}(t) X_{i,j}(t) \wedge Q_{i,j}(t) \Big) + \sum_{t=t_k}^{t_{k+1}-1} A_{i,j}(t), \tag{11}$$

where $(x \wedge y) := \min\{x, y\}$. Note that the first term of (11) represents the number of vehicles that actually leave $Q_{i,j}$ during the k-th superframe, and the second term is the total number of external arrivals at $Q_{i,j}$ in the k-th superframe.

On the other hand, for any movement $(i,j) \in \mathcal{M}$ with link $i \notin \mathcal{L}_{\text{entry}}$, we have

$$\Delta Q_{i,j}(t_k) = -\sum_{t=t_k}^{t_{k+1}-1} \Big(S_{i,j}(t) I_{i,j}(t) X_{i,j}(t) \wedge Q_{i,j}(t) \Big) \tag{12}$$

$$+ \sum_{t=t_k}^{t_{k+1}-1} \sum_{m:(m,i)\in\mathcal{M}} \Big(S_{m,i}(t) I_{m,i}(t) X_{m,i}(t) \wedge Q_{m,i}(t) \Big) R_{i,j}(t). \tag{13}$$

Note that (13) represents the total number of vehicles coming from the upstream links of i during the k-th superframe.

To study the throughput performance, we analyze the Lyapunov drift over one superframe. Define a Lyapunov function

$$L(\mathbf{Q}(t)) := \mathbf{Q}(t)^{\mathsf{T}} \mathbf{Q}(t) = \sum_{(i,j)\in\mathcal{M}} Q_{i,j}(t)^2, \tag{14}$$

where $\mathbf{Q}(t)^{\mathsf{T}}$ is the transpose of the queue length vector. Define the Lyapunov drift over the k-th superframe as $\Delta L(t_k) := L(\mathbf{Q}(t_{k+1})) - L(\mathbf{Q}(t_k))$. Then, we have

$$\Delta L(t_k) = 2\mathbf{Q}(t_k)^{\mathsf{T}} \Delta\mathbf{Q}(t_k) + \Delta\mathbf{Q}^{\mathsf{T}} \Delta\mathbf{Q}(t_k), \tag{15}$$

where $\Delta\mathbf{Q}(t_k) := \mathbf{Q}(t_{k+1}) - \mathbf{Q}(t_k)$. Given $\mathbf{Q}(t_k)$, the size of the k-th superframe is known and therefore the conditional drift over the k-th superframe is well-defined. Note that it is actually not straightforward to calculate the conditional drift over one superframe for the following reasons:

- For any intersection, there could be multiple frames, and hence multiple phases are scheduled in a stochastic sequence in one superframe.
- Different intersections could possibly have totally different frame sizes in the same superframe.
- Given the queue length information at the beginning of a superframe, it is still not clear when switch-over will be triggered and which phase will be scheduled at each intersection, since the arrival and service processes are stochastic.

Despite the above challenges, the conditional drift over one superframe can still be upper bounded for the max-pressure-at-switch-over policies.

LEMMA 1. *Given any $\boldsymbol{\lambda} \in \Lambda$, under any max-pressure-at-switch-over policy with superframe structure, the conditional drift over one superframe is upper bounded as*

$$\mathbb{E}\Big[\Delta L(t_k) \,\Big|\, \mathbf{Q}(t_k)\Big] \leq -2\epsilon T_k \sum_{(i,j)\in\mathcal{M}} W_{i,j}(t_k)^+ \tag{16}$$

$$+ C_1 \sum_{v\in\mathcal{V}_C} M^v_k \Big(\sum_{(i,j)\in\mathcal{M}_v} W_{i,j}(t_k)^+ \Big) \tag{17}$$

$$+ C_2 \sum_{v\in\mathcal{V}_F} \sum_{(i,j)\in\mathcal{M}_v} W_{i,j}(t_k)^+ + C_3 T_k^2 + C_4 T_k \tag{18}$$

where C_1, C_2, C_3 and C_4 are finite positive constants and $x^+ := \max\{x, 0\}$.

PROOF. With the max-pressure-at-switch-over property, we are able to quantify the pressure of the scheduled phases at any $t \in [t_k, t_{k+1})$ even if the scheduling decision of each frame

is not known. The complete proof is provided in the Appendix A of [10]. □

REMARK 1. *Note that (16) represents the negative drift required for system stability. Also note that (17) and the first term of (18) represent the loss of service due to switch-over at connected intersections and fixed-time intersections, respectively. The second and third terms of (18) stand for the service loss due to possible emptiness of the scheduled queues.*

REMARK 2. *Note that in (17) the service loss due to switch-over is basically a direct sum of the service loss contributed by each connected intersection. In other words, the performances of any two connected intersections are completely decoupled. Due to this feature, Lemma 1 still holds if different connected intersections follow different max-pressure-at-switch-over policies with superframe structure.*

To show that B-MP is throughput-optimal, we introduce a sufficient condition for strong stability in the following lemma.

LEMMA 2. *For any max-pressure-at-switch-over scheduling policy with superframe determined by (6), if there exist some constants $B_0 > 0$, $\epsilon_0 > 0$ such that the conditional drift satisfies*

$$\mathbb{E}\Big[\Delta L(t_k) \,\Big|\, \mathbf{Q}(t_k)\Big] \le B_0 - \epsilon_0 \Bigg(\sum_{(i,j)\in\mathcal{M}} Q_{i,j}(t_k)\Bigg)^{1+\beta}, \quad (19)$$

then we have

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \sum_{(i,j)\in\mathcal{M}} \mathbb{E}\Big[Q_{i,j}(t)\Big] < \infty. \quad (20)$$

PROOF. Define $H(t_k) := \sum_{t=0}^{T_k-1} \sum_{(i,j)\in\mathcal{M}} Q_{i,j}(t_k+t)$. Then, we have

$$H(t_k) \le \sum_{t=0}^{T_k-1} \sum_{i\in\mathcal{L}_{\text{entry}}, j\in\mathcal{D}(i)} \Big(Q_{i,j}(t_k) + \sum_{s=0}^{T_k-1} A_{i,j}(t_k+s)\Big) + \sum_{t=0}^{T_k-1} \sum_{i\in\mathcal{L}_{\text{int}}, j\in\mathcal{D}(i)} Q_{i,j}(t_k).$$

After taking conditional expectation of $H(t_k)$, we have

$$\mathbb{E}\Big[H(t_k) \,\Big|\, \mathbf{Q}(t_k)\Big] \quad (21)$$

$$\le T_k^2 \Bigg(\sum_{i\in\mathcal{L}_{\text{entry}}, j\in\mathcal{D}(i)} \lambda_i^* r_{i,j}\Bigg) + T_k \Bigg(\sum_{(i,j)\in\mathcal{M}} Q_{i,j}(t_k)\Bigg) \quad (22)$$

$$\le B_1 \Bigg(\sum_{(i,j)\in\mathcal{M}} Q_{i,j}(t_k)\Bigg)^{1+\beta} \quad (23)$$

where $B_1 = 1 + \sum_{i\in\mathcal{L}_{\text{entry}}} \lambda_i^* r_{i,j}$ is a positive constant independent of $\mathbf{Q}(t_k)$. Then, by (19),

$$\mathbb{E}\Big[\Delta L(t_k) \,\Big|\, \mathbf{Q}(t_k)\Big] \le B_0 - \frac{\epsilon_0}{B_1} \mathbb{E}\Big[H(t_k) \,\Big|\, \mathbf{Q}(t_k)\Big]. \quad (24)$$

By summing (24) over all the superframes, we have

$$\sum_{k\ge 0} \mathbb{E}\Big[\Delta L(t_k) \,\Big|\, \mathbf{Q}(t_k)\Big] \le \sum_{k\ge 0} \Big(B_0 - \frac{\epsilon_0}{B_1} \mathbb{E}\Big[H(t_k) \,\Big|\, \mathbf{Q}(t_k)\Big]\Big). \quad (25)$$

Given a finite initial condition $\mathbf{Q}(0)$, we have $L(0) < \infty$ and $\sum_{k\ge 0} \mathbb{E}\Big[\Delta L(t_k) \,\Big|\, \mathbf{Q}(t_k)\Big] \ge -L(0)$. Hence, we conclude that

$$\limsup_{T\to\infty} \frac{\sum_{t=0}^{T-1} \mathbb{E}\Big[\sum_{(i,j)\in\mathcal{M}} Q_{i,j}(t)\Big]}{T} \le \frac{B_1\Big(B_0 + L(0)\Big)}{\epsilon_0} < \infty.$$

□

Next, since Lemma 1 involves both queue length and pressure, we provide a useful inequality between total queue length and total pressure as follows.

LEMMA 3. *For any queue length vector $\mathbf{Q} = (Q_{i,j})$ and its corresponding pressure vector $\mathbf{W} = (W_{i,j})$, there exists a constant $\delta > 0$ such that*

$$\sum_{(i,j)\in\mathcal{M}} {W_{i,j}}^+ \ge \delta \Bigg(\sum_{(i,j)\in\mathcal{M}} Q_{i,j}\Bigg). \quad (26)$$

PROOF. We provide a sketch of the proof. We first construct a new system by adding several dummy links and dummy movements to the original system and show that the new system is strongly connected, and that the corresponding routing matrix is invertible. By applying the Perron-Frobenius Theorem to the routing matrix, we obtain a strictly positive eigenvector with a positive eigenvalue. Based on the eigenvector properties, we show that there must exist a constant $\delta > 0$ such that the inequality (26) holds. The complete proof is provided in Appendix C of [10]. □

Note that B-MP is a max-pressure-at-switch-over policy and therefore Lemma 1 holds under the B-MP policy. To characterize the number of switch-over events in one superframe under the B-MP policy, we provide an upper bound on the size of each frame as follows.

LEMMA 4. *Under the B-MP policy, there exists a constant $C_5 > 0$ such that the length of each frame is lower bounded as*

$$T_{k,l}^{\upsilon} \ge C_5 B_{\upsilon}(t_{k,l}^{\upsilon}) \Bigg(\sum_{(i,j)\in\mathcal{M}_{\upsilon}} W_{i,j}(t_{k,l}^{\upsilon})^+\Bigg). \quad (27)$$

PROOF. The proof is provided in Appendix D of [10]. □

With Lemma 4, we are ready to provide a lower bound on the number of switch-over events in one superframe under B-MP.

LEMMA 5. *For any intersection υ under the B-MP policy with bias function defined by (9), we have $\forall k \ge 0$,*

$$M_k^{\upsilon}\Bigg(\sum_{(i,j)\in\mathcal{M}_{\upsilon}} W_{i,j}(t_k)^+\Bigg) = o\Bigg(\Bigg(\sum_{(i,j)\in\mathcal{M}} Q_{i,j}(t_k)\Bigg)^{1+\beta}\Bigg). \quad (28)$$

PROOF. The proof is provided in Appendix E of [10]. □

We are ready to show that B-MP is throughput-optimal.

THEOREM 4. *The B-MP policy is throughput-optimal for any $\alpha \in (0,1)$, $\beta \in (0,1)$.*

PROOF. Since B-MP is a max-pressure-at-switch-over policy with superframe structure, Lemma 1 holds under B-MP. Therefore, by Lemma 3 and the fact that $W_{i,j}(t)^+ \le Q_{i,j}(t)$ for any movement (i,j) and any time t, we have

$$\mathbb{E}\Big[\Delta L(t_k) \,\Big|\, \mathbf{Q}(t_k)\Big] \le -2\epsilon\delta_0 T_k \sum_{(i,j)\in\mathcal{M}} Q_{i,j}(t_k) \tag{29}$$

$$+ C_1 \sum_{v\in\mathcal{V}_C} M_k^v\Big(\sum_{(i,j)\in\mathcal{M}_v} W_{i,j}(t_k)^+\Big) \tag{30}$$

$$+ C_2 \sum_{v\in\mathcal{V}_F}\sum_{(i,j)\in\mathcal{M}_v} Q_{i,j}(t_k) + C_3 T_k^2 + C_4 T_k. \tag{31}$$

By Lemma 5 and the choice of T_k, we know $T_k \sum_{(i,j)\in\mathcal{M}} Q_{i,j}(t_k)$ is the dominating term in (29)-(31). Therefore, there exists a constant $B > 0$ such that

$$\mathbb{E}\Big[\Delta L(t_k) \,\Big|\, \mathbf{Q}(t_k)\Big] \le B - \epsilon\delta_0\Big(\sum_{(i,j)\in\mathcal{M}} Q_{i,j}(t_k)\Big)^{1+\beta}. \tag{32}$$

By Lemma 2, we know that the system is strongly stable under the B-MP policy for any external arrival rate $\boldsymbol{\lambda} \in \Lambda$. Hence, the B-MP policy is throughput-optimal. □

REMARK 3. *By Theorem 4, B-MP can achieve throughput-optimality for any choice of α between 0 and 1. The choice of α can indeed affect the average delay performance, and is a topic that is not addressed in this paper.*

REMARK 4. *The parameter β determines the superframe size for coordination among the intersections. To minimize the coordination overhead, it is recommended that β be close to 1.*

6 EXTENSIONS OF BIASED MAX-PRESSURE POLICY

6.1 Weighted Queue Length

The concept of pressure can be further generalized by using *"weighted queue lengths"*:

Definition 6.1. Let $q_{i,j} > 0$ be the predetermined weight factor of movement (i,j). For each movement (i,j), we define the "weighted queue length" as $\hat{Q}_{i,j}(t) := q_{i,j}Q_{i,j}(t)$, for all t. Then, the generalized pressure is defined as

$$\hat{W}_{i,j}(t) := \hat{Q}_{i,j}(t) - \sum_{k:k\in\mathcal{D}(j)} r_{j,k}\hat{Q}_{j,k}(t). \tag{33}$$

If one substitutes $\hat{W}_{i,j}(t)$ for $W_{i,j}(t)$, the B-MP policy continue to remain throughput-optimal:

THEOREM 5. *The B-MP policy using the generalized pressure in Definition 6.1 is still throughput-optimal for any $\alpha \in (0,1)$, any $\beta \in (0,1)$.*

PROOF. This can be proved by considering the drift of a Lyapunov function: $\hat{L}(\mathbf{Q}(t)) = \sum_{(i,j)\in\mathcal{M}} q_{i,j}Q_{i,j}(t)^2$. The rest of the proof is similar to that of Theorem 4 and hence omitted due to space limitation. □

One important usage of weighted queue lengths is to design a capacity-aware version of the B-MP policy that mitigates the *queue overflow effect* due to finite queue capacity. Queue overflow often occurs when the system operates under oversaturated traffic (even if only for a short period of time). The overflow effect can lead to significant service loss as well as severe delay. Given the information about queue capacity, we can choose $q_{i,j}$ appropriately for each movement (i,j) to reduce the chance of queue overflow. For example, choosing $q_{i,j}$ inversely proportional to the queue capacity of $Q_{i,j}$ is suggested in [9]. In Section 7, we provide an example of applying weighted queue length in simulation.

6.2 Estimated Queue Length With Bounded Error

In networked transportation systems, it might be difficult or expensive to obtain precisely accurate queue length information, due to latency in communication, or random errors in sensor detection. Let $Q^\dagger_{i,j}(t)$ and $W^\dagger_{i,j}(t)$ be the estimated queue length and the corresponding pressure, respectively. If the estimation error of queue length is always upper bounded, then the B-MP is still throughput-optimal with the estimated queue length. We still consider the Lyapunov function $L(\mathbf{Q}(t)) = \sum_{(i,j)\in\mathcal{M}} Q_{i,j}(t)^2$ and the corresponding drift conditioned on $\mathbf{Q}^\dagger_{i,j}(t_k)$. Then, we have the following upper bound on the conditional drift:

LEMMA 6. *Given any $\boldsymbol{\lambda} \in \Lambda$, under the B-MP policy using estimated queue length $(Q^\dagger_{i,j}(t))$, if there exists a constant $B > 0$ such that $\left|Q_{i,j}(t) - Q^\dagger_{i,j}(t)\right| \le B$ for all (i,j) and all t, the conditional drift over one superframe is upper bounded as:*

$$\mathbb{E}\Big[\Delta L(t_k) \,\Big|\, \mathbf{Q}^\dagger(t_k)\Big] \le -2\epsilon T_k \sum_{(i,j)\in\mathcal{M}} W^\dagger_{i,j}(t_k)^+ \tag{34}$$

$$+ C_1^\dagger \sum_{v\in\mathcal{V}_C} M_k^v\Big(\sum_{(i,j)\in\mathcal{M}_v} W^\dagger_{i,j}(t_k)^+\Big) \tag{35}$$

$$+ C_2^\dagger \sum_{v\in\mathcal{V}_F}\sum_{(i,j)\in\mathcal{M}_v} W^\dagger_{i,j}(t_k)^+ + C_3^\dagger T_k^2 + C_4^\dagger T_k \tag{36}$$

where $C_1^\dagger$, $C_2^\dagger$, $C_3^\dagger$ and $C_4^\dagger$ are finite positive constants.

PROOF. The proof is similar to that of Lemma 1. The main differences are: (i) Since the drift is now conditioned on $\mathbf{Q}^\dagger(t_k)$ instead of $\mathbf{Q}(t_k)$, the estimation error introduces an extra term in $\mathbb{E}\Big[\mathbf{Q}(t_k)^\intercal \Delta\mathbf{Q}(t_k) \mid \mathbf{Q}^\dagger(t_k)\Big]$. Due to the boundedness of estimation error, this extra term is at most of the same order as T_k. (ii) For connected intersections, B-MP using $\mathbf{Q}^\dagger(t_k)$ makes scheduling decisions based on $\mathbf{W}^\dagger(t_k)$. Therefore, B-MP is max-pressure-at-switch-over in terms of $\mathbf{W}^\dagger(t_k)$

instead of $\mathbf{W}(t_k)$. Since $Q_{i,j}(t) - Q^{\dagger}_{i,j}(t) \in [-B, B]$, we also have $W_{i,j}(t) - W^{\dagger}_{i,j}(t) \in [-2B, 2B]$, for all (i,j) and all t. As a result, the bounded error in pressure only affects the coefficients of the existing terms in the original drift expression. The complete proof is provided in Appendix F of [10]. □

Now, we are ready to prove that B-MP is throughput-optimal with estimated queue lengths.

THEOREM 6. *If there exists a constant $B > 0$ such that $\left|Q_{i,j}(t) - Q^{\dagger}_{i,j}(t)\right| \leq B$ for all (i,j) and all t, then B-MP is still throughput-optimal using estimated queue lengths $(Q^{\dagger}_{i,j}(t))$.*

PROOF. First, we have $Q_{i,j}(t) - Q^{\dagger}_{i,j}(t) \in [-B, B]$ and $W_{i,j}(t) - W^{\dagger}_{i,j}(t) \in [-2B, 2B]$, for all (i,j) and all t, Also, Lemma 3 holds regardless of the scheduling policy. Therefore, we can rewrite the upper bound in Lemma 6 as

$$\mathbb{E}\Big[\Delta L(t_k) \,\Big|\, \mathbf{Q}^{\dagger}(t_k)\Big] \leq -2\epsilon\delta_0 T_k \sum_{(i,j)\in\mathcal{M}} Q^{\dagger}_{i,j}(t_k) \tag{37}$$

$$+ C^{\ddagger}_1 \sum_{v\in\mathcal{V}_C} M^v_k \Big(\sum_{(i,j)\in\mathcal{M}_v} W^{\dagger}_{i,j}(t_k)^+ \Big) \tag{38}$$

$$+ C^{\ddagger}_2 \sum_{v\in\mathcal{V}_F} \sum_{(i,j)\in\mathcal{M}_v} W^{\dagger}_{i,j}(t_k) + C^{\ddagger}_3 T_k^2 + C^{\ddagger}_4 T_k, \tag{39}$$

where $C^{\ddagger}_1$, $C^{\ddagger}_2$, $C^{\ddagger}_3$, $C^{\ddagger}_4$ are finite positive constants. Furthermore, with a slight modification of the proof we know that Lemma 4 and Lemma 5 still hold when $\mathbf{W}(t_k)$ is replaced by $\mathbf{W}^{\dagger}(t_k)$ under B-MP. By the same argument as that in the proof of Theorem 4, we know that $-2\epsilon T_k \sum_{(i,j)\in\mathcal{M}} Q^{\dagger}_{i,j}(t_k)$ is the dominating term in (37)-(39). Therefore, there must exist a constant $B^{\dagger} > 0$ such that

$$\mathbb{E}\Big[\Delta L(t_k) \,\Big|\, \mathbf{Q}^{\dagger}(t_k)\Big] \leq B^{\dagger} - \epsilon\delta_0 \Big(\sum_{(i,j)\in\mathcal{M}} Q^{\dagger}_{i,j}(t_k) \Big)^{1+\beta}. \tag{40}$$

By a similar procedure as in Lemma 2, we know that (40) is also a sufficient condition for strong stability. Hence, we conclude that B-MP remains throughput-optimal when the error in queue lengths is bounded. □

From Theorem 4, we know that B-MP is also robust to estimation error in queue length information.

6.3 Limitations on Green Period

Conventionally, the timing plan of traffic signals includes a minimum green time to accommodate the vehicle startup delay. Under the B-MP policy, the minimum green time can be easily incorporated by introducing a minimum frame size $T_{G,\min} > T_S$. Then, (27) in Lemma 4 would become

$$T^v_{k,l} \geq \max\left\{ T_{G,\min}, C_5 B_v(t^v_{k,l}) \Big(\sum_{(i,j)\in\mathcal{M}_v} W_{i,j}(t^v_{k,l})^+ \Big) \right\}. \tag{41}$$

With a slight modification of the proof of Lemma 5, the B-MP policy with a minimum frame size still remains throughput-optimal. On the other hand, a maximum green time is sometimes applied in the actuated version of fixed-time policy to avoid excessive delays of minor roads. While this can also be included in B-MP by introducing a maximum frame size $T_{G,\max}$, setting a maximum frame size can result in loss of system throughput since the fraction of time spent on switch-over would always be greater than or equal to $\frac{T_S}{T_{G,\max}}$.

7 SIMULATIONS

We evaluate the proposed policy in VISSIM [18], which is a standard microscopic traffic simulator for transportation systems. In addition to the built-in features for conventional traffic signal control, VISSIM also provides programming integration with MATLAB to support user-customizable traffic control algorithms.

We consider a system of six signalized intersections as shown in Figure 2. In total, there are 10 entry links (4 major entries from the East and the West along with 6 minor entries from the North and the South) and 10 exit links. The number of lanes of each through-traffic link and left-turn link are 3 and 1, respectively.

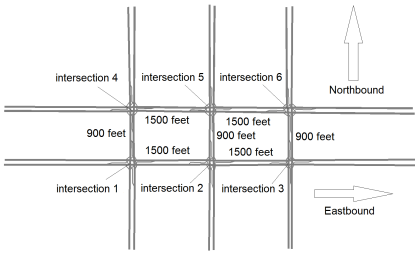


Figure 2: System topology in VISSIM.

Conforming to the official statistics [24], the saturation flow of each link is set to be 1900 vehicles per hour per lane. Vehicles enter the system from the entry links and are routed towards an exit link in a probabilistic manner. We set the routing probabilities to be 0.2 and 0.8 for left-turn movement and through movement, respectively. We use λ_{E}, λ_{W}, λ_{N}, and λ_{S} to denote the arrival rates of the entry links coming from the East, West, North, and South, respectively. We use the default driver behavior and lane-change model provided in VISSIM. The speed limit of each vehicle is 40 miles per hour. Each intersection has four admissible phases as described in Figure 1. Throughout the simulation, we choose the slot time to be 1 second which is sufficient for updating the scheduling decisions. The switch-over delay is set to be 5 seconds, which includes an amber period of 3 seconds and an all-red period of 2 seconds. An important feature of our VISSIM simulation is that we consider the effect of finite buffer size. When a link is fully occupied by vehicles, VISSIM will prohibit the entry of new vehicles, either from the external or from upstream links, and hence lower the throughput.

We compare the B-MP policy against the conventional fixed-time policy, Max-Pressure (MP) policy, and the Variable Frame-Based Max-Weight (VFMW) policy. For the fixed-time policy, the timing plan is calculated by Synchro [21], which is a widely-used optimization tool for timing plan design in transportation research. Throughout the simulation, we assume that the

fixed-time policy has perfect knowledge of the average traffic statistics of each link, and is therefore able to optimize the timing plan accordingly. For VFMW, we choose the frame size to be $T_S + \left(\sum_{(i,j)\in\mathcal{M}_v} Q_{i,j}(t_k)\right)^{0.9}$ as suggested in [3]. For the B-MP policy, we choose $\alpha = 0.01$ and $\beta = 0.99$ as discussed in Section 5.2. To mitigate possible queue overflow due to finite queue capacity, we use weighted queue lengths with $q_{i,j} = 3$ for through-traffic queues and $q_{i,j} = 1$ for left-turn queues, as discussed in Section 6.1. First, we consider the following arrival traffic pattern:

Scenario 1: $\lambda_E = \lambda_W = \bar{\lambda}$ and $\lambda_N = \lambda_S = 0.5 \cdot \bar{\lambda}$ (veh/hr).

Under this traffic pattern, the maximum achievable $\bar{\lambda}$ is about 2600 veh/hr according to the traffic equations given by (2). The total simulation time is 1800 seconds. Figure 3 shows the total number of vehicles in the system with $\bar{\lambda} = 2400$ under the four policies. We observe that B-MP indeed achieves the smallest total queue length while the total queue length is much larger under the other three policies.

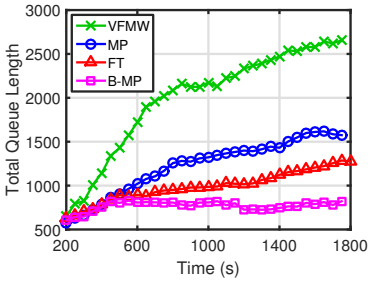


Figure 3: Total queue length of the system under the four policies with $\bar{\lambda} = 2400$.

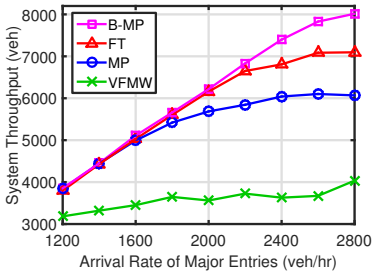
Next, we measure the performance with $\bar{\lambda}$ between 1200 and 2800. Figures 4(a) and 4(b) show the system throughput and average delay for different arrival rates. Note that the average delay here is defined as the difference between the actual travel time and the travel time without any stoppages at the intersections. In Figure 4(a), we see that under B-MP the throughput grows linearly with the arrival rate for $\bar{\lambda}$ up to 2600. At $\bar{\lambda} = 2800$, the throughput under B-MP gets saturated simply because the arrival rate is already beyond the capacity region. Concerning the fixed-time policy, it can support $\bar{\lambda}$ only up to 2200 due to the capacity loss resulting from the switch-over delay. Both MP and VFMW suffer from severe capacity loss due to frequent switching of traffic signals. In Figure 4(b), the B-MP still achieves the smallest delay for every $\bar{\lambda}$. For the heavy traffic condition with $\bar{\lambda} = 2600$, compared to the fixed-time policy with perfect knowledge of traffic statistics, B-MP reduces the average delay by more than 40% without any arrival rate information. For VFMW, we only show the average delay for $\bar{\lambda}$ below 1800 because it performs much more poorly than the other three policies for $\bar{\lambda}$ above 2000.

Next, we consider time-varying arrival rates.

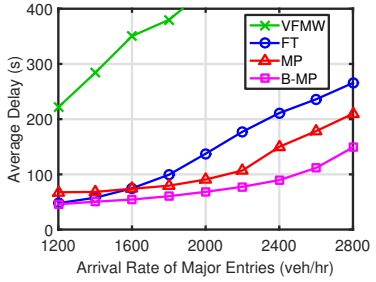
Scenario 2:

- 0 s to 1200 s: $(\lambda_W, \lambda_E, \lambda_N, \lambda_S) = (2000, 2000, 1000, 1000)$.
- 1201 s to 2400 s: $(\lambda_W, \lambda_E, \lambda_N, \lambda_S) = (2500, 1500, 1500, 500)$.
- 2401 s to 3600 s: $(\lambda_W, \lambda_E, \lambda_N, \lambda_S) = (1500, 2500, 500, 1500)$.

Note that the total arrival rate of the whole system remains the same under the above traffic pattern. Figure 5 shows the total queue length under the three policies. Here we omit the VFMW policy simply because it has a much larger total queue length. Again, B-MP still achieves the smallest total queue length at any time. It is notable that the total queue length under B-MP does not change much under the time-varying pattern. In contrast, the fixed-time policy suffers from much more congestion during the period 1200 s to 3600 s. This is because the fixed-time policy optimizes its timing plan based on the average arrival rates and thus fails to accommodate traffic dynamics. Similar to Figure 3, MP still performs quite poorly due to the service loss incurred by the switch-over delay.



(a) System throughput



(b) Average delay

Figure 4: Delay and throughput performance under the four policies for different arrival rates.

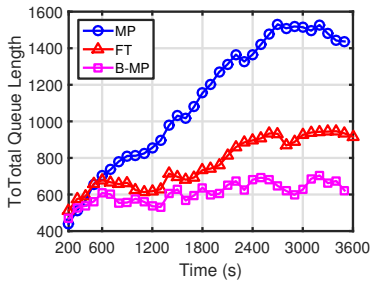
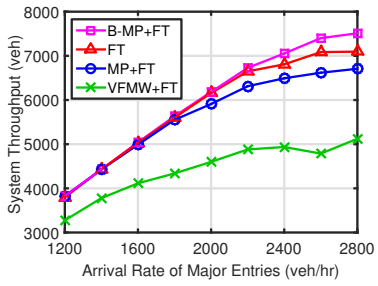


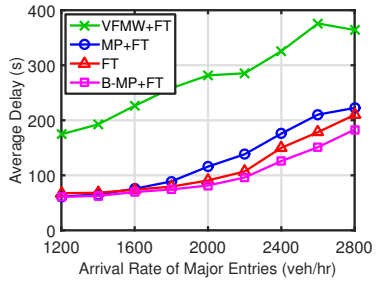
Figure 5: Total queue length under time-varying traffic.

Last, we consider a partially-connected system where three of the intersections are connected under a user-customized

policy (B-MP, MP, or VFMW) and the rest are fixed-time intersections as usual. Figures 6(a) and 6(b) show the average delay and system throughput of the partially-connected system for different arrival rates. Compared to the pure fixed-time system, even partial inclusion of the B-MP policy still provides improvement in both throughput and average delay. Also, B-MP still outperforms the other two policies by a large margin in the partially-connected system. Through the above simulation, we see that B-MP indeed provides significant improvement over the other three popular policies.



(a) System throughput



(b) Average delay

Figure 6: Delay and throughput performance under the four policies in the partially-connected system.

8 CONCLUSION

In this paper, we study the scheduling problem for networked transportation systems with switch-over delay. We propose a distributed scheduling policy that is throughput-optimal with switch-over delay without requiring knowledge of traffic demands. Moreover, the proposed policy still remains optimal when there are both fixed-time intersections and connected intersections in the overall system. Hence, the proposed policy can still perform well in partially-connected systems. Simulation results show that the proposed policy indeed outperforms the other current policies.

ACKNOWLEDGMENTS

This material is based upon work partially supported by NSF under Contract Nos. ECCS-1646449, CCF-1619085, and NSF Science & Technology Center Grant CCF-0939370, USARO under Contract W911NF-15-1-0279, and NPRP Grant 8-1531-2-651 from the Qatar National Research Fund (a member of Qatar Foundation).

REFERENCES

[1] M. Armony and N. Bambos. 2003. Queueing dynamics and maximal throughput scheduling in switched processing systems. *Queueing Syst.* 44, 3 (2003), 209–252.
[2] G. Celik, S. C. Borst, P. A. Whiting, and E. Modiano. 2016. Dynamic Scheduling with Reconfiguration Delays. *Queueing Syst. Theory Appl.* 83, 1-2 (Jun 2016), 87–129.
[3] G. D. Celik and E. Modiano. 2015. Scheduling in networks with time-varying channels and reconfiguration delay. *IEEE/ACM Trans. Netw.* 23, 1 (2015), 99–113.
[4] C. W. Chan, M. Armony, and N. Bambos. 2016. Maximum Weight Matching with Hysteresis in Overloaded Queues with Setups. *Queueing Syst. Theory Appl.* 82, 3-4 (Apr 2016), 315–351.
[5] H. Chen and D. Yao. 2001. *Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization*. Springer-Verlag.
[6] F. M. David, J. C. Carlyle, and R. H. Campbell. 2007. Context switch overheads for Linux on ARM platforms. In *Proc. of Workshop on Exp. Comput. Sci.* ACM, 3.
[7] A. Ghavami, K. Kar, and S. Ukkusuri. 2012. Delay analysis of signal control policies for an isolated intersection. In *Proc. of ITSC*. 397–402.
[8] J. Gregoire, E. Frazzoli, A. de La Fortelle, and T. Wongpiromsarn. 2014. Back-pressure traffic signal control with unknown routing rates. *IFAC Proceedings Volumes* 47, 3 (2014), 11332–11337.
[9] J. Gregoire, X. Qian, E. Frazzoli, A. De La Fortelle, and T. Wongpiromsarn. 2015. Capacity-aware backpressure traffic signal control. *IEEE Trans. Control Netw. Syst* 2, 2 (2015), 164–173.
[10] P.-C. Hsieh, X. Liu, J. Jiao, I.-H. Hou, Y. Zhang, and P. R. Kumar. 2017. Throughput-optimal scheduling for multi-hop networked transportation systems with switch-over delay. *Tech. Report, https://arxiv.org/abs/1701.03991* (2017).
[11] Y.-C. Hung and C.-C. Chang. 2008. Dynamic scheduling for switched processing systems with substantial service-mode switching times. *Queueing Syst.* 60, 1-2 (2008), 87–109.
[12] S. Lämmer and D. Helbing. 2008. Self-control of traffic lights and vehicle flows in urban road networks. *J. Stat. Mech: Theory Exp.* 2008, 04 (2008), P04019.
[13] X. Liu, K. Ma, and P.R. Kumar. 2015. Towards provably safe mixed transportation systems with human-driven and automated vehicles. In *Proc. of CDC*. 4688–4694.
[14] T. J. Lomax. 1997. *Quantifying congestion*. Number 398. Transportation Research Board.
[15] P. R. Lowrie. 1982. The Sydney coordinated adaptive traffic system-principles, methodology, algorithms. In *Proc. of International Conference on Road Traffic Signalling*.
[16] M. P. McGarry, M. Reisslein, and M. Maier. 2008. Ethernet passive optical network architectures and dynamic bandwidth allocation algorithms. *Commun. Surveys Tuts.* 10, 3 (2008), 46–60.
[17] V. Navda, A. P. Subramanian, K. Dhanasekaran, A. Timm-Giel, and S. Das. 2007. MobiSteer: using steerable beam directional antenna for vehicular network access. In *Proc. of MobiSys*. ACM, 192–205.
[18] PTV VISSIM. 2017. Transportation planning, traffic engineering and traffic simulation. (2017). Retrieved May 24, 2017 from http://vision-traffic.ptvgroup.com/en-us/products/ptv-vissim
[19] D. Schrank, B. Eisele, T. Lomax, and J. Bak. 2015. 2015 Urban Mobility Scorecard. (2015).
[20] SCOOT. 2014. SCOOT - The world's leading adaptive traffic control system. (2014). Retrieved May 24, 2017 from http://www.scoot-utc.com
[21] Synchro Studio. 2017. Synchro Studio: Planning & Analysis Software. (2017). Retrieved May 24, 2017 from http://www.trafficware.com/synchro.html
[22] L. Tassiulas and A. Ephremides. 1992. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Trans. Automat. Control* 37, 12 (1992), 1936–1948.
[23] K. Tiaprasert, Y. Zhang, X. B. Wang, and X. Zeng. 2015. Queue Length Estimation Using Connected Vehicle Technology for Adaptive Signal Control. *IEEE Trans. Intell. Transp. Syst.* 16, 4 (2015), 2129–2140.
[24] Transportation Research Board. 2000. *Highway capacity manual*. National Research Council.
[25] P. Varaiya. 2013. Max pressure control of a network of signalized intersections. *Transp. Res. Part C: Emerg. Technol.* 36 (2013), 177–195.
[26] T. Wongpiromsarn, T. Uthaicharoenpong, Y. Wang, E. Frazzoli, and D. Wang. 2012. Distributed traffic signal control for maximum network throughput. In *Proc. of ITSC*. 588–595.
[27] N. Xiao, E. Frazzoli, Y. Li, Y. Wang, and D. Wang. 2014. Pressure releasing policy in traffic signal control with finite queue capacities. In *Proc. of CDC*. IEEE, 6492–6497.