Self-Organizing Traffic Signals for Arterial Control

A Dissertation Presented

by

Burak Cesme

to

The Department of Civil and Environmental Engineering

in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in the field of

Transportation Engineering

Northeastern University
Boston, Massachusetts

March, 2013
ABSTRACT

Actuated signal control is very efficient for isolated intersections, but along arterials it lacks the means to synchronize signals, leading to frequent stops. Industry practice in the United States is to use fixed cycle coordination for arterial control, with signals running with a common cycle and offsets that ensure good progression. However, fixed cycle coordination has many limitations. Among them is the inability to respond to variations in traffic demand, increased delay for non-coordinated movements, and a low degree of flexibility for accommodating transit signal priority (TSP).

This research proposes a new paradigm for arterial signal control, “self-organizing traffic signals”. The proposed logic begins with a foundation of actuated control, but with added rules that can lead signals to synchronize with their “neighboring” intersections to provide coordination. The objectives of this research are:

1. To develop control algorithms that are free of any cycle length, but still have coordination mechanisms through communication among neighboring intersections, making them self-organizing at an arterial level.

2. To develop control algorithms to improve the efficiency of actuated control on local level, particularly with respect to gap-out on multi-lane approaches.

3. To achieve a flexible signal control framework that has the ability to respond to fluctuations in traffic demand and recover from TSP interruptions.

4. To develop control policies for oversaturated arterials, which focus on maximizing the throughput to limit oversaturation and manage growing queues.

To test the performance of self-organizing logic, three real corridors and a benchmark network, were modeled in a micro-simulation model, VISSIM. Simulation results indicated the success of
the proposed self-organizing logic. During under-saturated periods, self-organizing signals resulted in delay reductions of up to 14% compared to an optimized coordinated-actuated scheme without TSP. When TSP was applied, self-organizing signals yielded delay reductions of 19% to 50% in transit delay with very little impact to private traffic.

During oversaturation, simulation tests using the benchmark network showed 45% less delay than coordinated control and 4% less than a real-time optimizing control method designed for oversaturated arterials. Simulation tests on two corridors also showed delay reductions of 8% and 35% compared to coordinated control.
ACKNOWLEDGMENTS

I would like to express my deep appreciation and gratitude to my advisor, Prof. Peter G. Furth, for his continuous guidance, supervision, and support throughout my research. I feel very fortunate to have worked with one of the most brilliant researchers in transportation engineering. Without his guidance and persistent help, this dissertation would not have been possible, and it has been my privilege to work with him.

I would like to express my appreciation to Dr. Daniel M. Dulaski for his assistance and encouragement throughout my research. I would also like to thank to the thesis committee members, Dr. Haris N. Koutsopoulos and Emanuel Melachrinoudis for serving on my Ph.D committee and for their suggestions that greatly improved the quality of my dissertation.

A very special thank you to my friends in Boston for the support they have provided over all these years. Also, many thanks to my friends in Turkey, part of my “big family”. And of course a heartfelt thanks goes out to my girlfriend Jana for all her love, encouragement and patience even during hard times of this study.

Last but not least, I am profoundly thankful to my family for the unconditional love and support throughout my life. I am indebted to my mother Nejla Cesme and to my father Yusuf Cesme for their loving care and understanding. Finally, I owe special thanks to my brother Murat Cesme for his love, support, and humor.
TABLE OF CONTENTS

Chapter 1: Introduction ........................................................................................................... 1
  1.1 Actuated Traffic Signal Control ..................................................................................... 2
  1.2 Fixed Cycle Length Coordination ............................................................................... 4
    1.2.1 Limitations of Fixed Cycle Coordination ............................................................... 5
  1.3 Self Organization ......................................................................................................... 9
    1.3.1 Actuated Control’s Self-Organizing Potential ...................................................... 10
  1.4 Research Objectives .................................................................................................... 11

Chapter 2: Review on Arterial Traffic Signal Control ......................................................... 13
  2.1 Actuated Signal Control ............................................................................................. 13
  2.2 Pre-Timed Signal Control ......................................................................................... 14
  2.3 Coordinated-Actuated Signal Control ..................................................................... 18
  2.4 Adaptive Traffic Control Systems ............................................................................. 21
    2.4.1 Adaptive Traffic Control Systems (ATCSs) with Fixed Cycle ......................... 22
      2.4.1.1 Oversaturation Logic under ATCSs with Fixed Cycle Length ....................... 31
      2.4.1.2 Transit Signal Priority (TSP) Logic under ATCSs with Fixed Cycle Length .... 32
    2.4.2 ATCSs Free of Any Cycle Length ......................................................................... 34
      2.4.2.1 Transit Signal Priority (TSP) Logic under ATCSs that are Free of Any Cycle Length ... 48
  2.5 Summary and Conclusions ......................................................................................... 50

Chapter 3: Sources of Inefficiency in an Actuated Control at Local Level ...................... 52
  3.1 Stop-line vs. Upstream Detector ................................................................................ 52
  3.2 Short Critical Gap ....................................................................................................... 53
  3.3 Short and Variable Minimum Green Interval ............................................................ 56
  3.4 Simultaneous vs. Non-simultaneous Gap-Out .......................................................... 57
  3.5 Gap-out on Multi-lane Approaches .......................................................................... 58
    3.5.1 The Impact of Headway Variability on Actuated Signal Control Performance ... 58
      3.5.1.1 Type I and Type II Errors ............................................................................... 59
    3.5.2 Traditional Gap-out Logic on a Single Lane vs. Multi-Lane Approach .......... 61
    3.5.3 Lane-By-Lane Detection for Multi-Lane Approaches ........................................ 63
  3.6 Inefficiencies of Lagging Left in Actuated Signal Control ........................................ 65

Chapter 4: Multi-Headway Gap-Out and Efficient Lagging Left Turn Logic ..................... 68
  4.1 Multi-Headway Gap-Out Logic for Actuated Control on Multi-Lane Approaches .... 68
    4.1.1 Efficiency of Single Headway and Multi-Headway Gap-Out Logic ................. 70
Chapter 5: Arterial Signal Control Strategies ................................................................. 93

5.1 Secondary Extension Logic: Progression vs. Cycle Length Tradeoff ...................... 94
   5.1.1 Importance of an Arriving Platoon in Providing Secondary Extension ............. 97
   5.1.2 Use of Excess Capacity for Secondary Extension ............................................ 100
   5.1.3 The Maximum Allowed Secondary Extension ............................................... 102
   5.1.4 Granting a Secondary Extension - Criteria ..................................................... 104

5.2 Dynamic Coordination for Coupled Intersections .................................................. 105
   5.2.1 Coordination Logic at Coupled Intersections .................................................. 105
   5.2.2 Power of the Critical Intersection in a Coupled Zone ..................................... 109
   5.2.3 Lead-Lag Phasing to Improve Coordination .................................................... 110
   5.2.4 Secondary Extension Logic for Coupled Intersections .................................... 112

Chapter 6: Control Logic for Oversaturated Arterials .................................................... 114

6.1 Research Papers on Oversaturation ...................................................................... 115

6.2 Control Policies Designed for Oversaturation ....................................................... 116
   6.2.1 Efficient Actuated Control .............................................................................. 116
   6.2.2 Green Truncation in Response to Spillback ..................................................... 117
   6.2.3 Inhibiting Secondary Extensions .................................................................... 119
   6.2.4 Turn Pocket Spillback Prevention .................................................................. 119
   6.2.5 Dynamic Offsets for Coupled Intersections .................................................... 122
   6.2.6 Maximum Green Time .................................................................................... 125

Chapter 7: Self-Organizing Control Logic Applications ................................................ 127

7.1 Under-Saturation Conditions – Beacon Street – Test Case I ................................... 128
   7.1.1 Network Implementation in VISSIM ............................................................... 130
      7.1.1.1 Impact of Overcrowding on Dwell Time .................................................... 131
      7.1.1.2 Varying Demand Conditions and Development of Control Logic ............ 133
   7.1.2 Determination of Coupled Intersections ........................................................ 134
   7.1.3 Transit Signal Priority (TSP) ........................................................................... 135
      7.1.3.1 TSP Application for Beacon Street ............................................................ 136
   7.1.4 Random Interruptions ..................................................................................... 137
   7.1.5 Simulation Results and Analysis ..................................................................... 138
LIST OF FIGURES

Figure 1-1: Actuated Traffic Signal Control [2] ................................................................. 3
Figure 2-1: Signal Coordination and Offsets on a One-Way Street .................................. 15
Figure 2-2: Two-Way Coordination .............................................................................. 16
Figure 2-3: Fixed and Floating Force-Offs in a Coordinated-Actuated Control [12] ......... 20
Figure 2-4: Subsystem Structure in SCATS [20] ............................................................ 27
Figure 2-5: Rolling Horizon Approach .......................................................................... 37
Figure 2-6: Concept of Time Tunnels in InSync [6] ......................................................... 44
Figure 3-1: Detector Placement in an Actuated Control ................................................ 53
Figure 3-2: The Relation between Gap Time and Time Headway ................................... 54
Figure 3-3: Expected Cycle Length (sec) as a Function of Demand, Critical Headway H (sec), and Detector Setback S (ft) [35] ................................................................. 55
Figure 3-4: Intermediate Queue When an Upstream Detector Is Used for Extension ........ 56
Figure 3-5: Low Headway Variability during Saturated Flow Regime ......................... 60
Figure 3-6: High Headway Variability during Saturated Flow Regime ......................... 61
Figure 3-7: Two Extreme Arrival Patterns that are Identically Perceived by Signal Controller under Traditional Detection Scheme .................................................. 63
Figure 3-8: Gap-Out Logic under Lane-By-Lane Detection and Traditional Detection [39] ................................................................. 64
Figure 3-9: Dual Ring with a Lagging and a Leading Left Turn Phase ......................... 65
Figure 4-1: Cumulative Headway Distribution on a Single Lane Approach for Saturation Flow and 50% Saturation Flow ................................................................. 73
Figure 4-2: Cumulative Headway Distribution on a Three-Lane Approach for Saturation Flow and 50% Saturation Flow using Single Headway, Three Headways, and a Six Headway Detection Scheme .......... 74
Figure 4-3: Long Headways during Queue Discharge due to Turning Vehicles Leaving through Lane ...... 77
Figure 4-4: VISSIM Simulation Geometry, Traffic Volumes, and Eight-Phase Dual Ring Control Used to Test Multi-Headway Logic ................................................................. 79
Figure 4-5: Average Delay and Cycle Length under Single Headway (traditional), Lane-by-Lane, and Multi-Headway Detection Logic with Varying Traffic Flow Rates and Detector Setback ......................... 83
Figure 4-6: Dual Ring with a Lagging Left Turn Phase and a Leading Through Phase ......................... 86
Figure 4-7: Slack Time at the Barrier Goes to a Leading Through Phase ......................... 90
Figure 4-8: Spillback Detector to Prevent Turn-Pocket Spillback ................................... 91
Figure 5-1: Arrival Profile for an Approaching Platoon for Secondary Extension Logic ......................... 95
Figure 5-2: Communication from an Upstream Intersection to Obtain Arrival Profile When Secondary Green Extension Horizon is Limited due to Close Intersection Spacing ................................. 97
Figure 5-3: Secondary Green Extension Time versus Cumulative Number of Expected Arrivals for Granting Secondary Extension for an Arriving Platoon ......................................................... 99
Figure 5-4: Affordable Lost Time per Vehicle with Respect to Intersection Volume to Capacity (v/c) Ratio ......................................................................................................................... 101
Figure 5-5: Natural Cycle Length Comparison at Neighboring Intersections to Determine the Maximum Allowed Secondary Green Extension for the Non-Critical Direction ......................................................... 103
LIST OF TABLES

Table 1-1: Natural Cycle Length vs. Coordination Cycle Length .......................................................... 6
Table 2-1: Control Logic of ATCSs Operating with a Common, Fixed Cycle Length [6] ......................... 23
Table 2-2: Transit Signal Priority Logic under ATCS with Fixed Cycle Length .................................. 33
Table 2-3: Control Logic of ATCSs that are Free of any Cycle Length ............................................... 35
Table 4-1: Vehicle Arrivals and Headways on a Three-Lane Approach ............................................. 69
Table 7-1: Arterial Intersection Spacing and Coupled Intersections along the Simulated Segment of Beacon Street, Brookline, Massachusetts .................................................................................. 134
Table 7-2: Simulation Results for Beacon Street With and Without Transit Signal Priority ............... 139
Table 7-3: Average Cycle Lengths by Intersection along Beacon Street under Self-Organizing Logic for each Demand Profile ........................................................................................................ 141
Table 7-4: Average Cycle Length at the Study Intersections along Rural Road under Self-Organizing Logic ........................................................................................................................................... 145
Table 7-5: Simulation Results for Rural Road With and Without Transit Signal Priority (With Actuated Pedestrian Phases) ........................................................................................................ 148
Table 7-6: Simulation Results for the Hypothetical Rural Road: No Pedestrian Minimum Green Constraints and No Signal Priority for Buses ........................................................................ 151
Table 7-7: Simulation Results for the Hypothetical Rural Road: With Pedestrian Minimum Green Constraints and With Signal Priority for Buses Traveling along the Corridor ................................................. 153
Table 7-8: Simulation Results for the Hypothetical Rural Road: With Pedestrian Minimum Green Constraints and With Signal Priority for Cross Street Buses Only ........................................... 155
Table 7-9: Comparison of Total Vehicle Delay for Beacon Street by Time Period for Self-Organizing Control versus Coordinated-Actuated Control Optimized for the Period of Oversaturation (30-60 minutes) .................................................................................................... 161
Table 7-10: Average Cycle length (s) by Intersection and by Period along Rural Road ...................... 162
Table 7-11: Simulation Results for Columbia Pike With and Without Transit Signal Priority .......... 171
Table 7-12: Average Cycle Length under Self-Organizing Signals (No TSP) at the Study Intersections along Columbia Pike ........................................................................................................... 174
Table 8-1: Simulation Results for Rural Road with Different Affordable Lost Time per Vehicle Curves .. 179
Table 8-2: Comparison of Simulation Results for Rural Road under Different $L_{max}$ and $X_{min}$ Combinations .............................................................................................................................................. 181
Table 8-3: Comparison of Simulation Results with Scaled Down Volumes for Rural Road under Different $L_{max}$ and $X_{min}$ Combinations ........................................................................................................ 182
Table 8-4: Comparison of Simulation Results for Beacon Street under Different $L_{max}$ and $X_{min}$ Combinations ........................................................................................................................................ 183
Table 8-5: Simulation Results for Rural Road and Beacon Street under Different Maximum Secondary Extension Lengths ..................................................................................................... 185
Table 8-6: Fraction of Cycles with Secondary Green Extension when the Maximum Secondary Extension Length is 10, 20, and 30 Seconds .................................................................................. 186
Chapter 1 : Introduction

Transportation professionals are constantly faced with the challenge to provide/increase traffic capacity in order to reduce traffic delays while minimizing the magnitude of construction. One approach to deal with that challenge is to improve the efficiency of traffic signal controllers to ensure more efficient use of existing roadway capacities.

Efficient signal control plays a primary role in increasing traffic capacity and reducing signalized intersection delay. Increased traffic capacity helps reduce congestion, which in turn eliminates the need to widen roadways, thereby improving efficiency while maintaining a smaller roadway footprint.

Moreover, efficient signal control can include transit signal priority (TSP), which has the potential to promote the operation of public transportation, thereby reducing auto-dependency and alleviating traffic congestion. TSP is a signal control strategy that provides priority to transit vehicles over non-transit vehicles at signalized intersections to reduce signal delay [1]. TSP increases operating speed and thus decreases passenger riding time. It also improves transit reliability, which decreases passenger waiting time. As a result, the benefits achieved through TSP increase the relative attractiveness of transit.

Efficient signal control can further reduce dependency on automobiles by encouraging walking through providing shorter cycle lengths. Shorter cycle lengths reduce pedestrian delays and encourage better pedestrian signal compliance, which can in turn promote walking.

At isolated intersections, actuated signal control usually provides the most efficient operation by providing lower delay and higher capacity. It can also accommodate effective TSP, which encourages use of public transportation. However, the main disadvantage of actuated
control is that it does not work well on arterial corridors as it cannot provide signal coordination (known as “green waves” or progression), which limits their implementation.

Industry practice in the United States is to use coordination with a fixed cycle length for arterial control, with signals running with a common cycle and offset (i.e., reference point) that provides progression. However, fixed cycle coordination has many limitations. Among them is the inability to adjust to variations in actual traffic demand, longer cycle lengths due to the requirement for a common cycle, which increases delay for non-coordinated movements (e.g., cross traffic, turning traffic, pedestrians), and a low degree of flexibility (due to rigid reference points in the cycle to maintain coordination) for accommodating TSP.

### 1.1 Actuated Traffic Signal Control

The operating principle of actuated traffic signal control is to assign right-of-way to traffic movements until the movements are no longer discharging at or near the saturation flow rate, and then assign right-of-way to the next movement(s).

With actuated control, a phase is served when its turn comes if a vehicle call is registered (i.e., a call detector is occupied) and a phase may be skipped if there is no registered call. Green phases are terminated when either the maximum green time is exceeded (“max-out”) or when the time since the last vehicle detection exceeds a threshold (“gap-out”). The controller continuously monitors the time that the detector has been unoccupied (“gap time”), and, subject to other constraints such as minimum green, ends the green when it exceeds a controller parameter variously called the critical gap, passage time, and vehicle extension. **Figure 1-1** illustrates the extension mechanism of an actuated signal controller.
Figure 1-1: Actuated Traffic Signal Control [2]

This extension strategy of actuated control keeps cycle lengths to a minimum and prevents overflow (a remaining queue when a signal turns red), resulting from random fluctuations in demand. One measure of delay, control delay, which accounts for drivers’ decelerating, stopping, and accelerating at a signal, tends to increase with cycle length and to increase dramatically when there is overflow. Thus, actuated control is far more efficient than pre-timed control, and results in near-minimal delay [3].

Actuated traffic signal control is amenable to aggressive TSP. The most common TSP tactics include green extension and early green. The former tactic extends green phase when an approaching transit vehicle is detected, while the latter tactic shortens the green time of preceding phases (non-transit phase) when a transit vehicle is detected in order to speed up the return to green for the transit phase. If a phase is delayed or shortened due to a TSP interruption, actuated control will naturally compensate by allowing it, on its next green phase, to run until vehicles are no longer discharging at or near saturation flow rate. The signal can quickly recover from priority interruptions, making their impact on traffic delay negligible, and thus allowing
aggressive priority tactics that result in near-zero delay for transit. Actuated operation with aggressive priority is commonly used in Zurich, the Hague, and many other European cities known for fast and reliable tram and bus service.

The only limitation of actuated control at a single intersection level is the difficulty in discriminating low flow rates from saturation flow on multi-lane approaches. This results in inefficient gap-out on multi-lane streets, in which a controller extends green interval well past the end of saturated flow, resulting in wasted green time. The detailed discussion of actuated controllers serving multi-lane approaches is provided in Chapter 4 and 5.

1.2 Fixed Cycle Length Coordination

Along arterials, actuated operation has an important disadvantage – it doesn’t provide good progression (green waves). If signals can be coordinated so the platoon of traffic discharged from one intersection arrives when the signal is green at the next intersection, delay will be minimal regardless of cycle length. The easiest way to provide this type of progression is to use fixed cycle coordination, in which every signal along the arterial operates with a fixed, common cycle length and a pre-timed offset (difference in green start time for the arterial movement) that corresponds to the travel time between intersections.

Fixed cycle coordination can take the form of pre-timed coordinated control, coordinated-actuated control, and adaptive traffic control systems that follow a common cycle length (e.g., SCOOT, Section 2.1). Pre-timed control uses predetermined, fixed cycle length, splits (allocation of the cycle to conflicting phases), and offsets. Coordinated actuated control also has a predetermined, fixed cycle length, splits, and offsets. However, the controller allows unneeded green (slack) to be reallocated from the uncoordinated movements to the coordinated movements
(usually the arterial through movements), but it generally does not allow reallocation in the other direction.

Adaptive control systems with a common cycle length employ standard coordination logic (i.e., sharing a common cycle and having fixed offsets), yet they periodically update timing parameters based on detector data. Typically, an online computer continuously monitors traffic flows over the network and makes small adjustments to cycle length, splits, and offset every few minutes.

Fixed cycle coordination has become the dominant method used for arterial control in the United States. It can easily be configured to provide a green wave in one arterial direction, giving the through traffic in that direction near-zero delay, making it especially suitable for one-way arterials. On two-way arterials, green waves can be provided in both directions if intersection spacing is such that travel time between intersections equals a multiple of half the cycle length. Where spacing is not ideal, finding a compromise between the two directions and using flexibility in whether left turn phases lead or lag can sometimes lead to reasonably good progression. As a result, it has been estimated that coordination can reduce delays and stops by between 10 and 40 percent, depending on the prior method of signal control, traffic flows, and road layout [4].

1.2.1 Limitations of Fixed Cycle Coordination

Even though fixed cycle coordination can significantly improve traffic operations through signal coordination, it also has significant limitations, which are summarized in the rest of this section.

First, the need for a common cycle length results in longer cycle lengths, because it forces all the intersections to adopt the cycle length needed by a single “critical” intersection. This is not a problem for traffic that benefits from a green wave, but for road users that don’t –
pedestrians, transit vehicles that make stops, cross traffic, and turning traffic – longer cycles generally mean longer delay. Because each intersection along an arterial has different cross-traffic and other characteristics, each has its own natural cycle length, meaning the cycle length it would have if operated in isolation. Table 1-1 demonstrates how coordination results in longer cycle lengths at non-critical intersections (i.e., intersections #1, #2, and #4).

**Table 1-1: Natural Cycle Length vs. Coordination Cycle Length**

<table>
<thead>
<tr>
<th>Intersection</th>
<th>Natural Cycle (s)</th>
<th>Coordination Cycle (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>70</td>
<td>100</td>
</tr>
<tr>
<td>#2</td>
<td>60</td>
<td>100</td>
</tr>
<tr>
<td>#3*</td>
<td>100*</td>
<td>100*</td>
</tr>
<tr>
<td>#4</td>
<td>45</td>
<td>50</td>
</tr>
</tbody>
</table>

*Note: * indicates critical intersection.

Intersection #1 and #2 require shorter cycle lengths (70 s and 60 s, respectively) than the critical intersection (100 s); however with fixed cycle coordination; all three would have to run with a 100 s cycle. Since coordination is still possible with double cycling (i.e., an intersection operates at one-half of the cycle length of the remaining intersections of the system), the coordination cycle length for intersection #4 is 50 s, therefore the increase in cycle length is very marginal. However, it should also be noted that double cycling occurs very rarely, because it requires big differences in natural cycle lengths, which is not very common along arterials, particularly when pedestrian minimum green times are accounted for.

Second, with fixed cycle coordination, there is always a considerable chance of overflow unless slack is built into the cycle, and so the “needed” cycle length with fixed cycle control will be longer than the average cycle length that will occur with efficient actuated control. Again, the longer cycle increases delay for pedestrians and others not benefitting from a green wave.
Because coordinated-actuated control allows unneeded green to be used by the coordinated phases but not by the non-coordinated phases, it, too, needs slack in the cycle length to limit the probability of overflow.

Third, fixed cycle coordinated schemes are not usually designed to deal with periods of oversaturation. The primary goal of fixed cycle coordination is minimizing delay or the number of stops through good progression, while the goal of an efficient controller during oversaturation should be maximizing system throughput/capacity (or utilization of capacity), which requires different offsets. As a result, fixed cycle coordination plans generally result in inefficient operation when intersections become oversaturated.

Fourth, fixed cycle coordination is poorly suited to TSP because it has a limited ability to handle priority interruptions. Fixed cycle coordination offers a limited range within the signal cycle in which transit vehicles can get a green extension or an early green start. Equally importantly, it lacks mechanisms needed to compensate movements affected by a priority interruption (e.g., giving them a longer green period in the next cycle) [5]. Coordinated-actuated operation offers some flexibility for compensating a coordinated phase interrupted by a cross-street, but it has no mechanism to compensate the cross street or a turning movement affected by a priority intervention favoring transit operating along the arterial [5]. If a priority interruption is so severe that it “knocks” an intersection out of coordination, the process used to recover synchronization can lead to capacity shortfalls and long delays. Because of this inherent lack of flexibility, many TSP tactics are limited/less aggressive (e.g., allowing only short green extensions, not allowing priority in consecutive cycles) and they show little benefit – often less than three seconds delay reduction per transit vehicle per intersection.
Fifth, longer cycles (e.g., 120 s) that accompany fixed cycle coordination can lead to long periods of unsaturated green, when the signal remains green even though the queue has discharged. Some consider this a safety problem, believing that it promotes speeding and pedestrian non-compliance (higher pedestrian delays associated with longer cycles and gaps in traffic flow due to unsaturated flow encourage pedestrians to make crossings when walk signal is not displayed).

Sixth, the timing plan used is typically optimized for traffic volumes collected at a previous date and time, and therefore can run much (or all) of the day with a plan that is poorly adapted to current traffic conditions. Adaptive control systems with common cycle length overcome this weakness by constantly measuring traffic volumes and updating the signal timing parameters frequently [6].

Finally, local road and traffic characteristics can make it such that the coordination benefits of fixed cycle coordination fall far short of the ideal of both arterial approaches benefiting from a green wave. The irregular and non-ideal intersection spacing prevalent on many arterials makes it such that good two-way coordination is not achievable; progression will either be good for one direction and poor for the other, or a compromised timing plan will be identified in which stops are frequent in both directions. Furthermore, in urban settings, a significant portion of arterial traffic does not continue along the corridor. Traffic turning off the arterial reduces platoon density, which undermines the efficiency of a green wave, and suggests that optimal control should involve stopping mainline platoons in order to restore their density.

The goal of adaptive control systems that employ fixed cycle coordination is to overcome the limitations of pre-timed coordinated and actuated-coordinated control by periodically updating the cycle length, splits, and offset to respond to changes in the current traffic situation;
however they are limited as well. Operating under a fixed cycle length to maintain coordination and allowing only small transitions in every timing update (incrementally updates the cycle length, split and offsets based on the flow information) limit their ability to respond to fluctuations in demands. One other drawback of adaptive control with fixed cycle operation is limited optimization. By the time the controller selects a new plan and starts transitioning, the traffic pattern which initiated this selection has left the system, resulting in a timing plan that is no longer optimized. Finally, they still lack the inherent flexibility due to the fixed cycle operation, not allowing them to apply aggressive TSP.

1.3 Self Organization

Self-organizing systems can provide a useful approach for developing efficient signal control strategies that can address the shortcomings of existing signal controllers. A system is described as self-organizing when local elements interact with each other in order to achieve a global function or behavior, dynamically. This function or behavior is not imposed by one single or several elements, nor is it determined hierarchically; rather it is achieved as the local elements interact with one another. Those interactions produce feedback that regulates the system. Then, the solution is dynamically sought by the elements of the system. In this way, systems can adapt quickly to unforeseen changes as elements interact locally [7] [8].

Self-organization can also be applied to traffic control problem, because traffic signal control can also be treated as an adaptation problem, in which traffic signals should adapt themselves to changing traffic conditions. In fact, any signal control method that relies on local rule-based, decentralized control rather than a centralized approach (e.g., a central computer makes decisions for all intersections in a network) can be considered as self-organizing. In
theory, a centralized approach could also solve traffic control problem, but in practice such an
approach would require too much time to compute the solution.

1.3.1 Actuated Control's Self-Organizing Potential

On the single, isolated intersection level, actuated signal control is self-organizing. Its only
global control is establishing the phase sequence to be followed every cycle; otherwise, rules
govern each signal phase directly, telling them when to start and stop based on information from
phase-specific detectors. What emerges from this decentralized control is cycle lengths and splits
that adjust themselves to actual demand as if they had been globally optimized.

At the arterial level, too, actuated control has at least one self-organizing coordination
mechanism. It comes through the information carried downstream in platoons between adjacent
intersections. Actuated signals are programmed to end their green upon detection of the end of a
passing platoon (gap-out logic). If the offset between two neighboring signals ever becomes such
that a platoon discharged from the first signal arrives while the second signal is green, the second
signal will hold its green until the platoon passes, and thus the two signals will end their green
with an ideal progression offset. Suppose now that the cross-traffic demands at the two
intersections are such that the red interval at the first intersection is just as long as or a little
longer than the second intersection’s red interval. The next platoon released from the first
intersection will again arrive on green at the second intersection, maintaining an ideal end-of-
green offset. However, if the first intersection’s red interval is shorter than the second
intersections, or is much longer, then this synchronization can be lost.

Gershenson and Rosenblueth have further demonstrated that actuated control has the
potential for achieving self-organization in street networks [7]. Their test network was a grid of
one-way streets, with alternating northbound and southbound streets crossed by alternating
eastbound and westbound streets. A set of local rules governs each intersection approach based on local detectors. Visualizations from their simulation model show that strong coordination is achieved, with platoons advancing with little or no delay through successive intersections in all directions. This approach shows lower average delay than with a globally optimized coordination plan that provides green waves for all directions, because the self-organizing system adapts to random fluctuations in demand.

However, it is not clear to what extent the remarkable coordination achieved by Gershenson and Rosenblueth is an artifact of their simplified model with one-way streets, equal average traffic demand on all streets, no turning traffic, and no lost time when switching control. Still, their results raise the question of whether other coordination mechanisms can be developed for actuated control that will lead to self-coordinating behavior (coordination achieved organically through self-organizing local rules) in a realistic network with realistic flows, while maintaining the essential benefits of actuated control.

1.4 Research Objectives

This research is focused on developing a new paradigm for arterial traffic signal control, “self-organizing traffic signals”. This approach begins with a foundation of actuated control, but with added rules that can lead signals to synchronize with their neighbors to provide coordination. The proposed coordination logic under self-organizing signals does not require a fixed, common cycle length or a fixed reference point. The coordination of signals is achieved dynamically, based on real-time traffic information, and through communication between adjacent intersections (“peer-to-peer communication”). The objectives of this research are described as follows:
1. To develop signal control algorithms that are based on the addition of new rules to actuated operation to enable coordination of intersections along an arterial. Coordination mechanism relies on simple, local intersection rules and communication among neighboring intersections.

2. To develop control algorithms to improve the efficiency of actuated control on local level, particularly with respect to gap-out on multi-lane approaches.

3. To achieve a flexible signal control framework that has the ability to respond to fluctuations in traffic demand and recover from TSP interruptions. This will allow signal controller to apply aggressive TSP without causing major impacts to non-transit traffic.

4. To develop control policies for oversaturated arterials, which focus on maximizing the capacity/throughput to limit the extent of oversaturation and manage growing queues.

5. To develop generic self-organizing control logic that can be applied to every intersection regardless of the layout (e.g., closely or largely-spaced intersections) and traffic volume.

6. To perform a comprehensive evaluation of the proposed self-organizing control logic on a simulation environment using a widely-accepted traffic micro-simulation software.
Chapter 2: Review on Arterial Traffic Signal Control

Arterial traffic signal control aims at providing progressive traffic flow for through vehicles, which is accomplished by the coordination of traffic signals. With signal coordination, platoons of vehicles can travel through multiple intersections with almost zero delay.

The three most common arterial traffic control strategies are pre-timed, coordinated-actuated, and adaptive control. This chapter presents the pros-and cons of pre-timed, coordinated-actuated, and adaptive traffic control strategies. While actuated control is typically not implemented on arterial level, a very brief discussion of actuated control is also included.

2.1 Actuated Signal Control

Actuated control is an efficient mode of operation at isolated intersections (typically spacing is larger than 0.50 miles can be taken as a threshold), where coordination between adjacent signals is not a concern. At isolated intersections, actuated signals provide greater efficiency than the fixed cycle timing plans by servicing phases only when there is traffic demand, minimizing cycle length, and ultimately reducing average delay.

Along arterials, coordination can arise with actuated control where natural cycle lengths vary little at adjacent intersections. When cross-streets have similar demands, therefore similar red period lengths, if vehicles released from an upstream signal arrives the next intersection before the arterial phase gaps-out, standard extension rule will hold the signal green until the platoon clears the intersection, spontaneously forming green waves. Minimum green interval may also help improve progression by providing a buffer for the front side of a green wave. If the red period is shorter than usual due to low cross street demand, minimum green constraint may protect the arterial green phase from gapping-out, which prevents the arriving platoon from stopping and waiting a long red period.
However, there are many situations in which actuated control’s coordination mechanism is not sufficient to guarantee progression. If a downstream intersection has a shorter cycle length than its upstream neighbor and minimum green is not long enough to provide buffer, green will be ended at the downstream signal before the platoon arrives. If the downstream signal cycles slower, it may not be ready to serve the arriving platoon.

With each intersection operating at its own cycle, it becomes difficult to create a regular pattern of green waves for arterial traffic. Therefore, actuated operation is generally not implemented along arterials.

2.2 Pre-Timed Signal Control

Pre-timed control employs pre-determined and fixed cycle length, splits, and offsets, determined based on historical data, to enable signal coordination. Figure 2-1 describes the concept of signal coordination on a one-way street. Off\(_{2,1}\) is the offset between intersections 1 and 2, and Off\(_{3,1}\) is the offset between intersections 1 and 3. With this approach, vehicles released from intersection 1 can progress through intersections 2 and 3 without having to make stops due to signal coordination. Bandwidth is the width of the interval each cycle within which traffic can advance at the progression speed without hitting a red signal.
Coordination on a one-way street is relatively simple as the objective is only to provide good progression for one direction. However, providing good progression for both directions along a two-way arterial can be difficult to achieve. If intersections are irregularly or closely spaced, there is typically a compromise between coordinating for both directions (generally comes with the cost of decreased bandwidth) and coordinating for the direction with higher volumes (e.g., coordinating inbound traffic in the morning peak and outbound traffic in the afternoon peak).

Figure 2-2 shows three intersections in which intersection spacing is not ideal for providing good two-way coordination. Northbound (shown by solid line) is the favored direction with good progression, while southbound (shown by dashed line) has a very short bandwidth, which leads to long delays except during periods of low traffic. One way to enhance progression
(to increase bandwidth) for southbound would be giving an earlier green start at intersection 3 relative to intersection 1 and 2; however doing so would reduce the bandwidth for the northbound direction.

![Diagram of two-way coordination](image)

**Figure 2-2: Two-Way Coordination**

To handle the challenge of two-way coordination along irregularly spaced arterials, several signal timing optimization programs were developed to analyze and optimize signal timings on arterials. The two most popular ones are TRANSYT-7F (TRAffic Network Study Tool, version 7, Federal) and PASSER (Progression Analysis and Signal System Evaluation Routine). Both software packages follow fixed, common cycle coordination. The timing plans are generated based on off-line demand and is not responsive to real time demand fluctuation.
TRANSYT 7-F is used for delay optimization. It uses a combination of exhaustive, hill climbing, and genetic algorithm optimization methods to calculate optimum cycle length, splits, phase sequence, and offsets [9]. Selection of splits and offsets can also be performed for a user-specified cycle length. In performing the optimization, TRANSYT-7F has the ability to offer double-cycling, overlaps, and user-specified bandwidth constraints [10].

PASSER is an arterial based bandwidth optimizer, which determines phase sequences, cycle length, and offsets. Splits are determined using an analytical (Webster's) method, but are fine-tuned to improve progression [11]. PASSER maximizes bandwidth efficiency by finding the highest value of summing the bandwidth for arterial phases divided by twice the cycle length. No double cycling is permitted.

The main problem with PASSER is that optimizing bandwidth is not the direct objective of traffic signal control, because minimized bandwidth does not result in minimal delay. The objective of efficient control should be minimizing delay and stops during under-saturated conditions (oversaturation environment has different objectives, explained later). Also, bandwidth optimization may lead to unnecessarily long cycle lengths, causing very long delay for cross-street, turning traffic, and pedestrians.

Pre-timed control is often desirable within a downtown area, which has closely spaced intersections with limited queuing space, because having fixed splits and offsets help controllers manage and control queues at closely spaced intersections, avoiding spillback and gridlock.

The main limitation of pre-timed control (including TRANSYT-7F and PASSER) is that it has the disadvantages of fixed cycle coordination, which is summarized as follows (also discussed in Chapter 1).
• The need for a common cycle length to maintain coordination and slack to avoid overflow cause long cycles, which in turn causes higher delay for movements that cannot benefit from coordination (e.g., turning traffic, cross-street traffic, pedestrians).

• The rigidity of fixed-cycle coordination makes it ill-suited to TSP. As a result, it only allows limited TSP (e.g., short green extension, not giving priority in every cycle), which generally results in very little benefit to transit vehicles (less than two seconds delay reduction per intersection).

• When intersection spacing is not ideal for two-way progression, good progression is provided for only one direction, causing long delay and stops for the other.

• Pre-timed control often leads to long unsaturated green due to long cycles. This may lead to safety issues by promoting speeding and encouraging pedestrians to make risky crossings.

• Finally, during oversaturation, pre-timed control (or control methods that employ fixed cycle coordination) performs poorly. Pre-timed control is designed for under-saturated traffic environment, in which the control objective is to minimize delay, while the main objective during oversaturation is to maximize throughput, which requires different timing parameters.

2.3 Coordinated-Actuated Signal Control

Coordinated-actuated control, the current industry standard along arterials, tries to combine the benefit of actuation with coordination benefits to respond to demand fluctuations. It uses a common, fixed cycle length and fixed offsets. Non-coordinated phases are set on actuation and follow standard actuation rules (e.g., skipping and gap-out); therefore their green phase can be
terminated early. No detection is provided for coordinated phases, and they are guaranteed to have at least their pre-specified green time.

For non-coordinated phases, controllers use force-off, that is, points in the cycle where non-coordinated phases must end their green time even if there is continued demand. Force-off ensures that coordinated phases will receive their allocated (pre-specified) green time and ensures that intersection offsets and cycle length will be maintained.

There are two options for programming force-offs in controllers, fixed or floating. The fixed force-off maintains the phase force-off point within the cycle. If a previous non-coordinated cycle ends its phase early, any following phase may use the extra time up to that phase’s force off. As a result, a phase later in the sequence (before the coordinated phase) may receive more than its split time (provided the maximum green is not reached). However, a phase after the coordinated phase will never have an opportunity to receive time from a preceding phase, regardless of the method of force-offs [12].

Floating force-offs are limited to the duration of the splits that were programmed into the controller. The force-off maintains the non-coordinated maximum times for each non-coordinated phase in isolation of one another. Floating force-offs are more restrictive for the non-coordinated phases. If a phase does not use all of the allocated time, then all extra time is given to the coordination phase. Fixed and floating force-offs are illustrated in Figure 2-3.
In the given example, phase 2 is the coordinated phase. With the current demand, phases 1 and 3 gap-out, and phase 4 maxes-out due to high demand (pre-specified split is 25 seconds where demand needs 40 seconds). With fixed force-offs (scenario c), the green time for phase 4 is extended to serve an increased demand up to the force-off point, receiving additional time from phase 3. The additional time from phase 1 due to gap-out is given to the coordinated phase. Under floating force-offs (scenario b), phase 4 would be forced off even when with the higher demand at its split value, 25 seconds, and all extra time would be given to the coordinated phase.

Fixed force-offs allow the extra green time to be used by non-coordinated phases (except the phase(s) following the coordinated phase) having excess demand, increasing the demand-responsiveness of coordinated-actuated control. Fixed force-offs may also reduce the early return to coordinated phases, which can be helpful in a network with closely spaced intersections.
However, limiting early return can also be a disadvantage under congested conditions on the arterial, because an early return can help queue clearance for coordinated phases.

Even though coordinated-actuated control tries to combine coordination with actuation, it is essentially pre-timed control with only a marginal level of actuation. Because the cycle length is still fixed and common along an arterial and non-coordinated phase’s green time may not be extended in case of high demand, coordinated-actuated control cannot address the limitations of fixed cycle-coordination. It still causes long cycles, leading to long delays for non-coordinated movements; it still lacks the inherent flexibility and does not provide compensation mechanism, resulting in less effective TSP; and it still cannot deal with oversaturation.

2.4 Adaptive Traffic Control Systems

Adaptive traffic control systems (ATCSs) adjust signal timings online based on the current traffic conditions, demand, and system capacity [6]. Their primary objective is to be able to respond to within-day and day-to-day traffic fluctuations.

Although adaptive controllers share similar goals (e.g., efficient traffic control that can accommodate fluctuations in traffic demand), the features and the limitations of each ATCS differ significantly with regards to traffic control logic, detection requirements, and system architecture. Note that standard actuated operation can also be considered as adaptive, because it provides variable green times and cycle length based on traffic demand at isolated intersections. However, this section focuses on ATCSs that can provide regular pattern of green waves along an arterial, and, thus actuated signal control is excluded in the analysis.

Many ATCSs use fixed, common cycle length and offset, which is covered in Section 2.4.1. In order to cope with fluctuations in traffic demand, they periodically and incrementally update the signal timing parameters (i.e., cycle length, splits, and offsets). Although the
incremental changes in timing plans introduce certain flexibility in the system in responding to demand fluctuations, they still employ fixed, common cycle coordination. As a result, their control logic is not any different than coordinated actuated control. The extra green time (slack) is still used by the coordinated phases and non-coordinated phases still use fixed green end points, in which they are forced to end their green even when there is high demand.

Other ATCS do not use a fixed cycle operation, fixed splits, and offsets. Cycle length varies from one intersection to another. Those types of ATCSs are defined as systems that are free of cycle length in this study and discussed in Section 2.4.2.

2.4.1 Adaptive Traffic Control Systems (ATCSs) with Fixed Cycle

ATCSs with fixed cycle can be defined as control systems in which intersections along an arterial or in a network follow a common, fixed cycle length, with fixed splits and offsets, but with a layer of adaptive control that makes incremental adjustments to cycle length, splits, and offsets based upon traffic data collected over a period of few minutes. Traffic data are generally obtained using loop detectors or video cameras. Detection requirements (number and location of detectors) vary based on operational objectives of control logic.

The important features and main limitations of some of the more notable ATCSs are described below in details. The following table provides a brief description of ATCSs operating under a fixed cycle length and discusses their operational principles.
Table 2-1: Control Logic of ATCSs Operating with a Common, Fixed Cycle Length [6]

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th>Control Logic</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFAULT</td>
<td>Default ATCS uses fixed, common cycle at all intersections, fixed phase sequence, and standard gap-out logic for non-arterial phases. Any control logic different than default is presented below.</td>
</tr>
<tr>
<td>ACS LITE</td>
<td>Updates split and offset based on measured demand every five to 10 minutes. Cycle length is not adjusted, but selected according to time-of-day schedule (off-line).</td>
</tr>
<tr>
<td>LA ATCS</td>
<td>Updates cycle length, split, and offset every cycle. Double cycling is permitted.</td>
</tr>
<tr>
<td>SCOOT</td>
<td>Updates cycle length every five minutes; updates splits and offset every cycle. Updates cycle length, split, and offset every cycle. The whole system is divided into a large number of smaller subsystems ranging from one to 10 intersections each. Each subsystem requires common cycle length, but cycle length at different intersections can be different. Subsystems are linked and unlinked to each other adaptively.</td>
</tr>
<tr>
<td>SCATS</td>
<td></td>
</tr>
<tr>
<td>MOTION</td>
<td>Updates cycle length, split, and offset in every 10-15 minutes. Number of phases is also determined, rather than being fixed. Allows variable phase sequence.</td>
</tr>
<tr>
<td>BALANCE</td>
<td>Selects the program with the best cycle time from a prearranged set of signal programs based on real-time demand. Determines the splits and cycle length according to the selected cycle length. Allows variable phase sequence. Cycle length, offset, and split are determined every five minutes.</td>
</tr>
<tr>
<td>VFC-OPAC&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Calculates cycle length once every few minutes as specified by the user. Calculates splits and offset every cycle.</td>
</tr>
</tbody>
</table>

Note:
1. The earlier versions of OPAC do not follow a fixed common cycle length (i.e., free of any cycle length). However VFC-OPAC, developed for arterials and networks, employs standard coordination rules, and is therefore included in the table.

**Adaptive Control Software Lite (ACS Lite)**

ACS Lite uses standard coordinated-actuated logic, and updates splits and offsets based on real-time information. Cycle length is currently not adjusted by ACS Lite (although future enhancements are planned). It relies on stop line detectors to monitor volume and occupancy on green [13]. On approaches where progression is desired, advance loops are used to monitor cyclic flow profile (i.e., arrival of vehicles during red and green) so that offsets can be adjusted a few seconds earlier or later to increase the fraction of arrivals on green. ACS Lite makes incremental adjustments to split and offset parameters as often as every 5 to 10 minutes. This improves the efficiency of signal controller if the traffic pattern changes are gradual, however
when changes occur abruptly (e.g., sporting events or incidents), adjustments in every 5 or 10 minutes limit ACS Lite’s responsiveness to traffic.

The cycle length under ACS Lite is dictated by the “underlying” or “baseline” timing plan, which is selected according to the time-of-day schedule (based on off-line data) [14]. Split adjustments (typically limited to 2-5 seconds) are based on balancing the “green utilization” of each phase (i.e., estimation of degree of saturation of each phase). The system provides an optional “progression biasing mechanism”, giving higher priority to designated progression phases by giving them extra or slack green time, if available [15].

The research and development of ACS Lite has been done by Siemens, with collaboration from other partners including major signal control manufactures Eagle, Econolite, McCain, and Peek, who accepted an invitation from Federal Highway Administration (FHWA) to participate in the project.

**LA Adaptive Traffic Control System (LA ATCS)**

Like ACS Lite, LA ATCS also operates under coordinated actuated control (fixed duration cycle, fixed phase sequence). A centralized computer controls all intersections in the network. The controller collects and analyzes detector data (e.g., volume and occupancy data are obtained in every second, but utilized every cycle) to determine real-time traffic demand and adjusts cycle length, phase split, and offset on a cycle-by-cycle basis [16]. This gives the system the ability to respond to spikes in traffic demand. To avoid overreacting, when adjustments are minor, it does not perform any optimization, but it applies heuristic formulas based on extensive operational experience. Detectors are usually located 200 to 300 ft upstream of the intersection.
The system allows for limitation in variation of cycle lengths by providing an upper and a lower limit. The control logic does not vary its phase sequence, but phases can be omitted based on the real-time traffic demand. Double cycling is permitted at minor intersections.

LA ATCS, developed by Los Angeles Department of Transportation (LA DOT), was first deployed as a part of the Automated Traffic Surveillance and Control (ATSAC) Center in 1984 for the Los Angeles Olympic Games [6].

**Split Cycle Offset Optimization Technique (SCOOT)**

Based on detector measurements upstream of an intersection, SCOOT computes the cyclic flow profile for every traffic link and it projects those profiles to the downstream intersection assuming a pre-defined cruising speed as well as a dispersion model in order to estimate queue profiles at the downstream intersection. The primary objective is to minimize the sum of the average queues in the area, which is expressed as a performance index (PI).

Advanced detectors transfer vehicle information (which is then converted to queue estimates) to a central computer and the computer makes control decisions periodically and incrementally to minimize sum of average queues. The optimization process of SCOOT is summarized as follows [17] [18].

- A few seconds before every phase change, SCOOT determines whether it is better to advance or retard the schedule change by four seconds, or leave it unaltered.
- Once per cycle, SCOOT determines whether the PI can be improved by reducing or increasing each offset by four seconds.
- SCOOT varies the cycle time in four, eight or 16 second increments once every five minutes, or every two and a half minutes when cycle times are rising rapidly. It identifies the critical intersection within the region (any of the intersections in a region can
determine the system cycle length; SCOOT is not constrained by a “master” intersection in determining system cycle length) and will attempt to adjust the cycle time to maintain the critical intersection at 90% saturation on each critical phase. For example, if saturation exceeds 90%, the cycle optimizer will add four, eight, or 16 seconds depending on level of saturation. The main advantage of small cycle adjustments is that they can always be accommodated immediately in the following cycle without creating any disruption in progression. However, small adjustments reduce the demand responsiveness of SCOOT.

**Sydney Coordinated Adaptive Traffic System (SCATS)**

SCATS was developed by the Roads and Traffic Authority (RTA) of New South Wales, Australia. In SCATS, traffic control is governed by two levels, strategic (network) and tactical (local).

Network control logic is managed by the regional computers and is known as the Masterlink mode of operation. Using real-time flow and occupancy data collected at the intersection from detectors, the computers determine, on an area basis, the optimum cycle time, phase splits, and offsets [19]. Network controller provides overall system control of cycle time, phase split, and offset. The local controller provides flexibility within the constraints of predetermined network control parameters. As a result, local control logic primarily allows for green phases to be terminated early when the demand is low and for phases to be omitted entirely from the sequence if there is no demand (i.e., follows conventional vehicle actuation rules, unless prohibited by instruction from network control decisions). The local controller bases its decisions on information from vehicle loops at the intersection, some of which may also be used for network detectors.
SCATS manages groups of intersections that are called subsystems (predefined). Each subsystem consists of a number of intersections, usually between one to 10. Figure 2-4 illustrates SCATS subsystem structure used along State Route 224 (SR 224) in Park City, Utah [20]. Ellipses indicate each subsystem and dots show intersection within each subsystem.

One of the intersections in a subsystem is designated as the controlling or critical intersection. The intersections in a subsystem form a discrete group, which are always coordinated together (i.e., follows standard fixed cycle coordination, all intersections share a common cycle, variable phase sequence is allowed).

![Subsystem Structure in SCATS](image)

Figure 2-4: Subsystem Structure in SCATS [20]
To provide coordination over larger groups of signals (i.e., not only coordination within a subsystem), subsystems link and unlink with other subsystems to form larger systems, all operating on a common cycle length. This is important because when the subsystems are unlinked and can operate independently, only intersections within one subsystem are forced to share a common cycle length. Therefore, one or more subsystems can operate more efficiently at a lower cycle time. However, intersections within one subsystem still operate under a common cycle length dictated by the critical intersection, causing longer cycles for non-critical intersections. A SCATS regional computer has a maximum of 250 subsystems.

SCATS network control bases its decision on degree of saturation (known as X in Highway Capacity Manual). Using loop detectors at the critical intersections, the local controller collects flow and occupancy data during the respective green phase, and these data are sent to the regional computer to calculate X. The X is used as a basis for determining whether an increase or decrease in both cycle time and phase split is required.

Cycle time increases or decreases to maintain X at a user-definable value (90% typically) on the lane with the greatest degree of saturation. A lower limit for cycle time (usually 30 to 40 seconds) and an upper limit (usually 100 to 150 seconds) are specified by the user. Cycle time can normally vary by up to six seconds each cycle, but this limit increases to nine seconds when a trend is recognized.

Phase splits can be varied by up to four percent of cycle time each cycle so as to maintain equal degrees of saturation on the competing approaches. The minimum phase split is either a user definable minimum or a value determined from the local controller’s minimum phase length which may vary according to whether there is a pedestrian call.
In selecting offsets, the higher traffic flow links (based on traffic monitoring) determine the offsets that provide good progression for that link. The goal is achieving better coordination for the links with high volumes to minimize the total number of stops in the system. Other links carrying lower flows may not receive optimum coordination.

The local controller may skip a phase if not demanded or may terminate early (before the expiry of the time allocated by network control decisions) under control of the gap timers. As in coordinated-actuated operation, one key phase, usually the arterial phase, cannot skip and cannot terminate early by action of gap timers. This maintains the common cycle time for coordination. Any time saved (slack) during the cycle as a result of other phases terminating their green early or being skipped may be used by subsequent phases or is added on the key phase to maintain cycle length. Note that the control logic with SCATS is similar to a standard coordinated-actuated operation in which extra green time (slack) can generally used by coordinated phases and non-coordinated phases use fixed green are forced to end their green even when there is high demand.

**Method for Optimization of Traffic Signals in Online-Controlled Network (MOTION)**

The goal of MOTION is to combine the advantages of well designed coordination plan with the flexibility of an immediate response of the local signals to the actual state of traffic. Using online (real-time) data, MOTION determines a common cycle length, number of phases and phase sequence (no fixed phase sequence), offset, and phase split.

MOTION makes the network decisions in the central computer, while local decisions take place in the local controller. The network control logic restricts the local control only to the extent that is necessary to guarantee a good coordination [21], as in coordinated-actuated operation with the difference that MOTION allows variable phase sequence.
Cycle length, phase sequence, and offset are determined on the network level, every 10 to 15 minutes. Cycle length is determined according to the traffic volumes, obtained in real time, at the critical intersection (i.e., the intersection that requires the highest cycle length). A common network cycle is calculated from the individual minimum cycle time of all intersections in the network. Coordination is optimized by using model of delay and number of stops. The optimization allows assigning individual link weightings to selected links (e.g., arterial links that carry higher traffic or links that are used by public transit).

MOTION is an adaptive traffic control strategy developed by Siemens, Germany.

**BALANCE**

BALANCE selects the program with the best cycle time for current traffic demand from a prearranged set of signal programs (based on stored programs). The selection is done at the same time for all traffic signals of one pre-defined control group. A control group is defined as a subset of traffic signals in the network (e.g., a group of signals in close proximity).

The optimization model of network control (using Genetic Algorithms in real-time) determines the length of splits and the offsets according to the selected, common cycle length [22]. Performance measures for optimization are delays, stops, and queue lengths for all intersection approaches. Queue lengths are used to determine timing parameters. They are estimated based on fill-up time, that is, the time needed to fill the area between the intersection stop line and the end of the detector located upstream (30-150 ft). BALANCE runs every five minutes and calculates optimized signal plans for the following five minutes period.

The optimization creates signal timing plans for every intersection in the network. Created timing plans based on the optimization determine, for every intersection, the earliest and latest points at which a phase could start. The interval between the two points is available to the
local traffic actuated controller in order to provide flexibility and responsiveness on the local control level, constrained by optimization decisions.

BALANCE was developed at Technische Universitat Muenchen and later also at the companies GEVAS software and TRANSVER. The system has been implemented in several German cities including Remscheid, Hamburg, and Ingolstadt.

2.4.1.1 Oversaturation Logic under ATCSs with Fixed Cycle Length

Oversaturation occurs when the average traffic demand exceeds the intersection capacity. During oversaturation, traffic queues develop and grow in length, which may lead to queue spillback to adjacent upstream intersections, causing excessive delays.

An oversaturated traffic environment requires different control policies than the policies developed for under-saturated environment. In the former, the control objective should be maximizing system capacity/throughput, while in the latter, the objective should be minimizing delay or number of stops, because capacity is adequate and queues can generally be stored within approaches and between intersections, avoiding queue spillback. As a result, an efficient coordination plan during under-saturated conditions may cause poor traffic performance when intersections are oversaturated. This section focuses on control logic of ATCSs operating under a fixed cycle length in addressing oversaturated conditions.

To handle oversaturation, all ATCSs first increase cycle length. In theory, longer cycles increase capacity by lowering the proportion of lost time in a cycle. In practice, however, longer cycles often fail to yield greater capacity, because longer red periods associated with long cycles allow for formation of longer queues, which increases potential for capacity-reducing queue interactions (e.g., spillback to an upstream intersection or spillback from a turning lane to a through lane).
To deal with oversaturation, ACS-Lite provides longer green times for arterial phases in order to limit oversaturation along an arterial and store queues on cross street. This is an important strategy during oversaturation, because giving higher priority to the movements having a greater saturation flow rate, due to having more lanes (e.g., arterial phases), reduces queuing delay. For example, time given to an approach (e.g., 15 seconds) with two lanes reduces delay more than the same amount of time given to a single lane approach.

SCOOT has several methods to handle oversaturated conditions. It uses gating or action at a distance, which allows the restriction of the green times for entry links to the congested area. This is one of the key control strategies during oversaturation, because it limits residual queues (i.e., queues from the previous cycle) and queue spillback, which in turn prevents capacity reductions. SCOOT also provides more green times at the exit links at the congested zone (i.e., at the bottleneck) to limit the extent of oversaturation along the arterial. SCOOT also allows users to specify a congestion importance factor (pre-defined) for each link to influence split calculations in favor of the link (with higher importance) when congestion is detected.

2.4.1.2 Transit Signal Priority (TSP) Logic under ATCSs with Fixed Cycle Length

TSP is an important component of efficient traffic signal control with the potential of increasing transit ridership by reducing transit travel time and improving service reliability. The following table shows the TSP logic of ATCSs considered in this section.
Table 2-2: Transit Signal Priority Logic under ATCS with Fixed Cycle Length

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th>Transit Signal Priority (TSP) Logic</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACS LITE</td>
<td>No logic is provided to accommodate TSP.</td>
</tr>
<tr>
<td>LA ATCS</td>
<td>Provides green extension and early green for late buses. TSP is granted at local intersection level.</td>
</tr>
<tr>
<td>SCOOT</td>
<td>The detection of a bus gives a higher priority to the bus phase in making split decisions, though there is no guarantee of TSP.</td>
</tr>
<tr>
<td>SCATS</td>
<td>No logic is provided to accommodate TSP.</td>
</tr>
<tr>
<td>MOTION</td>
<td>When expected TSP disruption is significant (e.g., one priority request per two cycles), TSP is taken into account in the offset determination based on expected transit arrival time. When priority requests are not too frequent to disrupt progression, phase splits and sequence are modified to accommodate TSP.</td>
</tr>
<tr>
<td>BALANCE</td>
<td>TSP is provided locally, at intersection level, through modification of phase splits and sequence.</td>
</tr>
<tr>
<td>VFC-OPAC</td>
<td>No logic is provided to accommodate TSP.</td>
</tr>
</tbody>
</table>

Four ATCSs described in this section provide some type of priority for transit vehicles. The priority is either provided at the local controller’s level (intersection level) or it is integrated into the incremental adjustment structure of the signal timing plan.

LA ATCS and BALANCE provide priority at local level by adjusting the timing plans in response to bus detection. LA ATCS grants priority to late buses by extending bus green phases and shortening the green time of preceding phases (therefore shortening bus red phase’s red time). BALANCE not only changes phase splits when a bus is detected, but it can also modify phase sequence (known as phase rotation) to reduce bus delay.

MOTION has two main levels of priority that can be pre-selected by the user including absolute priority and relative priority. Absolute priority indicates priority to transit independent of the impact to private traffic, and is given when the expected TSP disruption is not significant (e.g., less than one priority request per two cycles). When the expected impact of TSP to non-private traffic is large, relative priority is given to public transport.
Under relative priority, there is no “active priority” (i.e., changing signals in response to real-time detection), but only “passive priority”. TSP is taken into account in the offset determination (network control logic), limiting the range of options for optimizing phase sequence, splits, and offset to those which provide good progression for public transit.

SCOOT weakly accommodates bus priority within its optimization procedure. With each split decision, SCOOT takes into account the percent saturation of all approaches. The detection of a bus merely gives a higher priority to that phase, increasing its green time slightly. Such an extended green may be completely unnecessary (the bus may have passed already), or may be insufficient to let the bus clear the intersection. This is in contrast to active priority logic that tries to make the signal green when the bus needs it.

In conclusion, only three of the seven ATCSs in Table 2.1 have active TSP tactics. One of them, LA ATCS, keeps overall coordination by limiting length of green extension and early green. The two German systems, that is, BALANCE and MOTION, allow more aggressive, disruptive priority interruption, letting coordination disrupt and allowing increased delay to private traffic.

### 2.4.2 ATCSs Free of Any Cycle Length

This section discusses control strategies of the most widely used cycle free ATCSs and highlights their limitations. ATCSs that are free of any cycle length distinguish themselves from traditional fixed, common cycle length strategies by dropping the concept of cycle. The cycle length varies from cycle-to-cycle, and from intersection to intersection, depending on traffic demand. There is no explicit coordination, that is, there are no clock-based offsets and no fixed, common cycle lengths. Rather, the coordination is typically provided implicitly and dynamically as adjacent intersections communicate with each other.
Table 2-3 gives a brief description of ATCSs that are free of any cycle length and their operational principles. Note that the oversaturation comparison is not provided in this section, because typically ATCSs that do not follow a fixed, common cycle length address oversaturation inherently by adjusting the green times in order to spread the queues and bound them within the available storage capacity. Therefore, only TSP logic is discussed.

Table 2-3: Control Logic of ATCSs that are Free of any Cycle Length

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th>Control Logic</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPAC-3</td>
<td>Predicts future arrivals over the next one to two minutes period (defined as horizon). Updates prediction every five seconds. Allows variable phase sequence. Does not allow gap-out logic at intersection level. Though OPAC-3 was designed for arterial-network control, it had to include a common cycle length and fixed phase sequence logic for coordination, which led to VFC-OPAC.</td>
</tr>
<tr>
<td>RHODES</td>
<td>For network control, projects arrivals over the next 200-300 seconds. Peer-to-peer communication and data exchange between intersections is used to provide “implicit” coordination. For intersection control, predicts arrivals over the next 45-60 seconds. Updates prediction every three seconds. Network control decisions are used as constraints to intersection control logic. Uses variable phase sequence. Does not allow gap-out logic at intersection level.</td>
</tr>
<tr>
<td>UTOPIA/SPOT</td>
<td>Predicts arrivals over the next 120 seconds with prediction updates in every three seconds. Uses peer-to-peer communication to build dynamic coordination. Allows variable phase sequence and gap-out logic at intersection level.</td>
</tr>
<tr>
<td>INSYNC</td>
<td>Uses peer-to-peer communication between adjacent intersections for progression. Each green signal is successively turned based on the expected arrival time of vehicles from upstream intersection. Critical arterial direction is required to be defined by the user. Local control logic is constrained by network decisions. Allows variable phase sequence and standard gap-out logic at intersection level.</td>
</tr>
<tr>
<td>SOTL(^1)</td>
<td>Uses only local rules (each controller makes its own decisions independently) regardless of the state of adjacent intersections. Applies fundamentals of actuated control (skipping, gap-out logic) with added spillback rules. Uses fixed variable sequence. Does not require future arrival prediction.</td>
</tr>
</tbody>
</table>

Note:
1. SOTL refers to the self-organizing logic developed by Gershenson [7].

The Optimized Policies for Adaptive Control (OPAC)

OPAC is a real time, demand responsive traffic signal timing optimization algorithm which was developed by Nathan Gartner at the University of Massachusetts, Lowell.
OPAC is a distributed control strategy (i.e., intersection controller is responsible for control at the intersection rather than a centralized system in which a central computer makes control decisions of individual intersections) featuring a dynamic optimization algorithm to minimize a performance function of total intersection delay and stops. The original versions of OPAC (OPAC-1, OPAC-2, and OPAC-3) were free of any cycle length and phase sequence, however the latest version of OPAC (VFC-OPAC), which is designed primarily for arterial and network control, follows cyclic operation with a common cycle length and fixed phase sequence. [23] [24] [25].

OPAC-1, the first generation of OPAC, requires complete knowledge of arrivals over the entire control period and uses dynamic programming, ensuring a global optimum solution. However, this approach cannot be used for real-time implementation due to the amount of computational complexity involved as well as the lack of available real-time information (due to the inability to predict the complete knowledge of arrivals over the control period with certain precision).

OPAC-2 is a simplification of OPAC-1. The control period was divided into projection horizon (one-two minutes time periods, comparable to a cycle length for a pre-timed traffic signal, the period over which traffic patterns are projected). The number of signal phase changes is limited to be at least one and no more than three at each horizon. For every switching sequence, a delay function is defined for each approach based on the projected arrivals. The control problem for each horizon then becomes finding the optimal switching sequence which minimizes total vehicular delay. However, this method still requires knowledge of arrivals over the whole horizon, which is difficult to obtain with accuracy in practice.
OPAC-3, the further simplification of OPAC-2, uses a rolling horizon approach to make use of flow data that are readily available without degrading the performance of the optimization. With a rolling horizon, the roll period consists of k stages. The projection horizon is typically equal to an average cycle length for a pre-timed traffic signal (one-two minutes) with stages of four or five seconds. Actual arrival data for k stages can be obtained for the beginning, or head, portion of the horizon from detectors placed upstream of each approach (Figure 2-5). For the remaining n-k stages, the tail of the horizon, flow data may be obtained from a model. An optimal switching policy is calculated for the entire horizon, but only those changes which occur within the head portion (0 - k) are actually implemented. In this way, the algorithm can dynamically revise the switching decisions as more recent real-time data become available.

Figure 2-5: Rolling Horizon Approach

OPAC-4 is a network version of OPAC. Earlier versions of OPAC were developed as a cycle-free controller. No explicit coordination features were embedded; however, the algorithm has inherent self-coordination capabilities because of the tail model in the projection horizon, which transmits information between adjacent intersections (implicit coordination). As part of FHWA’s Real Time Traffic-Adaptive Signal Control System (RT-TRACS) project, the control logic was altered to include an explicit coordination and synchronization strategy that is suitable
for implementation in arterials and in networks. This version is referred to as virtual fixed-cycle OPAC (VFC-OPAC, as given before in ), because from cycle-to-cycle the yield point, or local cycle reference point, is allowed to range about the fixed yield points dictated by the virtual cycle length bounds and offsets [26]. This allows the synchronization phases to terminate early or extend later to better manage dynamic traffic conditions. VFC-OPAC consists of three layer control architecture outlined as follows:

- **Layer 1**, the local control layer, implements OPAC 3 rolling horizon procedure. It continuously calculates optimal switching sequences for the projection horizon subject to VFC constraint communicated from Layer 3.
- **Layer 2**, the coordination layer, optimizes the offset at each intersection (once per cycle).
- **Layer 3**, the synchronization layer, calculates the network-wide VFC (once every few minutes, as specified by the user). The VFC is calculated in a way that provides sufficient capacity at the most heavily loaded intersections while maintaining suitable progression opportunities among adjacent intersections. The VFC can be calculated separately for groups of intersections as desired. Over time, the flexible cycle length and offsets are updated as the system adapts to changing traffic conditions.

The adaptive control logic of VFC-OPAC is described as follows:

- Performance function used by the optimizer is a weighted function of total intersection delay and stops. The weights are configurable to eliminate delays or stops, or set their relative importance.
- The central system optimizes the cycle length for each section or group of intersections. A “critical intersection” is determined periodically and the cycle length calculated based on data from the critical intersection. The goal of cycle length optimization is to meet
phase switching timing determined by local conditions, while maintaining a capability for coordination with adjacent intersections. However, the cycle length can start or terminate only within a prescribed range. As a result, all VFC-OPAC controlled intersections can oscillate with a common frequency.

- Split optimization is achieved for up to eight phases in a dual ring configuration. Minor phases, for example, can be left to the default control (local control), whereas only major phases are optimized. Phases with no detectors can also be left out of the optimization, because it would be based on unreliable estimates of demand.

- Where there is no cycle or offset constrains, OPAC may operate under free mode. Split optimization is constrained only by phase-specific minimum and maximum green.

Even though OPAC 3 has inherent self-coordination capabilities, it is important to point out that the early versions of OPAC had to include a common cycle length and fixed phase sequence logic to make the controllers suitable for implementation in arterials and networks. VFC-OPAC offers more flexibility compared to a standard actuated-coordinated scheme as it can update coordination parameters incrementally to adapt to current traffic conditions, however it still has the limitations of fixed cycle coordination.

**Real-Time Hierarchical Optimized Distributed Effective System (RHODES)**

RHODES is a real-time traffic adaptive signal control strategy developed at the University of Arizona.

The RHODES system takes input from detectors, predicts future traffic streams, and outputs optimal signal control settings that respond to these predictions. The optimization criteria can be any provided by the jurisdiction using the system, but must be based on traffic measures of effectiveness (e.g., average delays, stops, throughputs) [27].
The RHODES hierarchical architecture decomposes the traffic control problem into several subproblems that are interconnected in a hierarchical fashion. The subproblems include network control problem and intersection control problem.

The network flow control logic is based on an approach called “REALBAND” [28]. The REALBAND approach first identifies platoons based on real-time information and predicts their movement in the network (i.e., their arrival times at intersections, their sizes, and their speeds). The signals are set so that the predicted platoons are provided appropriate green times to optimize a given performance criterion.

Two platoons demanding conflicting movements may arrive at an intersection at the same time. In that case, one will be given priority on the green time, or one of the platoons will be split to maximize the given measure of performance. Optimally resolving such conflicts in real-time is the main objective of the REALBAND approach.

The described coordination mechanism (network control logic) under RHODES is not explicit coordination, that is, there are no clock-based offsets and no fixed cycle lengths. The cycle length is allowed to vary from cycle-to-cycle. The peer-to-peer communication and data exchange between intersections is able to provide “implicit” coordination that varies based on demand within the network. However, it is worth noting that this type of coordination is also different and more complicated than the coordination under self-organizing logic, which simply relies on local intersection rules and short-term predictions. Unlike self-organizing logic, RHODES requires the prediction of movement of platoons (including prediction of turning ratios at intersections and platoon dispersion) over a longer time periods (projection of 200-300 seconds), which may be difficult to obtain with accuracy.
At the lowest level of the RHODES hierarchy for a surface street network, that is, at the intersection control level, RHODES uses a dynamic-programming (DP) based algorithm, Controlled Optimization of Phases (COP) [29]. Using DP in a rolling horizon framework (45-60 seconds of horizon), RHODES determines the optimal phase sequence and green times that will minimize the delay incurred by vehicles passing through the intersection. The REALBAND decisions for network control problem are used as constraints to the intersection control logic (COP) that specify the winning phase from the outcome of each conflict resolution on the optimal root-to-leaf path in the decision tree. RHODES does not allow the local controller to execute gap-out decisions. As a result, if a green phase is not used by vehicles, the local controller cannot terminate the green phase until the next optimization is performed, resulting in wasted green time.

The success of RHODES (similar to earlier versions of OPAC) relies on an accurate prediction of traffic arrivals over the entire decision horizon, which is difficult in practice. This undermines the efficiency of RHODES and limits its implementation.

**Urban Traffic Optimization by Integrated Automation (UTOPIA) / System for Priority and Optimization of Traffic (SPOT)**

The system uses a two level, hierarchical and decentralized control strategy, involving intelligent local controllers to communicate with other signal controllers as well as with a central computer. UTOPIA/SPOT does not have a fixed cycle length. Coordination logic takes into account the state of adjacent intersections (through communication of upstream and downstream intersections) to build dynamic signal coordination [30].

The higher level in control hierarchy is responsible for setting the network control strategies, whereas the lower level (i.e., SPOT), implements signal timings according to the actual local traffic conditions constrained by the network control strategy from the higher level.
SPOT operates by performing a minimization of local factors such as delays, stops, excess capacities of links, stops by public or special vehicles, and pedestrian waiting times on the time horizon of the next 120 seconds and is repeated every three seconds. With each repetition, all SPOT units exchange information on the traffic state with their neighboring SPOT units. This permits the application of a “look-ahead” principle that accounts for the traffic forecast (realistic arrival predictions from upstream intersections) and “strong interaction” that accounts for the expected delay at the downstream intersections experienced by vehicles leaving the intersection under consideration.

At the area level (i.e., central level), the optimal network traffic control problem is developed based on the macroscopic traffic model of the network (representing the behavior of traffic over the whole controlled area), and control strategies such as minimum and maximum length of each stage, offsets, and weights for all the elements constituting the objective function which are optimized locally, are defined for each intersection. The network control strategy is optimized over the next 30 to 60 minutes time horizon (depending on the size of the network controlled) and is updated every five minutes. Each strategy is effective for not more than five minutes, and is then replaced by a new strategy.

In case of oversaturation, UTOPIA, on the network level control, requests that a downstream signal increase throughput or that an upstream controller decrease demand (similar to gating strategy applied by SCOOT). These requests are realized respectively by relaxing or tightening green time constraints.

UTOPIA/SPOT (Urban Traffic Optimization by Integrated Automation/System for Priority and Optimization of Traffic) is a traffic signal control strategy developed by Mizar Automaziona in Turin, Italy.
**INSYNC**

InSync is an adaptive traffic signal system developed by Rhythm Engineering (Lenexa, Kansas). The system operates on two levels; the global level (for network control) and local level (for intersection control).

On the global level, traffic signals under InSync are linked together with wireless broadband radio communications, allowing them to “talk” with each other and to provide progression along an arterial [31]. The coordinated directions are required to be determined by the user.

For coordination, the facilitator intersection (i.e., the critical intersection or the intersection that requires the longest cycle length) decides a time at which it will serve a green band for the coordinated phases and communicates that time with the adjacent intersections. Green bands, called “time tunnels”, are then guaranteed by successively turning each signal green at the projected arrival time of vehicles from upstream intersections (Figure 2-6).
Figure 2-6: Concept of Time Tunnels in InSync [6]

Intersection F, the middle of five intersections, is the facilitator intersection. The lines represent the required start of green for each direction of travel. Start times are adjusted by (-travel time) so that vehicles are released from the upstream intersection in time to reach the facilitator intersection when it initiates its tunnel time. Start times for downstream tunnel phases from the facilitator at the adjacent intersection are adjusted by (+travel time) so that vehicles arrive from the facilitator intersection when the signal is green. As a result, InSync does not require a fixed cycle length and offsets to a fixed point in a cycle to create coordination.
Time tunnels can be expanded or contracted by the facilitator to provide efficient green waves for each phase along the arterial. At the end of every period, each local processor is “polled” by the facilitator and “reports in” if it needs more time, the same time, or less time.

InSync automatically extends green signals beyond the set parameter if it observes that the moving platoon has not sufficiently gapped out at a user changeable percentage of occupancy (calculated every second by InSync) or set gap time.

On the local level, beyond the constraints communicated by the facilitator as tunnel messages that guarantee coordination for main arterials, local intersection control operates in a standard actuated mode. If the period between time tunnels is 90 s in duration and a green signal is guaranteed for 10 s for the main direction at each intersection, then 80 s are available for the local optimizer to schedule states at each intersection according to its intelligent scheduling, which is described as follows:

1. Scheduling of states is performed based on closeness to the initiation of a new tunnel. To illustrate, if the controller is close to the initiation of a new tunnel, it will schedule a main street sequence of states.

2. If a tunnel has recently ended, it will schedule a cross street sequence as its priority. If there are no cross street queues, it will schedule a miscellaneous main state.

3. A miscellaneous main state is scheduled for phases with queues that have been waiting the longest. Wait times being equal, the phase with the largest queue is scheduled.

InSync also allows arbitrary phase sequence. A particular phase can be served more than once or not at all during a period if demand warrants. For each sequence, a minimum total time solution satisfying minimum green restrictions, determined by either queue service time or pedestrian
phase, and the total delay for queues are calculated. The sequence with the least total delay is chosen. If no solution satisfying the restrictions exists for any permitted sequence, then the lowest priority of the restrictions are relaxed until a solution is possible.

In case of limited queue storage at the downstream link, controller prevents early release of a platoon to inhibit spillback. The optimizer has the configuration option of restricting early release of a tunnel phase at an intersection.

**Self-Organizing Traffic Signals (SOTL)**

A system described as self-organizing is one in which elements interact in order to achieve dynamically a global function or behavior. Self organization is a process whereby pattern at the global of a system emerges solely from interactions among the lower level components of the system.

Gershenson et al. applied the concept of self organization to traffic control problem [7] [8] [32] [33]. SOTL does not need a central computer or a hierarchical framework. No optimization is performed to reach the global solution. Only local rules (i.e., each local controller makes its own decision independently) are implemented regardless of the state of adjacent intersections. Yet, the system is able to achieve global coordination of traffic.

Gershenson et al. [7] tested self-organizing logic in an abstract traffic grid network with one-way streets and two phase operation per cycle. The self-organizing rules, where high numbered rules have higher priority; used in the model can be described as follows:

1. An upstream detector counts vehicles approaching an intersection and switches the signal to green when the counter exceeds a pre-specified threshold (subject to override by subsequent rules). This rule helps platoons form and gives green signals quicker to the
movements having dense platoons. It also prevents long waiting times when the other
rules have failed to force a switch, while allowing longer than minimum greens.

2. Signals must stay green for a minimum time. This prevents platoons approaching the
same intersection from conflicting directions from triggering repeated switching that
would immobilize traffic.

3. If a few vehicles (m or fewer, but more than zero) are approaching to cross a green signal
at a short distance \( r \), do not switch the signal. This rule prevents the “tails” of platoons
from being cut, but allows the division of long platoons to promote the integrity of
platoons.

4. If no vehicle is approaching within a certain pre-determined distance, and at least one
vehicle on a competing movement is approaching, then the signal is switched. This is
similar to gap-out logic, which terminates green when no flow is detected.

5. If there is a stopped vehicle a short distance beyond the intersection (i.e., just downstream
of an intersection), then the signal is switched to red at the upstream intersection to
prevent vehicles joining the queue and blocking intersection, which causes intersection
gridlock.

6. If there are vehicles stopped in both directions at a short distance beyond the intersection,
then both signals switch to red. Once one of the directions is no more blocked, controller
restores the green signal in that direction.

With the described traffic signal rules, self-organizing control applies fundamentals of actuated
control, adjusting cycle length and splits to the actual demand. For low densities, Rule 1
promotes formation of platoons, which help improve coordination. For high densities, Rules 5
and 6 prevent gridlock and promote formation of free spaces that can be used by competing platoons.

The self-organizing method was tested on a ten-by-ten homogenous grid network and the results of SOTL were compared to fixed time operation. The model provided almost zero delay for low to intermediate traffic flows, and less delay than fixed time operation when traffic flow gets higher by eliminating queue spillback.

However, the success of the self-organizing model was partly due to the simple network used: a grid of one-way streets without turning traffic, having regular intersection spacing and equal traffic demand. One-way traffic is easier to coordinate than two way streets. Moreover, equal demand means all intersections have the same “natural” cycle length, which makes signal coordination fairly easy. Finally, turning traffic reduces the density of platoons, which undermines the efficiency of coordination as less number of vehicles benefit from green waves.

2.4.2.1 Transit Signal Priority (TSP) Logic under ATCSs that are Free of Any Cycle Length

TSP reduces delays (therefore passenger riding time) for transit vehicles at signalized intersections by extending the green signal or shortening the red signal. It improves schedule adherence, which in turn reduces passenger waiting time, and ultimately promoting transit ridership and reducing congestion. This section summarizes TSP strategies that are considered under ATCSs with free of any cycle length.

RHODES and UTOPIA/SPOT are the only two systems that consider TSP explicitly in their control logic. RHODES gives each detected bus a variable weight that depends on the number of passengers it has and on how late is the bus, if it is behind the schedule [34]. With the described TSP logic, RHODES tends to give higher priority for late buses with high volume. If
passenger counts are not available, the system uses an algorithm that estimates passenger counts (based on historical ridership data). Needless to say, RHODES’s TSP logic improves the operations of public transit. However, its efficiency is rather limited, because the success of TSP logic also relies on the accurate prediction of traffic arrivals over the decision horizon as well as the predictions of bus occupancy, which are difficult to obtain in reality.

UTOPIA/SPOT is able to assign absolute, weighted, and selected priority to buses and trams at signalized intersections. To determine the level of priority, weights are assigned to specific vehicles (e.g., according to the line, direction, or vehicle adherence to the schedule) locally or centrally. TSP is achieved within the intersection optimization process. Because intersection optimization is carried out every three seconds, the system can react quickly to any changes in the predicted transit arrival time.

While performing intersection optimization, transit vehicles are considered in the same way as private vehicles, that is, they are represented by equivalent “vehicle platoons”, which appear as probability curves centered on the predicted arrival times. The prediction variation decreases as transit vehicles approach the intersection. Communication of adjacent intersections (used for coordination) is also important in giving effective TSP, because it allows longer advanced detection (e.g., not limited to intersection spacing in case of far-side stop) so that traffic signals can be adjusted by small increments to serve the approaching transit vehicle. In this way, TSP logic ensures that the traffic signal phases are managed in a way that limits the impact on non-transit vehicles.

To conclude, it is surprising that none of these cycle-free ATCSs take advantage of their flexibility to provide active priority, targeting green time to exactly the same time that a bus is expected to arrive at the stopline.
2.5 Summary and Conclusions

This chapter reviewed traffic signal control logic for arterials, including pre-timed control, coordinated-actuated control, and adaptive control. The primary focus of the review was adaptive signal control, which was analyzed in two parts: adaptive controllers that follow a fixed, common cycle length, and adaptive controllers that are free of any cycle length.

Arterial control focuses on signal coordination to make sure that vehicles can get through multiple intersections without having to make stops. Coordination on a one-way street is relatively simple. However, coordination along two-way arterials with irregularly or closely spaced intersections is difficult to achieve. The latter generally results in good coordination for one arterial direction and poor coordination for the other one. Moreover, because coordination requires a common, fixed cycle length, and fixed offsets, traditional coordinated control (including actuated-coordinated control) cannot adapt to variations in traffic demand.

Adaptive traffic controls were designed to overcome the fluctuations in traffic demand. ATCSs with fixed cycle length (e.g., ACS Lite, SCOOT, SCATS) still employ fixed offsets and phase splits, yet they incrementally update these parameters based on real time or projected traffic data, which adds more flexibility compared to pre-timed or coordinated-actuated control. The update frequency varies from every cycle to 10-15 minutes depending on the deployed ATCS. Their two main limitations are:

- Small transitions limit their demand responsiveness
- By the time the controller starts transitioning into a new plan, the traffic plan which initiated this selection may have left the system

Adaptive controllers that are free of any cycle length (e.g., RHODES, UTOPIA/SPOT) drop the concept of a fixed, common cycle and offset. They typically predict future traffic arrivals and
make control decisions in response to these predictions. Coordination along an arterial is generally accomplished through communication of adjacent intersections.

ATCSs that are free of a cycle length brings significant flexibility by allowing variable cycle lengths and phase sequence, which help respond to traffic fluctuations. However, they still have the following challenges:

- The computational burden involved in exploring the search space makes some systems not suitable for real-time implementations
- Their method relies on prediction of future arrivals over the next one-two minutes, which is difficult to obtain in practice

Self-organizing control methods developed by Gershenson differs from the abovementioned cycle free adaptive controllers. It uses standard actuation logic with simple rules added (rules that do not require prediction of traffic) to develop stronger coordination mechanisms. Their results showed significant improvement compared to traditional pre-timed coordinated control.

However, self-organizing logic was tested on a very simple abstract traffic grid network, with one-way streets (two-phase operation), regular intersection spacing, and equal traffic demand. These three network characteristics are well-suited for the described self-organizing rules. Therefore, the challenge still remains to develop signal control algorithms that add communication and coordination mechanisms, making local, rule-based control self-organizing on the arterial level with realistic networks and traffic patterns.
Chapter 3: Sources of Inefficiency in an Actuated Control at Local Level

The proposed self-organizing control logic in this study builds on actuated intersection control. In order to develop efficient self-organizing control logic, it is crucial to have a successful framework of actuated control.

The design of an actuated intersection including detector type, location, and traffic controller settings can impact the performance of actuated operation. The lost time (wasted green time or any green time that is not used at saturation flow) that occurs from not detecting and switching efficiently can substantially increase cycle length and delay, resulting in suboptimal operation [35].

This chapter provides a systematic review of actuated control to identify sources of inefficiency in actuated controllers, and proposes detection and controller settings to achieve a maximum efficiency of local controller for the self-organizing control logic. The review focuses on physical components of an actuated signal (e.g., detection) and some of the control logic used in an actuated operation.

3.1 Stop-line vs. Upstream Detector

At an actuated intersection, controllers typically use two types of detector placements to call and extend a phase. The first type relies on only stop-line detection both to register a vehicle call (i.e., for start of green) and to extend a green phase (for end of green) (Figure 3-1a). The second type employs stop-line detection for vehicle calls and upstream detection for green extension (Figure 3-1b).
a) Stop-line Detection Only    

b) Stop-line and Upstream Detection

Figure 3-1: Detector Placement in an Actuated Control

When a stop-line detector is used to extend a green phase, the green time while waiting to detect the critical gap is wasted. When the detector to extend a green phase is moved upstream such that the travel time from an upstream detector to the stop-line is roughly equal to the critical gap, the last approaching vehicle extending the green phase will reach the stop-line just as the yellow clearance period is started, avoiding wasted green time. As a result, upstream detection to extend a green phase results in lower cycle length and lower delay [36].

3.2 Short Critical Gap

One important control setting for efficient actuated operation is the critical gap. Short critical gaps result in “snappy” operation at an actuated intersection, reducing wasted green time and cycle length. In traffic flow theory, gap time is related to time headway (Figure 3-2), and can be translated using equation (3.1).
where \( h \) is time headway, \( L_d \) is detector length, \( L_v \) is vehicle length, and \( v \) is approach speed in the absence of queues. The second component of summation is also known as detector occupancy time. Therefore, headway is equal to “gap plus detector occupancy time”. The headway corresponding to the critical gap is called “critical headway”.

The critical gap is set to a value (accounted for occupancy time) long enough to limit the chance of termination of a green phase during queue clearance. Termination of a green phase while a queue is clearing can also be defined as “Type I error” or “false negative”. Type I error occurs when a controller concludes that saturation flow has ended when in fact vehicles are still discharging at saturation flow. In our analyses, the critical headway was set to 99.5% percentile headway during saturation flow, which corresponds to a 0.5% of Type I error.

With a longer critical gap or critical headway, an actuated control may result in long wasted green time, leading to longer cycle length and higher delay. Figure 3-3 shows how sensitive average cycle length is to detector setback and critical headway. Note that cycle length

\[
h = \text{Gap} + \frac{L_d + L_v}{v}
\]  

(3.1)
results were calculated theoretically using the formulas described in Furth et al. [35], and validated through simulation. For a given level of demand (i.e., flow ratios (∑v/s) summed over critical lane groups), using an upstream detector instead of a stop-line detector reduces average cycle length by approximately 20 seconds, and making critical headway one second shorter reduces average cycle length by approximately 12 seconds. This substantial reduction in average cycle length also reduces red time, and thus average delay.

Figure 3-3: Expected Cycle Length (sec) as a Function of Demand, Critical Headway H (sec), and Detector Setback S (ft) [35]

The actuated operation that forms the basis of the proposed self-organizing control logic uses upstream detection with a setback distance (in seconds) equal to the critical gap. For example, if the critical gap is 3.5 seconds and approach speed is 30 mph (44.1 ft/sec), extension detectors are located approximately 150 ft upstream of the stop line.
### 3.3 Short and Variable Minimum Green Interval

The minimum green interval is another controller setting that can impact the performance of an actuated intersection. It is defined as the shortest time that may be provided to a green phase. With actuated operation, the controller does not monitor the gap time until the minimum green time has expired. Therefore, gap-out cannot occur until after the minimum green time is over.

A minimum green interval is established to enable the first vehicle in the queue to clear an intersection using the sum of minimum green time and the change interval. When slow vehicles are considered (e.g., heavy trucks or bicycles), five to eight seconds of minimum green interval is typically enough to allow the first vehicle to clear an intersection.

Minimum green interval also allows the “intermediate” queue to clear, that is, the queue stored between a stop line and an extension detector (Figure 3-4). When a standing queue at the beginning of a green phase does not reach an extension detector located upstream, standard gap-out/extension logic cannot hold a phase green. In that case, the minimum green time should be long enough so that vehicles stored between the stop-line and the extension detector can clear the intersection.

![Figure 3-4: Intermediate Queue When an Upstream Detector Is Used for Extension](image)

However, in cycles where there are very few (e.g., one or two vehicles) arrivals, long minimum green time extends green unnecessarily, causing longer cycle lengths and reducing
intersection’s efficiency. Another solution to address the intermediate queue problem without dropping the efficiency of actuated operation is to use variable minimum green interval, where the initial minimum green time is increased by a certain amount by each vehicle actuation during the associated phase yellow and red intervals. The detailed explanation of variable minimum green interval is provided in Chapter 4.

### 3.4 Simultaneous vs. Non-simultaneous Gap-Out

Actuated controllers offer two options to ensure that the two phases that end each half-cycle (at the barrier, e.g., switching from East-West to North-South movements) terminate their green together. One is “simultaneous gap-out”, a default setting in American controllers, which requires that both phases must be gapped-out at the same moment to force an end to extension green. The other option, “non-simultaneous gap-out”, allows one phase to gap-out and “wait” for the other phase to gap-out, and then both phases terminate their green.

In general, non-simultaneous gap-out allows phases to end green sooner, and is therefore more efficient. Simultaneous gap-out is over-conservative, requiring gap-out of combined traffic volumes of both phases [35]. When both phases are combined, the probability of getting the critical gap (or critical headway) becomes smaller, resulting in longer phases and therefore longer cycles. Green time is wasted as the critical direction phase waits for a gap in the non-critical phases.

In order to limit the wasted green time, actuated controllers in this research use non-simultaneous gap-out logic for the two phases ending green at the barrier.
3.5 Gap-out on Multi-lane Approaches

The efficiency of actuated traffic signal controllers relies on quickly detecting when the standing queue has discharged and the flow rate has dropped below saturation flow in order to minimize wasted green time by assigning right-of-way to a conflicting phase with a standing queue. While a controller’s need is to detect a change in flow rate, what they actually measure is gaps and, implicitly, headways. The effectiveness of gap-out logic depends, therefore, on how closely related headways are to flow rate. On average, they are directly related, as average headway is simply the inverse of flow rate. However, in actuated control, controllers make decisions based on single headways, not average headways, and for traffic moving with a given flow rate, headways vary from vehicle to vehicle. How well a controller can detect a change in flow rate depends on how variable headways are within a given flow regime.

3.5.1 The Impact of Headway Variability on Actuated Signal Control Performance

The critical gap under actuated controllers must be set above average headway (adjusted for detector occupancy time) in order to avoid “premature” gap-out (i.e., early termination of green). If the variation in saturation headway is high, then a more generous critical gap (or headway) value must be defined as the threshold. This generous critical gap that is aimed at preventing premature gap-out may be so long that gaps smaller than the critical gap can also occur frequently during unsaturated flow, making it difficult for the controller to detect the change to unsaturated flow. The controller may continue to hold the green for several headways after queue discharge is complete, resulting in wasted green, which in turn increases average cycle length and delay.
3.5.1.1 Type I and Type II Errors

There are two types of detection errors that may occur while the controller is looking for the critical gap. As described in Section 3.2, the first type of error is Type I error, in which the controller concludes that saturation flow has ended when in fact vehicles are still discharging at saturation flow (premature gap-out). The second type of error is to conclude saturation flow is still continuing when in fact flow rate is lower (extending a green phase when flow is below saturation flow), known as “Type II error” or “false positive”. Premature gap-out leaves a portion of the stopped queue unserved and, thus, leads to increased delays and possible queue spillback. Moreover, it leads to driver frustration. Therefore, Type II errors are much preferred over Type I errors and the critical headway is set for low probability of Type I error. The following figure shows distribution of headways during saturated and unsaturated flow regime and describes Type I and Type II errors.

Consider a single lane approach. The top curve shows headway distribution during saturation flow, with mean headway is equal to two seconds (Figure 3-5). There is hardly any headway lower than one second and higher than three seconds, which leads to low headway variability during saturated flow regime. The bottom curve displays headway distribution when flow regime is unsaturated. It describes a case in which flow is approximately equal to 55% of saturation flow, and the mean headway is 3.7 seconds. Note the high variability during unsaturated flow, where headway values range from one to 5.5 seconds.

The critical headway is set for low probability of Type I error (premature gap-out). During unsaturated flow regime (bottom curve), headways that are smaller than the critical headway extend green (even though flow rate is below saturation flow), resulting in wasted green time. This is Type II error.
\( \mu_s \) = mean headway during saturation flow \( \mu_u \) = mean headway during unsaturated flow

**Figure 3-5: Low Headway Variability during Saturated Flow Regime**

**Figure 3-6** shows headway distribution curves for a two-lane approach. Headway distributions during both saturated and unsaturated flow regime have greater variability (relative to the mean headway) compared to a single-lane approach (the reasons for high headway variability for a two-lane approach are discussed in the following sections). Greater variability in saturation flow distribution forces the controller to have a longer critical headway (further to the
right on the curve). That, combined with greater variability in unsaturated headway distribution, results in higher probability of Type II error.

**Figure 3-6: High Headway Variability during Saturated Flow Regime**

3.5.2 Traditional Gap-out Logic on a Single Lane vs. Multi-Lane Approach

On a single lane approach, headways during saturation flow have relatively low variability, because each driver has to keep a safe following distance from the preceding vehicle, making very small headways (e.g., headways smaller than one second) physically impossible. This low variability in headways makes individual gap time a good indicator of flow rate, and so
traditional phase termination logic works efficiently, leaving only a small possibility of keeping the phase green long (Type II error) after flow rate has declined below saturation flow.

However on multilane approaches, traditional vehicle detection uses detectors in all individual lanes on a particular approach providing a single input into the same signal phase [38]. For example, if there are three lanes on an approach, there will be three detectors, one in each lane, and all three detectors will be wired together in series. The controller sees the combined gaps and headways across all lanes and determines gap-out decision based on the combined headway. As a result, headways which are zero or close to zero become not only possible but common.

For a given average headway, if there are very small headways, there must also be very large headways, resulting in high headway variability. Very small headways are physically impossible on single-lane approaches due to longitudinal spacing requirements for safety, but they are common on multi-lane approaches. Therefore, multi-lane approaches have greater variability both for saturated and unsaturated flow regimes. As a result, for multi-lane approaches, in order to be reasonably certain that early termination of a green phase does not occur (Type I error), as explained earlier, a generous critical gap value must be defined as the threshold. Then when the queue is discharged and the flow rate drops, the controller can obtain several headways smaller than the critical headway, extending green phase significantly beyond saturation, leading to longer cycle length.

**Figure 3-7** illustrates how traditional vehicle detection results in inefficient operation on multi-lane approaches. Case (a) represents saturation flow, with vehicles arriving in ranks. Case (b) represents traffic with a flow rate equal to one-third of the saturation flow rate. With traditional detection, because the controller sees the combined gaps and headways, those two
arrival patterns are perceived the same by signal controller. As a result, the controller will extend the green even if the flow is only one third of saturation flow rate, resulting in less efficient operation and higher delay.

![Image of two extreme arrival patterns](image_url)

Hₜ: Saturation Headway within a Single Lane

**Figure 3-7: Two Extreme Arrival Patterns that are Identically Perceived by Signal Controller under Traditional Detection Scheme**

### 3.5.3 Lane-By-Lane Detection for Multi-Lane Approaches

Another detection scheme proposed by Smaglik et al. [39] is lane-by-lane detection, in which the detectors monitor the gaps in each lane independently (i.e., each lane detector provides different inputs to the controller). When the critical gap is detected on a lane, the lane flags itself as gapped-out and starts waiting for other phases. When all lanes for that approach gap-out, the controller terminates the phase.

Comparison of lane-by-lane logic to the traditional detection can be illustrated using the following figure (Figure 3-8). Time t₀ is the beginning of green. Traces 1, 2, and 3 correspond to the state of inductive loop detectors in lanes L₁, L₂, and L₃. The signal is either TRUE (high, vehicle is over the detector) or FALSE (low). Vehicle arrivals are shown by dots on the trace line. Trace 4 is the OR function of Traces 1, 2, and 3 (i.e., it is FALSE only if all these signals
are FALSE when all lanes gap-out simultaneously), and corresponds to the traditional detection logic. Trace 9 shows the phase termination for traditional detection, after the critical gap, $t_g$, is observed for the approach as a whole. Traces 5, 6, and 7 represent the lane-by-lane termination requests for lanes L1, L2, and L3, respectively. Trace 8 (gap-out point with lane-by-lane detection) illustrates that once all three lanes have recorded a gap greater than $t_g$, the phase will terminate. Time $t_1$ is the moment that lane-by-lane detection gaps-out.

**Figure 3-8: Gap-Out Logic under Lane-By-Lane Detection and Traditional Detection** [39]

Clearly lane-by-lane detection logic is more efficient than the traditional gap-out logic as headways within each lane are more uniform, making it easier to identify the end of saturated
period within each lane. Their results indicated that lane-by-lane detection results in 3% to 5% reduction in wasted green time on multi-lane approaches compared to traditional detection [39].

However, the primary limitation of lane-by-lane detection is that, when the queue on one lane discharges before the other lane(s) (i.e., imbalance in lane utilization), lane-by-lane logic will still hold the phase. This still results in wasted green time on certain lanes, undermining a controller’s efficiency. Therefore, “multi-headway gap-out” logic was proposed in this research for actuated controllers on multi-lane approaches. The explanation of the proposed multi-headway gap-out logic is provided in Chapter 4.

### 3.6 Inefficiencies of Lagging Left in Actuated Signal Control

At signalized intersections where left-turn phases are protected from oncoming traffic through their own phases, left turns can be served either as “leading”, in which left turn precedes the opposing through movement, (Figure 3-9a), or as “lagging”, in which left turn follows the opposing through movement (Figure 3-9b).

---

**Figure 3-9: Dual Ring with a Lagging and a Leading Left Turn Phase**

Choosing leading vs. lagging schemes can be beneficial depending on traffic conditions. At isolated intersections (where coordination is not required) with actuated operation, two phases
ending at the barrier must terminate their green simultaneously. Leading left phases will pass the control to a through phase as soon as they gap-out, however lagging left phases often remain green after gap-out, waiting until the parallel phase in the other ring gaps-out or maxes-out. Therefore, it is preferred to use leading lefts at isolated intersections so that the slack time at the barrier goes to a through phase, which usually has a higher arrival rate than a left turn, and can therefore make better use of slack time.

Another advantage leading sequences offer when the left turn bay is subject to overflow is that they clear left turning vehicles out of an intersection approach earlier in the cycle [40]. For approaches with inadequate vehicle storage, this feature can allow relatively free movement of through traffic, leading to an increase in throughput capacity (important at isolated intersections with actuated control).

However, in a coordinated system, using lead-lag configuration (i.e., leading left in one intersection and lagging left in the other) at intersections with irregular spacing can improve two-way progression. Furthermore, lagging left is beneficial for minimizing delay to left turning vehicles on a coordinated arterial. With the leading left, vehicles arriving with the through band (particularly vehicles at the tail-end of an arriving platoon) and wanting to turn left must wait almost a complete cycle before receiving a leading green arrow signal [41] [42]. However, lagging lefts in a lead-lag arrangement are a source of inefficiency because the slack time at the barrier is used by lagging lefts, with generally lower flow rate compared to through phases.

Because self-organizing traffic signals (which use actuated control as the basis) aim to achieve good coordination, they will sometimes want to use lagging lefts in order to facilitate two-way progression and to minimize delay for left turning vehicles. In order to reduce the inefficiency of lagging lefts in a lead-lag arrangement, logic was developed to reduce the
likelihood of wasting green with lagging lefts. The proposed logic for lagging left turn phases is described in the next chapter.
Chapter 4: Multi-Headway Gap-Out and Efficient Lagging Left Turn Logic

Chapter 3 reviewed detection and controller settings to identify the sources of inefficiency in an actuated controller that can result in suboptimal operation. Based on the review, it is found that standard gap-out logic on a multi-lane approach and lagging left turn phase in a lead-lag arrangement in actuated operation cause wasted green time.

This chapter starts with the discussion of multi-headway gap-out logic to address the inefficient operation of gap-out logic on multi-lane approaches. Then it continues with the explanation of control strategies to improve the efficiency of actuated operation with a lagging left turn phase. The control logic for multi-headway gap-out is also reported in the paper published by Cesme and Furth [43].

4.1 Multi-Headway Gap-Out Logic for Actuated Control on Multi-Lane Approaches

The current practice in vehicle detection (i.e., traditional detection) on multi-lane approaches results in unnecessary green extensions, and thus is not efficient from an operation perspective. Lane-by-lane detection improves the efficiency of actuated control on multi-lane approaches by monitoring the gaps in each lane individually. However, when one lane gaps-out while other lane(s) has not gapped out yet, lane-by-lane control logic still extends the green phase for flow below saturation flow rate, causing suboptimal operation.

In order to overcome the inefficient operation of actuated controllers on multilane approaches, “multi-headway” gap-out logic is proposed as a criterion for detecting when saturation flow has ended on a multilane approach.
A multi-headway is the time it takes several vehicles to pass a detector. For a three-lane approach, for example, the controller looks at the time for three headways to pass while making the decision to gap-out. The concept of multi-headway logic can be described using Table 4-1. It shows a sample of vehicles arrivals on a three-lane approach and the associated headway under single headway (i.e., traditional), three-headways, and six-headways detection.

**Table 4-1: Vehicle Arrivals and Headways on a Three-Lane Approach**

<table>
<thead>
<tr>
<th>Vehicle Number</th>
<th>Time Vehicle Passed Detector</th>
<th>Single Headway (s)</th>
<th>Three Headways (s)</th>
<th>Six Headways (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>0.8</td>
<td>0.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>1.2</td>
<td>0.4</td>
<td>2.2</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>2.7</td>
<td>1.5</td>
<td>2.2</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>3.1</td>
<td>0.4</td>
<td>2.3</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>3.6</td>
<td>0.5</td>
<td>2.4</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>4.9</td>
<td>1.3</td>
<td>2.2</td>
<td>4.4</td>
</tr>
<tr>
<td>8</td>
<td>5.7</td>
<td>0.8</td>
<td>2.6</td>
<td>4.9</td>
</tr>
</tbody>
</table>

With multi-headway gap-out logic, if three-headways detection is used, the controller checks the time for three headways, for example the time between the first arrival and the fourth arrival or the time between the second arrival and the fifth arrival.

The underlying principle is that multi-headways have lower variability due to the constraints of longitudinal spacing within a lane, making it easier to distinguish saturated flow from unsaturated flow. During saturation flow, while different lane use afforded by multiple lanes creates high variability in the time between vehicles on the approach as a whole, longitudinal spacing needs within a lane keep the multi-headways more uniform, resulting in far less variability.
The criterion for multi-headway gap-out is specified in terms of the number of vehicles passing a detector and the critical interval; for example, allow up to 3.3 seconds for three vehicles to pass. An upstream detector records each vehicle passage. If the number of vehicle arrivals detected in the last 3.3 seconds is fewer than three, the phase gaps-out. Note the close relationship to flow rate, which is the number of vehicles in a given interval divided by interval length.

For practical purposes, if passage detection (pulse) is used, headways can directly be used in terminating a green phase. However, if a controller uses presence detection, the critical interval must be adjusted for the last vehicle’s occupancy time.

Because variability tends to decrease with sample size, it can be imagined specifying multi-headways for greater number of vehicles; for example, let the criterion for a three-lane approach be based on the time needed for six or nine vehicles to pass. However, the time it takes to measure, for example, nine headways becomes a source of inefficiency and requires detectors to be placed far upstream to minimize wasted green time during gap-out. Far upstream detection for through approaches may not be practical on urban streets due to physical constraints, which can be attributed to closely spaced adjacent intersections and short exclusive turn lanes (detectors should be placed after the start of the turn lane in order not to confuse different movements). Therefore the proposed logic focuses on multi-headways of size kL, where L is the number of approach lanes and k is one or two.

### 4.1.1 Efficiency of Single Headway and Multi-Headway Gap-Out Logic

The efficiency of gap-out logic on multi-lane approaches relies on headway variability during saturated flow. Low variability in headway makes it easier for the controller to detect the end of saturation flow, reducing wasted green time and improving controller’s efficiency.
In order to explore the headway variability of single headway and multi-headway detection at saturation flow rate and lower flow rates, the traffic micro-simulation model VISSIM was used. While using field data would have been preferable, finding data with the necessary level of detail and with significant periods of saturation flow on multilane roads proved too difficult.

To model saturation flow for one-, two-, and three-lane approaches to obtain a headway distribution during saturated flow, an approach was given a high input volume (5000 vehicles per hour per lane (vphpl)) and the signal was permanently held in green. That guaranteed an effectively infinite standing queue. A point (pulse) detector just downstream of the traffic signal was used to obtain headway data. For multilane approaches, point detectors in each lane provided headways on a lane-by-lane basis. Lane-level headway data were later combined to model single headway gap-out logic.

For model calibration, VISSIM’s Wiedemann 74 parameters [44] were adjusted to get saturation flow rate close to 1800 vphpl, which is a typical saturation flow rate value for through movements. The following formula describes the Wiedemann 74 car following model parameters in VISSIM and their relationship with discharge headway. An increase in discharge headway results in a decrease in saturation flow rate.

\[ d = ax + bx \]  \hspace{1cm} (4.1)

where \( d \) is the distance between two vehicles, \( ax \) is the standstill distance, and \( bx \) is the safety distance and calculated as follows:
\[ b_x = (b_{x_{add}} + b_{x_{mult}} \cdot z) \cdot \sqrt{v} \]  

(4.2)

where \( b_{x_{add}} \) is the additive part of desired safety distance, \( b_{x_{mult}} \) is the multiplicative part of desired safety distance, \( v \) is vehicle speed, and \( z \) is a value of range \([0, 1]\) which is normally distributed around 0.5 with a standard deviation of 0.15. Parameter values used for model calibration were as follows: for the average standstill distance \( (ax) \), 8.2 ft; for the additive part of desired safety distance \( (b_{x_{add}}) \), 2.3; and for the multiplicative part of desired safety distance \( (b_{x_{mult}}) \), 3.4.

Headway distributions for a single lane approach obtained from the calibrated simulation model are shown in Figure 4-1 for saturation flow and for 50% of saturation flow. The mean saturation headway was 1.98 seconds, corresponding almost exactly to the calibration target of 1800 vphpl, with a coefficient of variation (CV) of 0.188. For single headway gap-out, the critical headway was set to the 99.5-percentile value of the saturation headway distribution, which was 3.2 s. This implies a 0.5% chance (one out of 200 vehicles) of a “Type I” error. The headway distribution when the input flow was 50% of saturation flow is also shown in Figure 4-1. It is used to assess the probability of Type II error. The results show that when flow is at 50% of the saturation flow rate, 40.5% of the headways exceed the critical headway of 3.2 s, resulting in a gap-out.
H*: Critical Headway with 0.5% of Type I Error
μ: Mean Saturation Headway
CV: Coefficient of Variation of Saturation Headway

Figure 4-1: Cumulative Headway Distribution on a Single Lane Approach for Saturation Flow and 50% Saturation Flow

Note that unsaturated flow could be any level of flow less than saturation flow rate. The reason for choosing 50% of saturation flow is that it reflects a moderate level of demand, leading to long headways (e.g., headways greater than 3.2 seconds) that are necessary for gap-out as well as short headways that are necessary for Type II error.

The discrimination power of a gap-out test (i.e., the power of discriminating low flow rates from saturation flow) can be defined as the distance between the cumulative headway distribution (saturation flow and 50% saturation flow) at the critical headway, which is shown as black line in Figure 4-1. The discrimination power is given as “1 – p(Type I error) – p(Type II error)”. As shown in Figure 4-1, the discrimination power of a single-headway test on a single-lane approach with flow at 50% saturation is approximately 40%.
However, on multilane approaches, the greater variability in approach headways causes much lower discrimination power for a single-headway test. **Figure 4-2** shows results for a three-lane approach for single headway, triple-headway, and six-headway detection. In each case, the critical headway, indicated by the projection of the vertical black line to the horizontal axis (e.g., 5.3 s for six-headways detection), is the 99.5% (multi)headway during saturation flow. Therefore, Type I error is held constant at 0.5% for each detection scheme.

![Figure 4-2: Cumulative Headway Distribution on a Three-Lane Approach for Saturation Flow and 50% Saturation Flow using Single Headway, Three Headways, and a Six Headway Detection Scheme](image)

The power of single headway detection is low, less than 20%, because the cumulative distribution of headways during half-saturated flow at the critical headway is very close to the cumulative distribution at saturation flow. Therefore, with single-headway gap-out, if flow falls...
to 50% of saturation flow, four out of five arriving vehicles will extend the green phase, resulting in wasted green time.

With multi-headway detection, the discrimination power increases significantly. With three-headway detection, the power is increased to 55% and multi-headway coefficient of variation (CV) is decreased from 0.708 to 0.190. When six headways are used for gap-out, the discrimination power of the controller increases to approximately 80%, and CV decreases to 0.120. The incremental gain from using six-vehicle headways is not as dramatic, but is still significant.

It is also important to note that single-headways are not independent, but being heavily influenced by the need for longitudinal spacing in each lane, resulting in lower headway variation. According to statistical theory, if headways were independent, then the CV of multi-headways would be equal to:

\[ CV_n = \frac{CV_1}{\sqrt{n}} \]  \hspace{1cm} (4.3)

where \( n \) is the number of headways in the multi-headway analysis, \( CV_n \) and \( CV_1 \) are the coefficient of variations for \( n \)-headway and single headway detection, respectively.

If headways were independent, it would be expected to have a CV of 0.409 for three headways (0.708, CV for single headway, divided by square root of three). However, the obtained CV for three headways (0.190) is much lower, confirming that the reduction in variability from using multi-headway is not merely due to the aggregation effect. In fact, the CV for three headways on a three-lane approach is almost equal to headway CV for a single lane approach (0.190 vs. 0.188). This indicates that three-headway detection on a three-lane approach
is similar to traditional detection on a single lane approach, which works efficiently due to low headway variability.

If three vehicle headways were independent, then the CV for six-headways would be equal to 0.134 (0.190 divided by square root of two, equation 4.3). The obtained CV for six-headways (0.120) is very close to 0.134, which indicates that the reduction in variability from using six-headways is mostly due to the aggregation effect.

4.1.2 Detector Setback and Variable Green Time

Using an upstream detector to extend a green phase at an actuated signal reduces cycle length by avoiding wasted green time while waiting to detect the critical headway. The benefit of using an upstream detector becomes more pronounced with multi-headway detection logic. Because the critical headway under multi-headway gap-out logic is higher than the critical headway with traditional gap-out logic, the wasted green time while waiting to detect the critical headway would be higher with multi-headway logic if stop line detection was used to gap-out. Therefore, it is assumed that multi-headway gap-out logic will always be used in connection with upstream detection.

However, a primary concern with upstream detection, as explained in Chapter 3, is the risk of premature gap-out. To prevent early termination of green, it is proposed employing variable minimum green logic, which is also part of so-called volume-density control [37]. It means that an initially specified short minimum green is increased by a certain amount with each vehicle actuation during the associated phase yellow and red intervals. The variable minimum green used with multi-headway detection is set by the formula

\[
\text{MinGreen} = L_s + n \cdot h_{sat} \tag{4.4}
\]
where \( L_s \) = start-up lost time, \( n \) = number of queued vehicles per lane at the beginning of green, and \( h_{sat} \) = saturation headway for a single lane. The number of queued vehicles is obtained by placing a counting (pulse) detector upstream of the stopline, a counting detector after the stopline, and using the difference between vehicles counted at the upstream detector and vehicles counted when exiting the intersections. This detection method can also be considered as “trap logic” in which signal controller calculates the number of vehicles within the trap at the beginning of green.

4.1.3 Accounting for Turning Vehicles Leaving Through Lanes

When a through movement queue extends beyond the start of a turn lane (Figure 4-3a), through lanes adjacent to the turn lane may experience long headways [45] while the queue is discharging due to vehicles turning off into the turn bay from the through queue (Figure 4-3b).

a) Through queue extends beyond turn lane  

b) Turning vehicles leaving through lanes

Figure 4-3: Long Headways during Queue Discharge due to Turning Vehicles Leaving through Lane

If detectors are placed after the start of turn lane(s) (which is where they are usually placed so as not to confuse different movements) and the queue extends beyond the start of the
turn bay, leaving turning vehicles increase the headway on through lanes adjacent to the turn lane. That can increase the probability of a Type I error by making the controller think that the standing queue has been served.

One remedy is to use a modified headway distribution taken from a situation in which vehicles are turning out of through lane, and selecting the 99.5% headway from this distribution as the critical headway. Such a distribution can be obtained from an experiment where the through queue extends beyond a turn lane and some vehicles depart from the through lane during saturation. That will result in a more generous critical headway, one that will prevent a phase from gapping-out when a single turning vehicle departs from a through queue during saturation flow. Another solution is to use a headway distribution without leaving vehicles, but reducing the number of headways that must be detected by one (e.g., instead of measuring the time for six headways, measure the time for five headways), in effect allowing for a “hole” in the arriving vehicle stream. With either remedy, the controller should extend green phase when the queue is discharging and one vehicle turns off during a multi-headway period. However, if two consecutive vehicles turn off, the stream will probably fail to meet the multi-headway criterion. Both potential remedies (i.e., using a more generous critical headway threshold or reducing the number of headways to account for holes) were tested using simulation. Results are provided in the next sections.

4.1.4 Test Bed for Multi-Headway Logic

In order to evaluate the effectiveness of multi-headway gap-out logic, a simulation testbed was developed in VISSIM. The testbed consists of an isolated junction of a six-lane road with a four-lane road, with left turn bays on all approaches that have protected green only, and with signal control following the standard dual-ring, eight phase structure, as shown in Figure 4-4.
Base traffic volumes (Figure 4-4) were chosen so that the flow ratio ($\Sigma(v/s)$) for the critical movements would equal 0.72. Assuming 18 seconds of lost time per cycle (four seconds for the left turn phases and five seconds for through phases) and a target v/c ratio of 0.90, a flow ratio of 0.72 would reflect a scenario in which the average cycle length will be 90 seconds if flow is maintained at 90% of saturation flow during green periods.

The distribution between left and right-through volumes was done using a biproportional model of origin-destination distribution with a propensity ratio of 18 : 82, which has been shown to be consistent over a wide range of intersections [46] when factored biproportionally to account for input and departure volumes. The propensity model estimates turning flows at an intersection to match given inflow and outflow volumes when intersection specific counts are unavailable. In

Figure 4-4: VISSIM Simulation Geometry, Traffic Volumes, and Eight-Phase Dual Ring Control Used to Test Multi-Headway Logic

The distribution between left and right-through volumes was done using a biproportional model of origin-destination distribution with a propensity ratio of 18 : 82, which has been shown to be consistent over a wide range of intersections [46] when factored biproportionally to account for input and departure volumes. The propensity model estimates turning flows at an intersection to match given inflow and outflow volumes when intersection specific counts are unavailable. In
our testbed, inflow and outflow volumes were assumed, and turning flows were obtained through the biproportional model.

Base volumes were scaled up or down to create a heavy traffic scenario ($\Sigma v/s = 0.85$) and a lower traffic scenario ($\Sigma v/s = 0.60$). Treatments analyzed include single headway (traditional), lane-by-lane, and multi-headway gap-out logic.

Multi-headway logic was only tested with upstream extension detectors; however single headway and lane-by-lane gap-out logic were tested with both stop line and upstream extension detectors, because stop line detection is the most common type of detection at actuated intersections. For the east-west street (three lane approaches), multi-headway logic used six headways, and for the north-south street, it used four headways. For single lane turn lanes, traditional single headway gap-out logic was applied.

In the dual-ring structure tested, left turns were leading and could be skipped for lack of demand; through phases had recall (i.e., could not be skipped). Phases terminating at a barrier had non-simultaneous gap-out. For all phases, six seconds of minimum green interval was used unless superseded by a greater variable minimum green time (equation 4.4). Maximum green time was set arbitrarily as 55 seconds for east-west through phases, and 40 seconds for north-south through phases.

The controller logic was programmed in C++ and interfaced to VISSIM through its application programming interface (API). In each simulation time step, the vehicle simulator advances vehicles based on rules of vehicle behavior, and then passes status to the controller. The control program will then change the signal state based on that detector information, and then return control to the vehicle simulator. Time steps for detection and control logic was 0.1 seconds.
Tests compared average intersection delay, a direct indicator of traffic performance, and average cycle length, a measure of operational efficiency, for different gap-out treatments. Simulation model was used to evaluate different gap-out treatments.

4.1.5 Results and Analysis

Simulation experiments were based on five replications with different seeds (each seed generates different input flow arrivals for each simulation, introducing variability in the model). Each replication lasted 3600 s after a 300 s warm-up period. When intersections are under-saturated, queues clear in every cycle, and so each cycle is, in effect, a virtually independent experiment, reducing the number of simulation replications required to achieve a given level of confidence in the output.

For single and multi-headway logic, critical headway was selected so that the probability of a Type I error was equal to 0.5%. For three-lane approaches (east-west), the critical headway threshold was selected as 2.2 seconds for single headway gap-out and 5.3 seconds for six-headway gap-out. In order to account for departing turning vehicles, the controller watched for five headways during this six-headways threshold value (i.e., 5.3 seconds), as described earlier. The option of detecting six-headways with 6.5 seconds as critical headway (a more generous critical headway to account for turning vehicles) was also tested, where 6.5 seconds is the critical headway that corresponds to 0.5% Type I error when some vehicles in the left lane turn off during saturation flow. The results of these two options were found to be almost identical, and so the results reported come from five-headway detection. On two-lane approaches (north-south), critical headways were 2.5 seconds for single headway gap-out and 5.4 seconds for multi-headway (four headways). No provisions was made for vehicles turning out of the through lane
on the north-south street, because a long turn bay was used for the north-south street, limiting the chance of a through movement queue extending beyond the start of the turn bay.

With lane-by-lane detection, there is no direct way of applying the 0.5% Type I error criterion, since an error in one lane does not necessarily lead to gap-out. Therefore, a reasonable value of 2.8 seconds was assumed as critical headway.

For the primary comparison, upstream detection was used for all methods so as not to unduly distort the comparison. Simulation results are given in Figure 4-5, which provides a comparison of average intersection delay and average cycle length for the different gap-out treatments. The results are provided in stacked columns, in which the height of the lower column shows average delay and the height of the top column (not its length, but its final height) represents average cycle length. Simulation results indicated that when traffic flow is low (i.e., \( \sum v/s = 0.60 \)), the average delay reduction due to multi-headway gap-out logic compared to single headway and lane-by-lane logic was less than one second. Nevertheless, multi-headway logic resulted in a seven second (12%) reduction in cycle length compared to single headway, showing greater benefit to pedestrians and transit vehicles.

In moderate flow conditions (i.e., \( \sum v/s = 0.72 \)), the delay reduction associated with multi-headway logic became noticeable, though still small (1.6 seconds and one second reduction in average delay compared to traditional and lane-by-lane gap-out logic, respectively). The cycle length was reduced by 12 seconds (15%) compared to traditional detection and by eight seconds (10.8%) compared to lane-by-lane detection.
The greatest reduction in delay and cycle length came under high traffic flow (i.e., $\sum v/s = 0.85$). Compared to single headway upstream detection, multi-headway logic reduced average delay by four seconds (10%), and shortened the cycle by 16 seconds (15%). In this higher demand scenario, lane-by-lane gap-out resulted in greater delay than single-headway gap-out (42.0 vs. 39.8 seconds of average delay), and both were clearly outperformed by multi-headway gap-out (35.9 seconds of average delay). The poor performance of lane-by-lane gap-out appeared to be due the effect of turning vehicles leaving the leftmost through lane to enter the turn lane,
causing that leftmost through lane to gap out early and increasing the frequency of premature gap-outs for the approach as a whole. This is also supported by the cycle length values. Cycle lengths under lane-by-lane logic were substantially smaller than single headway detection; however lane-by-lane logic resulted in higher delays, because east-west phases often gapped-out prematurely, due to turning vehicles leaving the leftmost through vehicles to enter the turn lane.

Finally, the results show the strong benefit of upstream detection for both single-headway and lane-by-lane detection in comparison with stop line detection only. There are significant reductions in delay and cycle length compared to having the extension detector at the stopline, especially at higher volumes. Compared to traditional single-headway, stopline detection, more benefit is gained in going to upstream detection than the incremental benefit, once detectors are moved upstream, to using multi-headway detection.

4.1.6 Imbalance in Lane Utilization

When a standing queue in a multilane approach dissipates, some lanes are bound to empty out sooner than others. With snappier gap-out criteria, the probability increases that the phase will gap out because one or more lanes have no more flow while some queued vehicles remain in lanes that haven’t yet fully discharged. Simulation models capture the kind of end-of-phase imbalance that arises from randomness in lane selection and discharge rate. However, in many realistic situations drivers exhibit a clear preference for certain lanes that can make the end-of-phase imbalance greater. On one hand, signal control strategies that reward drivers for making better utilization of lanes are desirable; on the other, some accommodation may be desirable, at least for the first few seconds of unsaturated flow, to meet drivers’ expectancy.

The proposed multi-headway gap-out logic does not explicitly take into account imbalance in lane utilization. The length of the multi-headway threshold will affect performance
with respect to lane utilization imbalance. With a more generous critical multi-headway threshold selection, it is more likely that the controller will extend the green for flow below saturation, therefore accommodating uneven lane utilization to a certain extent. With snappier gap-out settings; it is more likely that the phase would gap-out when periods of uneven lane utilization starts in order to minimize wasted green time.

Some of the strategies to deal with poor lane utilization to be tested in future research can be described as follows:

- When an intersection operates below capacity, a more generous gap-out criterion can be used that allows uneven lane utilization to a certain degree. That way, some portion of slack time can be used to meet driver’s expectancy. However, when an intersection operates at or close to capacity (determined adaptively), snappy gap-out settings can be applied to avoid wasted green time and maximize intersection capacity.

- Another strategy could be, as explained earlier, using a more relaxed gap-out criterion at the beginning of a green phase, and transitioning into a stricter gap-out criterion after a specified green time. This logic is similar to so called volume-density control, which reduces the critical gap-time to a smaller value once the specified green time expires [37].

### 4.2 Efficient Lagging Left Turn Logic

Lagging left turn schemes tend to reduce the efficiency of actuated operation, limiting their use in actuated control of isolated intersections. Along arterials, on the other hand, lagging lefts are often used, because they help provide good two-way progression where intersection spacing is not ideal, and help reduce delay to left-turning traffic. Because the goal of self-organizing signals is to provide good coordination along arterials while using actuated control as the underlying framework, one of the objectives of this research is to develop efficient lagging left turn logic for
actuated control so that lagging left can be used to improve coordination without degrading the performance of actuated control.

### 4.2.1 Efficient Lagging Phase Skipping Logic

In an actuated signal, a phase is skipped if there is no vehicle detection (unless the phase is set to recall). Consider the phase scheme of **Figure 4-6**, with a lagging left for the eastbound approach. When the opposing through phase, that is, Westbound Through (WBT) gaps-out at time $t_1$, if there is no vehicle detection on Eastbound Left (EBL) movement, the controller skips EBL and extends the green for WBT until Eastbound Through (EBT) either gaps-out or maxes-out (until time $t_2$).

![Figure 4-6: Dual Ring with a Lagging Left Turn Phase and a Leading Through Phase](image)

**Figure 4-6: Dual Ring with a Lagging Left Turn Phase and a Leading Through Phase**

Phase skipping generally improves the efficiency of actuated controllers, because it eliminates unnecessary delays by not allocating green time to an uncalled (no vehicle detected) phase (i.e., EBL in this case). The decision to skip EBL benefits WBT vehicles that arrive between $t_1$ and $t_2$; however, it also increases delay for EBL vehicles arriving during this interval. Because actuated operation is typically used at isolated intersections, in which arrivals are random, if no EBL vehicles arrived in the entire previous red period, which led to skipping, chances are low that there will be more than one EBL arrival during the $t_1$ to $t_2$ interval. As a
result, phase skipping improves the efficiency of an intersection. Moreover, not skipping forces
the controller to provide minimum split for the lagging phase, which can increase cycle length
and delay.

However, along arterials with closely-spaced intersections, vehicles are expected to arrive
in platoons rather than arriving independently. If WBT gaps-out early due to low demand in one
cycle, late arrivals of left turning vehicles from an upstream intersection may result in skipping,
and because vehicles arrive in platoons, it is more likely that there will be multiple left turn
arrivals during the t₁ to t₂ interval. As a result, skipping would make those left turn drivers wait
about one entire cycle to clear the intersection.

Moreover, skipping may lead to spillback from a turning lane (e.g., EBL) into a through
lane (e.g., EBT) if a platoon arrives into a turning lane with limited storage capacity, causing a
drop in through lane throughput. The throughput reduction is especially undesirable if spillback
is from a turning lane into a critical through movement.

Therefore, this research proposes logic for overriding a skip decision before a ring
crosses the barrier. The proposed override logic is based on:

- The expected number of left turn arrivals occurring after skipping but before reaching the
  barrier (i.e., during time interval t₁-t₂)
- The spillback risk from a turn lane onto a through lane if the controller decides to skip
  the left turn phase (i.e., EBL in this case)
- Whether a potential spillback would affect a critical through or non-critical through phase

This override logic applies to a lagging left turn that is positioned in the dual ring to reach a
barrier at the same time as its same direction through phase. The developed control algorithms to
override a skip decision for a lagging left turn phase are explained as follows:
1. “Skip” lagging left if no vehicle has been detected on the lagging left turn phase when its opposing through phase has gapped-out (i.e., at time \(t_1\))

2. However, at any time after \(t_1\) and before \(t_2\), “Override Skipping Decision” IF
   a. Two or more vehicle calls are registered for the left turn AND the intersection volume to capacity (v/c) ratio, calculated adaptively, is below 0.90.
   b. Two or more vehicle calls are registered for the left turn before time \(t_2\) AND the intersection v/c ratio is greater than 0.90, but probability of spillback (spillback risk) from a turn lane to a through lane is greater than 0.25, AND spillback would affect a critical through phase. Spillback risk is calculated as follows:

\[
p(\text{spillback}) = p(x > n) = 1 - p(x \leq n) \tag{4.5}
\]

where \(p(\text{spillback})\) is the spillback risk, \(n\) is the left turn storage capacity in number of vehicles, specified by the user, and \(x\) is the expected number of arrivals. Therefore, \(p(x > n)\) is the probability that the expected arrivals are higher than the storage capacity, which is calculated using Poisson distribution:

\[
p(x > n) = \sum_{x=n+1}^{\infty} \frac{\lambda t^x e^{-\lambda t}}{x!} \tag{4.6}
\]

where \(\lambda\) is the average number of left turn arrival rate (calculated adaptively based on the previous five cycles) and \(t\) is the expected \(t_1\) to \(t_2\) time interval, which is also
calculated adaptively based on the average EBL split time in the previous five cycles, where it was not skipped.

For example, suppose a left turn can store five vehicles and there are already three vehicles registered for the left turn. The controller calculates the probability that there will be more than two additional arrivals using Poisson distribution. The main limitation of this calculation is that Poisson distribution assumes independent arrivals, while vehicle arrivals are affected by control decisions made at an upstream intersection.

The underlying principle of the proposed override logic is that once the left turn has registered two arrivals (which typically need four seconds of green assuming two seconds of saturation headway), the lost time associated with serving a lagging left would be only a little more than change interval, assuming a minimum green time for left turn phases of six seconds (if a higher minimum green interval is used, a different threshold can be used instead of two arrivals). When an intersection has extra capacity (i.e., v/c is below 0.90), that lost time can readily be accommodated.

However, when an intersection has little excess capacity (i.e., v/c > 0.90), concerns about delay to left-turning vehicles are less important than the need to use capacity efficiently by avoiding wasted green time. In this case, overriding is performed only when there is a significant chance of spillback into a critical through phase, because spillback reduces critical through lane capacity, also reducing intersection capacity.

**4.2.2 Efficient Lagging Green Start Logic**

In a dual ring phasing, one ring is critical, that is, it requires longer green time than a non-critical ring, leading to slack time in non-critical ring. Ideally, the slack time in a non-critical ring should
be given to a through phase rather than a left-turn phase, because through traffic usually has a higher arrival rate than a left turn, and can therefore use the slack time more efficiently. Where left turns are leading, leading left phases will pass the control to a through phase as soon as they gap-out, and therefore slack time will go to a through phase. However, where left turns are lagging, the slack time is used by a left turn phase, which reduces the efficiency of an intersection.

To make lagging left operations more efficient, when a lagging left phase is not skipped, slack time after the left turn queue has cleared but before the critical phase has reached the barrier should ideally go to the opposing through phase, which can usually use it more effectively because of its higher arrival rate. What is needed for efficient lagging left is a method that predicts how much time the critical through phase needs (i.e., predict time $t_2$) and how long a split the lagging left needs (Figure 4-7).

![Figure 4-7: Slack Time at the Barrier Goes to a Leading Through Phase](image)

The goal is to start the lagging left (in Figure 4-7, EBL) so that its queue discharge ends when its parallel through phase EBT gaps out. That way the slack time still goes to the opposite through phase, but in advance, making the intersection more efficient. The following algorithm
describes how the controller delays the start of lagging left (EBL) so that slack time goes to the leading opposing through (WBT). Note that priority decreases as the rule number increases (i.e., Rule 1 has the highest priority over all).

1. IF spillback is detected from a turn lane (turn pocket spillback) into a critical direction lane through spillback detectors (Figure 4-8) AND the intersection is operating under capacity, “Do Not Delay Lagging Left” (i.e., allow opposing through to gap-out as usual) in order to expedite start of green for the lagging left, limiting capacity reduction on the parallel through lane due to spillback. If spillback into a critical direction lane is detected during oversaturation, controller “Truncates” green for the opposing through after minimum green criteria is satisfied (the details of green truncation logic is given in Section 5.4)

![Figure 4-8: Spillback Detector to Prevent Turn-Pocket Spillback](image)

2. IF there is no spillback detection from a turn lane, “Extend Opposing Through (i.e., Delay Lagging Left)” to match common green end for the lagging left and its parallel through phase, but biased toward early green start for lagging left in order to avoid increasing average cycle length. Upstream detectors detect gaps in advance to predict when the parallel through phase will gap-out (EBT), and detector counts are used to estimate the required green time for the lagging left (EBL) to discharge its queue.
Due to prediction uncertainty, it is likely to observe situations in which a lagging left (EBL) extends its parallel through phase (EBT), because EBL has not discharged its queue yet when EBT is ready to terminate its green phase. In such a case, the delayed start of lagging left increases cycle length. To avoid late start of lagging left, a bias factor of 1.15 is used to increase the estimated green time by 15% to provide some slack to the lagging left to limit the chance that cycle length is increased due to late start of lagging left. Note that 1.15 used as the biased factor was arbitrarily selected, which suggests an area for future research.
Chapter 5 : Arterial Signal Control Strategies

At isolated intersections, an efficient actuated controller cycles as quickly as possible, avoiding wasted green time by holding the green signal only until saturation flow has ended, subject to minimum green times. As a result, variable cycle lengths are obtained at different intersections, because the amount of time a signal needs to serve the demands of its conflicting phases are different. Allowing every signal to operate at its own cycle length makes it difficult to create a regular pattern of green waves along an arterial as signal coordination requires a common cycle length.

Actuated control has one self-coordination mechanism which comes true standard green extension logic before a phase gaps-out. If a platoon released from an upstream intersection reaches the next intersection when the signal is green, standard gap-out logic will hold the signal green until the platoon clears the intersection, which promotes the formation of green waves. However, this type of natural progression with actuated operation is only realized when adjacent intersections tend to operate at similar natural cycle lengths. If adjacent intersections cycle significantly faster or slower than each other, they will not naturally coordinate under actuated control. Therefore, stronger coordination rules are needed in order to make actuated control suitable for arterials.

This chapter introduces signal control strategies which add communication and coordination mechanisms intended to make actuated control self-organizing at an arterial level. In Section 5.1, rules for secondary extension at largely-spaced intersections are introduced. In Section 5.2, dynamic coordination logic at closely-spaced intersections is presented.
5.1 Secondary Extension Logic: Progression vs. Cycle Length Tradeoff

A short cycle, resulting from an efficient actuated control, minimizes delay for pedestrians and non-arterial traffic. However, if a platoon is expected to arrive shortly after an arterial approach’s phase gaps out, it may be worth giving that phase a “secondary extension”, meaning some additional green time beyond gap-out, because such an extension will drastically reduce delay for the vehicles in the platoon while only slightly increasing cycle length (and therefore delay) for other traffic. Therefore, there is a tradeoff between cycling as quickly as possible and giving a secondary extension for an arriving platoon to improve progression at the cost of wasting green and increasing cycle length [47].

When a downstream intersection cycles faster than its upstream neighbor, secondary extension logic provides a buffer for the front side of a green wave, preventing downstream intersections from terminating the green phase if a platoon from the upstream intersection is due to arrive soon. That way, a secondary extension allows an intersection to wait for an arriving platoon, which facilitates coordination on the arterial level, without forcing intersections to have a fixed, common cycle length.

To facilitate secondary extensions, upstream detectors should be placed roughly 20 seconds of travel time upstream of an intersection. With these detectors, the controller monitors the arrival profile for an approaching platoon. The arrival profile is the cumulative expected arrivals at the stopline at a time $t$ in the future for $t = 1, 2, \ldots$ seconds. With every detector actuation, the controller updates the profile of arrivals using a known travel time offset (Figure 5-1). The proposed arrival prediction method ignores possible platoon dispersion.
Vehicles in an upstream intersection’s queue can be included in the profile (using communication between intersections) beginning from the onset of the change interval preceding its green start. This approach is based on knowing when the green will begin, an assumed saturation headway, and queue length information from the upstream controller. In the analyses, vehicles in an upstream intersection’s queue were included in the arrival profile wherever intersection spacing was less than 20 seconds. Depending on intersection spacing and queue length, the expected arrival profile may be known for 20 seconds or more.
The communication scheme used between intersections to get platoon information from an upstream intersection (INT 1, Figure 5-2) to a downstream intersection (INT 2) is described as follows:

1. When the change interval preceding INT 1’s green starts, the total number of queued vehicles in its trap (i.e., between stop line detector and upstream detector at the upstream intersection, Figure 5-2) is passed to INT 2. Let the time when the change interval starts at an upstream intersection \( i \) be \( t \) and let \( A(t) \) be the expected arrival time at a downstream intersection \( j \), which is given by:

\[
A(t) = t + Y_i + TT_{ij} + k * HSAT_i
\]  

(5.1)

where \( n_{Q,i} \) is the queued vehicles in the trap at intersection \( i \), \( P_{RT,i} \) is the proportion of right-turning vehicles for the through movement at intersection \( i \), which is pre-specified by the user, \( Y_i \) is the length of change interval for the through movement at \( i \), \( TT_{ij} \) is the travel time from \( i \) to \( j \), and \( HSAT_i \) is the saturation headway of the through movement at \( i \). Queued vehicles in the trap generally include right-turning vehicles unless there is a right turn bay. Therefore, \( P_{RT,i} \) is considered in the arrival profile estimation.

2. While INT 1 is green, detections at INT 1’s extension detector are used to update the arrival profile INT 2 with a travel time offset.
Figure 5-2: Communication from an Upstream Intersection to Obtain Arrival Profile When Secondary Green Extension Horizon is Limited due to Close Intersection Spacing

5.1.1 Importance of an Arriving Platoon in Providing Secondary Extension

A secondary green extension when an arterial phase gaps-out reduces through vehicle delay, but increases wasted green time and cycle length. Whether an extension is granted for an arriving platoon, therefore, becomes an important question in the tradeoff between arterial progression and delay to non-arterial phases.

The willingness to grant a secondary extension along an arterial should increase as the platoon of vehicles becomes:

- **Larger** (more vehicles benefit from a fixed charge of waiting until the platoon arrives)
- **Denser** (less wasted green time or lost time within the platoon)
- And **more imminent** (closer to the intersection, or shorter gap before it arrives, which results in less fixed charge of waiting).

A proposed measure that considers these three parameters is $L^*$, that is, the lost time (or wasted green time) per vehicle. $L^*$ is defined as the minimum ratio of lost time during a tentative
secondary extension to the number of arrivals during that tentative secondary extension, minimized over different potential lengths of secondary extension. Let time $t$ be initialized so that the current time is 0 (i.e., the time that a phase gapped-out), and let $L(t) =$ lost time per vehicle if the secondary extension’s length is $t$, given by

$$L(t) = \frac{t}{n(t)} - h_{sat}$$

and

$$L^* = \min_{t = 2, 4, ..., Sx_{max}} L(t)$$

where $n(t)$, the number of vehicles expected to pass the stop line if the green phase is extended by $t$ is known from the arrival profile, and $h_{sat}$ is the saturation headway. The first term on the right hand side of equation (5.2) is the average time headway, that is, $h_{avg}(t)$.

$L(t)$ is then calculated for discrete values of $t$ up to $Sx_{max}$, the maximum allowed length of secondary extension, to find the value of $t$ that minimizes $L(t)$ (equation 5.3), or equivalently minimizes $h_{avg}(t)$ since $h_{sat}$ is constant regardless of the value of $t$. Note that for low values of $t$, $h_{avg}(t)$ is high, because the calculation is done at the moment of gap-out, therefore few vehicles (if any) will be expected to arrive in the first few seconds.

In the common case in which the arrival profile includes a gap followed by a dense platoon of uniform density, $L^*$ will be minimized at the time $t$ at which the last vehicle in the platoon reaches the stopline. Observe that in such a case, $L^*$ will be smaller if the platoon is larger, denser, and more imminent (shorter gap before it arrives).
Calculation of $L^*$ is demonstrated in Figure 5-3. Potential lengths of a secondary extension ($t$) are shown on the vertical axis, expected arrivals from the arrival profile are displayed on the horizontal axis. Average time headway (s/vehicle) for any given extension interval can be calculated by finding the slope of the ray to the arrival profile for that extension. A steep slope indicates a long average time headway (or low flow rate), while a low slope indicates a short average time headway (or high flow rate). Therefore, saturation flow has the lowest slope and the shortest time headway (which is indeed saturation headway, $h_{sat}$).

$t^*$: Secondary extension length that minimizes lost time per vehicle, that is, $L(t)$

$n^*(t)$: Number of vehicles expected to pass stopline at the optimum secondary extension length

$h^*_{avg}(t)$: average time headway at the optimum secondary extension length

Figure 5-3: Secondary Green Extension Time versus Cumulative Number of Expected Arrivals for Granting Secondary Extension for an Arriving Platoon
The solid line indicates the predicted arrival profile for an approaching platoon. Time zero (origin) is the time at which a phase gaps-out, which is when the decision about secondary green extension must be made. As shown in Figure 5-3, the approaching platoon has different flow regimes. Until the time at which the platoon will arrive at the intersection at point A, the arriving flow is sparse, resulting in a steep slope. From then until when the platoon finishes passing the intersection (point C), slope decreases as the arrival rate is high. After the tail of platoon passes the intersection (point C), arrival rate decreases again (slope increases) and becomes less than the average arrival rate up to that point.

At any secondary extension length, the lost time per vehicle \((L(t)\) in equation (5.2)) can be obtained by calculating the slope difference between the ray to the arrival profile curve (average headway) and saturation flow curve. The optimal secondary extension length \((t^*)\) corresponds to the minimum slope for the arrival profile curve, leading to minimum slope difference between the arrival profile and saturation flow (Figure 5-3).

### 5.1.2 Use of Excess Capacity for Secondary Extension

During under-saturated periods, excess capacity (i.e., the difference between the intersection capacity and traffic demand) at an intersection can be treated as a resource that a controller can use, sacrificing a shorter cycle to extend the green for an arriving platoon’s passage. However, if an intersection becomes more saturated, the willingness to provide secondary extension decreases due to limited excess capacity. In case of oversaturation, secondary extension should be inhibited in order not to waste any green time.

In order to take into account excess capacity while making the decision of secondary extension for an arriving platoon, secondary extension will be granted for a platoon only if its lost time per vehicle, \(L^*\), does not exceed “affordable lost time”, that is, the green time that the
controller can afford to waste to accommodate arriving platoon, which varies with how much excess capacity an intersection has. Because there is insufficient basis for establishing an affordable lost time curve, an empirical curve is suggested, which is a function of intersection \( v/c \) (volume to capacity) ratio (Figure 5-4). The calculation of the intersection \( v/c \) ratio is given by

\[
v/c = \frac{\sum_{i=1}^{\text{critical}} \frac{v_i}{s_i} \left(1 - \frac{L}{C}\right)}{1 - \frac{L}{C}}
\]  

(5.4)

where \( v_i \) and \( s_i \) are a movement’s arrival rate and saturation flow rate, the sum in the numerator is over critical movements only, \( L = \) sum of the lost time for the critical movements, and \( C = \) maximum desirable cycle length. The calculation used \( C = 90 \) seconds, assumed that \( L \) was four seconds per critical phase, and measured \( v \) and \( s \) adaptively (i.e., using detectors within the simulation environment, updating estimates with every new cycle).

![Figure 5-4: Affordable Lost Time per Vehicle with Respect to Intersection Volume to Capacity (v/c) Ratio](image-url)
As the intersection v/c increases, affordable lost time per vehicle approaches zero, making it more difficult for an arriving platoon to secure a secondary extension, and if an intersection is over capacity, secondary extensions are not granted. For low levels of v/c ratio, a cutoff of two seconds of lost time per vehicle (typical saturation headway per lane value) was considered. As a result, even for very low values of v/c ratio (e.g., 0.40), the controller will not grant an extension that will waste more than two seconds per vehicle (corresponding to flow at 50% saturation assuming two seconds of saturation headway). The empirical formulation of affordable lost time per vehicle is given as follows:

\[
\text{Affordable Lost Time per Vehicle} = \min\{L_{\text{max}} \left( \frac{\text{Critical } v/c \text{ Ratio}}{\text{Intersection } v/c \text{ Ratio}} - 1 \right) * L_{\text{max}} \} \quad (5.5)
\]

where critical v/c ratio is specified as 1.0 and \(L_{\text{max}}\), that is, maximum affordable lost time per vehicle is specified as two seconds per vehicle. The affordable lost time is essentially the allowed lost time per vehicle with respect to \(L_{\text{max}}\) that brings the intersection v/c ratio to the critical v/c ratio. Because the critical v/c ratio is selected as 1.0, indicating that the intersection would operate just at capacity if the critical v/c ratio has been reached, the secondary extension is only given when accommodation of platoon will not result in over capacity in this cycle. The particular values described in this criterion were selected based on judgment. The sensitivity analysis of the values and detailed explanation are provided in Chapter 8.

**5.1.3 The Maximum Allowed Secondary Extension**

The maximum allowed extension, \(S_{X_{\text{max}}}\), is defined as the maximum allowed secondary extension length, \(t\), considered when choosing an optimizing value of lost time per vehicle, that is, \(L(t)\).
For phases serving the critical arterial through movement, as a rule of thumb, the maximum allowed extension, $SX_{max}$, was specified as 20 seconds in the analyses (note that sensitivity analysis of $SX_{max}$ was performed in Chapter 8). A generous limit (e.g., 20 seconds) was allowed for the critical direction because if a critical phase is extended in one cycle to serve a platoon that would otherwise have been served in the next cycle, the longer red interval that conflicting movements will face in the first cycle will be offset by a shorter red in the next cycle. The same is not true for the non-critical arterial through movement, and so for the non-critical direction a more stringent maximum is specified for its secondary extension:

$$SX_{max} = \min\{\max(10, \Delta C_n), 20\}$$

(5.6)

where $\Delta C_n$ is the difference between the intersection’s natural cycle length and neighboring intersections’ natural cycle length (Figure 5-5).

* $\Delta C_n$ is taken as zero when an intersection’s natural cycle length is higher than its neighbors

**Figure 5-5: Natural Cycle Length Comparison at Neighboring Intersections to Determine the Maximum Allowed Secondary Green Extension for the Non-Critical Direction**
The secondary extension for the non-critical arterial phase is limited to 10 seconds when the local intersection’s natural cycle length is close to or higher than the maximum cycle length of its neighboring signals (equation 5.6). That way, secondary extension for the non-critical direction is given only to arriving platoons with very little lost time per vehicle (i.e., large platoons with high density, and close to intersection). As a result, the increase in cycle length becomes small due to little lost time, and the small difference (which helps coordination) between cycle lengths at neighboring intersections is maintained.

However, if a neighboring signal has a much longer cycle length than the local intersection, then a longer extension will be allowed, because it will improve progression for the critical direction (by promoting closer cycle lengths at adjacent intersections) as well as non-critical direction.

5.1.4 Granting a Secondary Extension - Criteria

Each time an arterial phase gaps out, the lost time per vehicle in the arrival profile and the excess capacity at an intersection are calculated. If the calculated lost time per vehicles meets the criteria for affordable lost time (i.e., smaller than the affordable lost time), a secondary extension will be granted, with its length being the value of $t$ that minimizes $L^*$. When the calculated secondary extension ends (the controller extends green by $t$ seconds), normal gap-out logic governs, allowing the controller to extend the green even farther as long as the flow remains heavy. That way, an entire platoon can be served even if all of it was not detected because of a limited horizon or a limit in maximum allowed extension. A phase may receive only one secondary extension per cycle.
5.2 Dynamic Coordination for Coupled Intersections

When intersections are closely-spaced, queue management is needed in order to prevent spillback and starvation at the critical intersection to preserve intersection capacity. At closely-spaced intersections, the coordination cannot merely rely on the secondary extension, which may or may not hold a signal for an arriving platoon, because a failure to provide good progression may cause spillback with resulting losses in intersection capacity.

Closely spaced intersections also offer an opportunity for good two-way progression by starting their arterial greens at the same time. Good two-way progression between closely spaced (e.g., intersection spacing is less than 600 ft) intersections also improves transit service when there is no transit stop between intersections.

Therefore, it is proposed that closely spaced intersections be coupled using a special control logic that forces a coupled group of intersections to cycle together without any fixed cycle length. This approach was inspired by Zurich’s traffic signal control, which uses “dynamic coordination” within zones of two or three closely-spaced intersections, with zones separated by segments long enough to serve as buffers [48]. Using dynamic rather than fixed cycle coordination makes the control logic flexible in order to give priority to frequent trams, with green waves for vehicles often simply following the tram, which is usually allowed to progress through with no delay at all. The followed logic in this study likewise assumes that coupled zones are small, and is therefore not suitable for tight downtown grids with many closely spaced intersections.

5.2.1 Coordination Logic at Coupled Intersections

Each coupled zone is governed by a critical intersection, the intersection with the longest natural cycle length. Determination of which intersection in the zone is critical is done adaptively based
on volume and saturation flow measurements over periods every cycle; volume-capacity (v/c) ratio is likewise measured adaptively by tracking green and red times.

Dynamic coordination within the zone aims to follow the critical intersection’s critical arterial through phase (“mainline” phase). When the v/c ratio at the critical intersection is below 0.9 (no risk of oversaturation), coordination logic aims for simultaneous green start for arterial phases in the coupled zone, which also provides good two-way progression as a result of short intersection spacing. A simultaneous green start results in a small amount of wasted green time, because the downstream mainline green phase is starved until vehicles arrive from the upstream intersection. Therefore, simultaneous start is only applied when the critical intersection’s v/c ratio is below 0.9, allowing the controller to utilize excess capacity to enhance two-way progression. Above 0.9 (the risk of oversaturation starts), it prioritizes progression in the mainline direction with offsets designed to prevent starvation and spillback at the critical intersection by making sure that the upstream intersection(s) release in time for the critical intersection to have a ready queue, and that the downstream intersections clear in time for the platoon arriving from the critical intersection (explained in detail in Chapter 6).

To achieve dynamic coordination, an “earliest activation time” is estimated for each arterial phase at the critical intersection at every phase transition based on the current signal state, commitments made by the local controller (e.g., minimum green and pedestrian clearance), and the minimum green needed to discharge the standing queue based on trap counts. A phase is “activated” when the clearance interval preceding its green is initiated. Earliest activation time accounts for the possibility that an intervening phase that is not on recall might be skipped, unless a call has been registered. The critical intersection communicates its earliest activation time to its neighboring intersections, who continue to pass it along peer-to-peer until the entire
coupled zone has been reached. When the critical intersection v/c is less than 0.9, this adjusted earliest activation time then becomes the “scheduled mainline activation time” for the non-critical intersections in the zone.

Normally, the non-critical intersections will be ready to activate their mainline before their scheduled activation time, since non-critical intersection naturally cycle faster than the critical intersection. Where the mainline phase is immediately preceded by a cross-street phase, that cross-street phase will be held green after it gaps out until the scheduled activation time (Figure 5-6). This allows the slack time forced by dynamic coordination to be better used, both in reducing delay to cross-street vehicles who would otherwise have to wait for the next cycle, and by reducing cross-street demand for the next cycle to help ensure that a non-critical intersection won’t become critical.

![Diagram](image)

**Figure 5-6: Holding Cross Street Phase (Shown in Dashed Lines) at a Non-Critical Intersection until Scheduled Mainline (EBT) Activation Time**

If the mainline phase is immediately preceded by an arterial opposing leading left (WBL) as given in Figure 5-7, using the slack time to hold this leading left (WBL) is not an efficient use of time unless it has an unusually high demand. Instead, the controller estimates the needed split for the left turn (WBL) using a queue count amplified by 10% (arbitrarily selected) to allow for
late arrivals, and subtracts it from the scheduled mainline activation time to schedule an activation time for the leading left (WBL). The cross-street phase preceding that leading left will then be held after gap-out until this scheduled activation time.

![Scheduled Mainline (EBT) Activation Time](image)

**Figure 5-7: Holding Cross Street at a Non-Critical Intersection based on the Estimated Needed Split for the Arterial Opposing Leading Left Turn Phase (WBL) to Meet the Scheduled Mainline (EBT) Activation Time**

When the mainline platoon is released at the first intersection (upstream) of a coupled zone, the controller holds the green signal at the following intersections until the released platoon has had enough time to clear the coupled zone (information goes from upstream to downstream signals). To illustrate using **Figure 5-8**, when the mainline green phase starts at the first intersection (INT 1), the controller prevents the mainline phase of next intersection (INT 2) from gapping out until the time that the mainline green phase gaps out at INT 1 plus the travel time between intersections INT 1 and INT 2. In a coupled zone with three intersections, the first intersection (INT 1) extends the second intersection (INT 2) and the second intersection (INT 2) extends the third, thus allowing the passage of the mainline platoon through the coupled zone.
5.2.2 Power of the Critical Intersection in a Coupled Zone

Because scheduled activations in a coupled zone are not guaranteed due to variability in the earliest activation time calculation, additional rules are developed to ensure dynamic coordination. For example, a cross street phase at a downstream signal in a coupled zone may want to continue beyond what was predicted based on trap counts due to queue length exceeding trap or due to slower discharge then predicted. In that case, a mainline platoon released early at the first intersection of a coupled zone (and this platoon could be large, since the first intersection in a coupled zone has sufficient queue storage capacity) may arrive at the next signal when mainline phase is red, causing queue spillback. This is especially important to avoid for any coupled signal downstream of the critical intersection, because discharge from the critical intersection will then be limited by spillback, reducing the critical intersection’s capacity.

In order to limit the described spillback phenomena, a critical intersection located upstream of any non-critical intersections in the same zone has the power to truncate the downstream non-critical intersections’ cross street green (subject to minimum green time, including pedestrian minimum). Therefore, if the first intersection in a zone is the critical intersection (determined adaptively), platoons released early at the first intersection should be

---

**Figure 5-8: Forward Communication in a Coupled Zone**
able to advance through the coupled zone without having to make stops by forcing non-critical intersections to start mainline green phase. This inhibits capacity reduction at the critical intersection, and should have little impact at non-critical intersections, because non-critical signals have slack capacity that allows them to clear the overflow vehicles (i.e., vehicles that could not be served in the previous cycle due to truncation) in the following cycles.

If non-critical intersections located downstream of the critical intersection start their mainline green phase before the critical intersection releases its platoon, they will hold their green signal to guarantee the passage of the mainline platoon released from the critical intersection. As a result, the power of critical intersection in a coupled zone assures dynamic coordination, preventing capacity reduction due to spillback.

5.2.3 Lead-Lag Phasing to Improve Coordination

At closely-spaced signals with protected left turn phases, staggering arterial through phases using lead-lag configuration (i.e., allowing a leading turn phase at one intersection and lagging at another) rather than starting simultaneously, improves two-way coordination. Using lead-lag configuration also limits wasted green time, because the green time that is wasted (starved) until vehicles arrive at the next intersection (the disadvantage of simultaneous green start for good two-way progression) can now be efficiently used by serving left turn phases. Christopherson and Riddle [49] showed that when intersections are closely-spaced, a policy allowing lead-lag configuration saves drivers time in a wide variety of coordinated systems over a policy allowing just leading sequences.

For good two-way coordination along the coupled zone, when intersections operate with protected left turn phases, it is desirable to have leading lefts where the mainline movement enters the coupled zone, and lagging lefts where it exits (Figure 5-9).
At the first intersection of a coupled zone (INT 1), leading lefts (EBL) clear left turning vehicles before the opposing through phase starts, which diminishes the spillback risk onto a through lane (EBT) from a turning lane (EBL) and helps increase platoon size and density at the first intersection (INT 1) of the coupled zone. Achieving a sizable platoon at the first intersection (INT 1) is important to get benefits that can outweigh the cost of signal timing adjustments that are necessary for coupled coordination. Using lagging lefts when exiting the group (INT 2) provides good coordination for left turning vehicles (EBL) as well, because it allows them to be served in the same cycle instead of waiting till the next cycle, as is the case with leading lefts.

One drawback of lead-lag configuration that may limit its implementation when left turn lanes operate under protected-permissive phasing (i.e., left turns use green arrow indication with right of way during protected phase and go with opposing through traffic and look for gaps to clear intersection during permitted phase) is the risk of yellow trap. Yellow trap occurs when a turning driver waiting for a gap sees the signal is turning red for its parallel through phase and expects that opposing through traffic is also getting a red signal. As a result, the turning driver

Figure 5-9: Lead-Lag Phasing along Coupled Intersections
may attempt to complete its turn while the opposing phase has still green signal, which can result in an angle collision. In order to overcome the problem of yellow trap and improve safety, flashing yellow can be used for permitted left turning vehicles when opposing through has a green signal to eliminate confusion [50].

**5.2.4 Secondary Extension Logic for Coupled Intersections**

Secondary extension logic for an arriving mainline platoon (critical direction) is applied only at the first intersection of a coupled zone, because the green termination for the mainline platoon within a coupled zone is determined by the first intersection (mainline green phase at downstream intersections in a coupled zone are terminated through forward communication from the first intersection). If a controller decides to grant a secondary extension at the first intersection when its mainline phase gaps-out, that information is passed to downstream signals to ensure that green will not be terminated at the downstream intersections until the arriving platoon has time to reach all intersections within the coupled zone in order to inhibit spillback.

Secondary extension logic for a mainline platoon follows the same logic as described in Section 5.1. Because coupled intersections cycle together and giving secondary extension at the first intersection leads to extending all intersections within the coupled zone, the affordable lost time calculation uses the v/c ratio of the critical intersection within the coupled zone. If the critical intersection can afford wasting green time (based on the available excess capacity) to enhance progression, then wasting green time at non-critical intersections for better coordination should not be problematic. Also, because the first intersection holds the green for the second and later intersections in a coupled zone, there is no need for secondary extension at the downstream signals.
Unlike the critical direction arterial phase, which uses the critical intersection’s v/c ratio in a coupled zone as a secondary extension criterion, secondary extension logic for a non-critical arterial phase employs only the local intersection’s v/c ratio (the intersection subject to secondary extension). Therefore, a secondary extension at one coupled intersection does not necessarily extend green at downstream coupled intersections for the non-critical direction.

The logic for secondary extension logic for a non-critical direction platoon in a coupled zone is the same as described previously, except that the calculation of $\Delta C_n$, that is, the difference between intersection’s natural cycle length and neighboring intersections’ natural cycle length, which is as follows:

- If the intersection is the first or last intersection of the coupled zone, cycle length comparison is between the coupled zone’s natural cycle length (taken as the critical intersection’s cycle length) and the cycle length of the neighboring uncoupled intersection.

- If the intersection is in the interior of a coupled zone (for coupled zones with more than two intersections), cycle length comparison logic compares the coupled zone’s cycle length and local intersection’s (i.e., the intersection which is subject to secondary extension for the non-critical direction) natural cycle length.
Chapter 6 : Control Logic for Oversaturated Arterials

On unsaturated arterials, coordination between traffic signals is focused toward providing good progression (green waves) between successive intersections, which helps reduce delay to through traffic. However, when an arterial has an intersection that is oversaturated (i.e., demand exceeds capacity) for more than a few signal cycles, long queues can develop, creating queue interactions that reduce capacity. Some of the more common names types of negative queue interaction are spillback, starvation, and turn bay overflow (when turning vehicles cannot all fit into a turn bay, and block a through lane). Queue interaction becomes more pronounced at closely-spaced signals due to the limited storage capacity on the segments between signals, and can occur even when an arterial is under-saturated. Capacity reductions due to queue interactions exacerbate oversaturation and can lead to significant delays. Because many control algorithms designed for under-saturated conditions do not account for these queue interactions, they often perform poorly when an arterial approaches capacity.

A standard way to deal with oversaturation in traffic signal control is to have longer cycle lengths in order to lower the proportion of the signal cycle that is lost due to clearance and start-up time associated with phase changes (typically four seconds per phase change). However, beyond about 100 seconds, the gains from longer cycle lengths are small and diminishing (lost time per hour varies inversely with cycle length, therefore diminishing benefit from longer cycles). Moreover, in practice, longer cycles often fail to yield greater capacity, because the formation of longer queues during the longer red period increases the extent of capacity-reducing queue interactions. Denney et al. [45] showed that when longer cycle lengths lead to queues extending beyond turning bays (even when there is no turning bay overflow), vehicles destined for a turn bay will be trapped in through lanes upstream of the start of the turn bay. Once traffic
starts flowing, these cars will turn out of a through lane into a turn bay, creating “holes” in the traffic flow. This is a form of starvation that results in a drop in the saturation flow rate once the green time exceeds the value needed to discharge the front part of the queue (the part reaching back to the start of the turn bay).

### 6.1 Research Papers on Oversaturation

Numerous researchers have examined the question of arterial control during periods of oversaturation. Many of them employ global optimization techniques. This section reviews some of the research papers to deal with oversaturation.

Lieberman et al. [51] developed a model called RT/IMPOST that uses mixed integer linear programming. The objective of RT/IMPOST is to maximize system throughput by avoiding spillback and starvation. The optimization is performed every cycle to calculate optimal signal offsets along the arterial. The control policy adjusts the arterial green phase durations every cycle to control and stabilize queue lengths and to provide equitable service to competing traffic streams. Simulation tests under RT/IMPOST showed a large reduction in delay compared to standard arterial control.

Hu et al. [52] developed a forward-backward procedure that adjusts green times along an arterial in order to mitigate oversaturation. The forward process aims to increase green times by searching for available green time which can be taken from side streets or conflicting phases to limit the extent of oversaturation along the arterial. The backward procedure “gates” traffic at intersections to prevent residual queue and spillback when available green time increased by the forward process is not sufficient. Green duration changes are performed based on the impacts on upstream and downstream intersections. Cycle length is fixed and unchanged, but offsets are dynamic in response to changes in green time allocation. The developed control strategies were
tested using simulation and their results showed a 12% reduction in average vehicle delay compared to standard arterial control.

Lämmer and Helbing [53] explore the idea of using a decentralized control with acyclic operation (i.e., variable phase sequences and no fixed cycle length). Their findings indicate that during oversaturation, control strategies should be governed by the availability of downstream storage space more than by the demand from an upstream queue. They proposed a bi-layer control strategy, with a supervisory layer designed to account for queue interactions. Simulation tests for an arterial near saturation showed reduced delays.

6.2 Control Policies Designed for Oversaturation

When an arterial is oversaturated, efficient signal control should focus on maintaining and utilizing capacity at bottlenecks in order to maximize throughput, limit the rate at which queues grow, and hasten recovery after the period of oversaturation ends.

This chapter describes the queue interactions and control policies proposed for self-organizing signal control during periods of oversaturation to maximize system capacity/throughput. The proposed policies for oversaturation include control logic on the intersection level as well as on the arterial level. The material in this section is summarized in Cesme et al. [54].

6.2.1 Efficient Actuated Control

The principle of actuated control is to hold a signal green only while its queue discharges, and assign right-of-way to a conflicting phase when flow drops below the saturation flow rate. This scheme generally helps minimize wasted green time and delay at isolated signals by keeping signal cycles as short as possible while preventing overflow (vehicles being left in the queue
when the green ends). Shorter cycles in turn result in shorter queues, creating less potential for queue interactions that reduce capacity and therefore help delay the onset of overcapacity. Actuated control is also ideal for recovering from oversaturation, as it helps ensure that queue dissipation is detected so that time can be used by phases that still have standing queues.

Control settings and detector configuration can make a difference in whether actuated signals achieve their desired switching behavior, as described in Chapter 3. Using upstream detection, short unit extensions (critical gap), and non-simultaneous gap-out for phases ending at a barrier improve the overall efficiency of an actuated intersection. Failing to use these features results in wasted green time which lengthens the signal cycle and therefore lengthens queues. In addition, the proposed multi-headway gap-out logic (Chapter 4) on multi-lane approaches increases the efficiency of an actuated operation by reducing wasted green time and cycle length.

The proposed self-organizing logic, which uses standard actuated control as a basis, included all of these features of actuated control, which help delay the onset of oversaturation and help hasten recovery from oversaturation.

### 6.2.2 Green Truncation in Response to Spillback

Spillback is caused when a queue from a downstream intersection uses up all the space on a link and prevents vehicles from entering the upstream intersection while a signal is green (Figure 6-1), resulting in reduction in intersection capacity.

When a discharging queue is blocked by a downstream queue that has spilled back to or near an upstream intersection, conventional actuation logic will continue to extend the green, resulting in capacity loss, longer cycles, and greater delay [55]. In order to prevent spillback from a downstream intersection, the proposed control logic uses spillback detectors located just downstream of an intersection, as shown in Figure 6-1.
Several aspects of the proposed queue management logic aim to prevent spillback. However, if it occurs and spillback detector occupancy reaches a certain threshold (which can vary depending on the detector length, speed limit, and target speed) while a through phase is green, the controller will truncate the green subject to minimum green and clearance time needs. If a through phase feeding the downstream queue is about to turn green, its green start will be delayed (and therefore the preceding phase held in green) until the spillback detector occupancy falls below the threshold. In the analyses, five seconds (arbitrarily selected) was used as the occupancy threshold with 10 foot long detectors.

At coupled signals where intersection spacing is low and intersections cycle together to maintain dynamic-coupled coordination, spillback rules will be activated only when there is a vehicle stopped on a spillback detector and the downstream signal is red. Because travel time between coupled intersections is short (e.g., 5 to 10 seconds), if the signal is green at the next intersection, the standing queue will start moving in a short time, and rules for coupled intersection ensure that the green phase at the downstream intersections will be held until...
vehicles clear the coupled zone, ensuring that queue spillback will last a short time without any significant impact on capacity. If the last intersection of a coupled zone ends its green due to max-out before fully serving a released platoon, spillback control rules still apply. If a queue reaches back to a spillback detector, it will force the upstream signal to terminate its green.

6.2.3 Inhibiting Secondary Extensions

As described in Section 5.1.2, when an arterial is below saturation, excess capacity can be used to improve progression by providing a secondary extension after a phase gaps-out in anticipation of an arriving platoon.

This tradeoff is part of the self-organizing logic proposed for under-saturated conditions, using a scheme in which the amount of green time that can be wasted per vehicle declines as the intersection (v/c) approaches 1.0 (Figure 5-4). When v/c reaches 1.0 at any given intersection, the need to preserve capacity dominates, and secondary extensions at that intersection are inhibited to avoid wasted green time and maximize intersection capacity. However, at intersections that are not themselves at capacity, secondary extensions can still be allowed.

6.2.4 Turn Pocket Spillback Prevention

Another type of queue interaction which results in capacity reduction during a queue discharge is turn pocket spillback. Turn pocket spillback occurs when the queue length for turning vehicles extends beyond the length of a turn bay (Figure 6-2). During oversaturation, pocket spillback can block an adjacent through lane, drastically reducing the capacity of the through movement once the elapsed green time for the through movement reaches the time needed to discharge the through queue stored between the pocket spillback point and the stopline.
This capacity reduction is of greatest concern when spillback affects a critical direction through phase. It is most likely to occur when traffic volumes approach or exceed capacity and cycle lengths, in response, become longer.

Two strategies are proposed for avoiding pocket spillback into a critical through lane, depending on whether the left turn movement is leading or lagging. These tactics are only applied when an intersection has been determined to be near or over capacity (intersection v/c exceeds 0.90) in order to avoid unnecessarily interfering with the signal cycle. Note that a left turn movement that could spill back into a critical through lane will not in itself be a critical movement, since it runs in parallel (not in conflict) with the critical through movement. Pocket spillback is identified when a spillback detector, placed at the entrance of a turn bay (Figure 4-8), reaches a threshold occupancy.

If the left turn movement prone to spillback is lagging (Figure 6-3a), detection of pocket spillback will trigger green truncation for its opposing through movement, subject to minimum green constraints, in order to begin the left turn movement immediately. To illustrate using Figure 6-3a, if the EBL queue starts overflowing into the adjacent EBT through lane, which is

---

**Figure 6-2: Turn Pocket Spillback**

This capacity reduction is of greatest concern when spillback affects a critical direction through phase. It is most likely to occur when traffic volumes approach or exceed capacity and cycle lengths, in response, become longer.

Two strategies are proposed for avoiding pocket spillback into a critical through lane, depending on whether the left turn movement is leading or lagging. These tactics are only applied when an intersection has been determined to be near or over capacity (intersection v/c exceeds 0.90) in order to avoid unnecessarily interfering with the signal cycle. Note that a left turn movement that could spill back into a critical through lane will not in itself be a critical movement, since it runs in parallel (not in conflict) with the critical through movement. Pocket spillback is identified when a spillback detector, placed at the entrance of a turn bay (Figure 4-8), reaches a threshold occupancy.

If the left turn movement prone to spillback is lagging (Figure 6-3a), detection of pocket spillback will trigger green truncation for its opposing through movement, subject to minimum green constraints, in order to begin the left turn movement immediately. To illustrate using Figure 6-3a, if the EBL queue starts overflowing into the adjacent EBT through lane, which is

---

**Figure 6-2: Turn Pocket Spillback**

This capacity reduction is of greatest concern when spillback affects a critical direction through phase. It is most likely to occur when traffic volumes approach or exceed capacity and cycle lengths, in response, become longer.

Two strategies are proposed for avoiding pocket spillback into a critical through lane, depending on whether the left turn movement is leading or lagging. These tactics are only applied when an intersection has been determined to be near or over capacity (intersection v/c exceeds 0.90) in order to avoid unnecessarily interfering with the signal cycle. Note that a left turn movement that could spill back into a critical through lane will not in itself be a critical movement, since it runs in parallel (not in conflict) with the critical through movement. Pocket spillback is identified when a spillback detector, placed at the entrance of a turn bay (Figure 4-8), reaches a threshold occupancy.

If the left turn movement prone to spillback is lagging (Figure 6-3a), detection of pocket spillback will trigger green truncation for its opposing through movement, subject to minimum green constraints, in order to begin the left turn movement immediately. To illustrate using Figure 6-3a, if the EBL queue starts overflowing into the adjacent EBT through lane, which is
the critical phase, control logic terminates the green for WBT (subject to minimum green time), and starts EBL early. Note that the opposing through movement (WBT in the example) is not a critical movement. Also note that identification of what phases are critical is done adaptively, updated every cycle, so that if repeated truncation begins to cause long queues on the opposing through movement, it can become critical and if so, will then be protected from truncation.

(a) Lagging Left and Leading Through

(b) Leading Left and Lagging Through

Notes: * Indicates the critical arterial through movement. Dashed line indicates the dynamic second realization of the left turn phase in case of pocket spillback.

Figure 6-3: Dual Ring with Protected Left Turn Phasing

If the left turn prone to spillback is leading, it is proposed adding a dynamic second realization of the left turn phase, shown as dashed line (Figure 6-3b). That is, if pocket spillback is detected after the leading left turn phase (EBL) but while the same direction’s through movement (EBT) is still running, the opposing through phase (WBT) will be truncated and the left turn (EBL) given a second realization as a lagging phase, as illustrated in Figure 6-3b. The lagging second realization or reservice will only be employed in cycles in which pocket spillback is detected, and only when intersection v/c exceeds 0.90.
The proposed logic for protecting against pocket spillback was tested in simulation at the junction of George Mason Drive and Columbia Pike in Arlington, Virginia. During the a.m. peak, when eastbound through is critical, the short left turn pocket on Columbia Pike eastbound often spills back into the adjacent through lane. Eastbound left turns are normally leading. Results from simulation (VISSIM) showed that the proposed self-organizing logic resulted in an increase in intersection throughput of 316 vphpl (vehicles per hour per lane) for all critical movements combined compared to standard coordinated-actuated control optimized using Synchro. By turning on and off different aspects of the proposed self-organizing logic, it was determined that dynamic second left turn realizations increased critical movement throughput by 183 vphpl (the remainder of the gain was due to using efficient actuated control in place of fixed-cycle coordination.) This is a substantial capacity gain, comparable to adding a second left turn lane, which in turn can be expected to lead to a large reduction in delay. In this test, left-turn reservice occurred in approximately seventy percent of the cycles.

6.2.5 Dynamic Offsets for Coupled Intersections

Where intersection spacing is too short to store the queues that can develop during a signal cycle, allowing intersections to cycle independently can lead to spillback and starvation. When intersections are oversaturated, spillback and starvation become even more critical because they reduce the capacity when it is most needed.

In order to provide the necessary coordination and prevent capacity reductions due to spillback and starvation, as reported before, it is proposed to use dynamic coordination for groups of closely spaced intersections, also called coupled zones. Intersections within a coupled zone cycle together with specified, dynamic offsets for their start of green, but without cycle length being specified.
Within each coupled zone, the critical intersection is identified based on the intersection that requires the greatest cycle length. This identification is updated regularly based on continuous volume measurements. When the critical intersection v/c ratio is below 0.90, the arterial phases in a coupled zone are scheduled to start simultaneously in order to provide good progression in both directions. Simultaneous green start logic, however, results in a small amount of wasted green time due to starvation, and therefore is only allowed when intersections have extra capacity.

When the critical intersection v/c exceeds 0.90, offsets are calculated to provide ideal progression (i.e., no capacity reductions at the critical intersection due to spillback and starvation) for the critical direction arterial phase (“mainline phase”). Ideal offsets for the mainline phase are calculated based on travel time between intersections, adjusted for the time needed to discharge the queue waiting at each signal. Queue size is determined using trap logic, with a pulse detector at the entry and exit of each road segment.

A phase is said to be activated the moment the change interval preceding it begins. Target offset (to prevent starvation at the critical intersection) for an intersection \( i \) upstream of the critical intersection \( j \) is given by

\[
\text{Offset}_{ij} = -TT_{ij} + \text{QueueLen}_{ij} HS_{at_{j}} + Y_{j} - Y_{i}
\]

where \( \text{Offset}_{ij} \) = difference when \( i \)'s mainline phase is activated and when \( j \)'s mainline phase is activated, \( TT_{ij} \) = travel time from \( i \) to \( j \), \( \text{QueueLen}_{ij} \) = number of vehicles stored between \( i \) and \( j \), and \( HS_{at_{j}} \) = saturation flow headway at \( j \), and \( Y_{i} \) = length of the change interval (yellow plus all-red) at \( i \). When \( \text{Offset}_{ij} \) is positive – likely because distance between coupled intersections is
small – mainline green at \( i \) should be activated after the critical intersection’s mainline green, making it easy for \( i \) to follow \( j \) in time. If, at intersection \( i \), the phase preceding the mainline gaps out before the scheduled mainline activation, the preceding phase is held in green (“holding extension”) until the scheduled mainline start. If the preceding phase has not yet gapped out when the mainline phase is scheduled to be activated, the preceding phase is not truncated; however, this should happen rarely because non-critical intersections tend to cycle faster than the critical intersection.

If \( \text{Offset}_{ij} \) is negative, scheduled activation at \( i \) is calculated as an offset from the earliest time \( j \)’s mainline green can be activated. This earliest activation time is calculated based on the existing signal state, queue counts for the intervening phases which are translated into estimated minimum green intervals by multiplying by the saturation headway, and commitments such as pedestrian clearance. Because queue counts are used to update minimum green each cycle, earliest activation time is a relatively good predictor of the actual activation time, especially during periods of oversaturation, when long queues make it such that minimum green becomes equal to maximum green.

If, in any cycle, \( j \)’s mainline green starts later than anticipated, there may be spillback as the queue advances from \( i \) without an open path, but such spillback will be temporary and can be tolerated because it doesn’t affect bottleneck capacity (spillback affects intersection \( i \), which is not the critical intersection). In such an event, the rule for truncating green in case of spillback is suspended until the downstream intersection has ended its anticipated green interval.

Target offset (to prevent spillback to the critical intersection) for an intersection \( k \) downstream of the critical intersection \( j \) is given by
so that the queue at \( k \) will be cleared just in time for the arrival of the platoon released from \( j \). If the offset is positive, \( k \)’s activation can easily follow \( j \)’s; if not, \( k \)’s activation time is scheduled as mainline \( j \)’s earliest possible activation time plus the (negative) offset. If \( j \) is activated later than expected, there will be some starvation at \( k \), but it can be tolerated because it doesn’t affect bottleneck capacity (i.e., the capacity of the critical intersection). Minor starvation at a downstream non-critical intersection is accepted to protect the critical intersection from spillback, just as minor spillback at an upstream non-critical intersection is accepted to protect the critical intersection from starvation.

### 6.2.6 Maximum Green Time

When an intersection is oversaturated, demand on the critical phases exceeds the intersection’s capacity to process them, and so vehicles have to be stored on those critical approaches or on the segments feeding them. Maximum green times will determine the relative share of queuing on the critical approaches. They can be set for various strategic reasons, such as which approach has more storage capacity before causing secondary queuing, or to keep streets used by transit from being blocked. Average delay will be smaller if queues are smaller on approaches with more lanes (therefore higher saturation flow); however, general attitudes toward equity limit the degree to which the public will tolerate one approach getting a long queue while another has a short one.

In the analyses, maximum green times (MaxGreen) were set as follows. First, a target maximum cycle length, \( C_{\text{Target}} \), was established \textit{a priori} (in the tests, it had a value of 105 seconds, which was arbitrarily selected). It is not enforced \textit{per se}, but is only used to determine
maximum green times (which make a maximum cycle length implicit). For single lane movements such as most left turns,

\[
MaxGreen = CTarget \times \frac{v}{s}
\]  

(6.3)

where \( v \) = volume and \( s \) = saturation flow rate. Both \( v \) and \( s \) are updated regularly (every cycle) based on detector measurements. For multilane approaches,

\[
MaxGreen = CTarget \times \frac{v}{s} \times \max(1, \frac{X}{XTarget})
\]  

(6.4)

where \( X \) is the movement’s degree of saturation, given by \( X = \frac{(v/s)}{(g/C)} \), where \( g \) is green time and \( C \) is cycle length, averaged over the last five cycles. \( XTarget \) equals 1.0 for the arterial mainline and 1.1 for other multilane approaches, giving a small degree of priority to the arterial mainline. If a movement’s upstream detector normally used to measure \( v \) is blocked by a standing queue for more than five seconds (which is an indication of very long queues), its \( X \) is assigned the value 1.2. The parameter values given in this section were chosen based on a judgment, and could be refined by performing sensitivity analysis.
Chapter 7: Self-Organizing Control Logic Applications

This chapter starts out with a discussion on microscopic simulation, and then it describes the case studies developed to test the strength of self-organizing control logic.

Microscopic traffic simulation models are computer programs which simulate the behavior and flow of traffic by representing and keeping track of the maneuvers of individual vehicles [56]. Microscopic traffic simulation models can represent more complicated situations (e.g., transit priority, incident conditions) and generally yield more precise and correct simulation results (if calibrated and validated properly to reflect site conditions) than macroscopic models, which utilize relationships between aggregate descriptors (e.g., concentration, average speed). Moreover, microscopic models are less expensive and more flexible than field implementation. As a result, microscopic traffic simulation programs are widely used by engineers and practitioners in assessing proposed traffic operation modifications.

In order to test the performance of the proposed self-organizing logic, case studies were developed using VISSIM, a microscopic simulation program for multi-modal traffic flow modeling developed by Planung Transport Verkehr (PTV) AG in Karlsruhe, Germany [44]. The key features of VISSIM which make it an ideal traffic simulation software in evaluating the effectiveness of self-organizing signals are:

- VISSIM allows changing the driver behavior parameters (e.g., reaction time, average standstill distance) to calibrate models to observed saturation flow rate and lost time.
- VISSIM provides a signal control application programming interface (API) for customized external signal control. The API allows for any kind of control strategies to be tested through the simulation including actuated control, adaptive traffic control, ramp metering, TSP, and signal preemption.
VISSIM is very effective in simulating transit behavior. It is able to model transit routes, various transit vehicle types, schedules, stops, number of passengers boarding, alighting, transit volume as well as dwell times. As a result, the impacts of signal control strategies on transit performance can be analyzed.

For testing the effectiveness of self-organizing traffic signals, several case studies were developed using VISSIM. Arterials with different characteristics (e.g., roadway geometry, mix of road user types) were studied in order to determine the robustness of self-organizing signals under different conditions. For example, one case study is an urban arterial with closely-spaced intersections, having high pedestrian activity and frequent transit service, while another one is an auto-dominated suburban arterial with more uniform intersection spacing making it more amenable to fixed cycle coordination.

The remainder of this chapter is divided into five sections. The next two sections describe the case studies used to test under-saturated arterials with self-organizing traffic signals. They are followed by the application of self-organizing control logic to oversaturated arterials in order to test the control logic during periods of oversaturation. The last section focuses on testing self-organizing logic on an arterial which experiences both under-saturated and oversaturated traffic environment.

7.1 Under-Saturation Conditions – Beacon Street – Test Case I

Beacon Street is an east-west arterial in the state of Massachusetts (MA), running through several towns/cities including Newton, Brookline, Brighton, and Boston. The modeled segment of Beacon Street falls within the limits of Town of Brookline, MA (Figure 7-1).
Figure 7-1: Study Corridor along Modeled Beacon Street, Brookline, Massachusetts

The modeled Beacon Street corridor is approximately 1.3 miles long and has twelve signalized intersections. It is bordered by Brandon Hall transit stop on the west and the Park Drive/Beacon Street intersection on the east. The corridor also serves a light rail transit (LRT) line, the “C” branch of Massachusetts Bay of Transportation Authority’s (MBTA’s) Green Line, which runs in a dedicated median along Beacon Street. The modeled corridor has six transit stops in both inbound (eastbound) and outbound (westbound) directions.

Weekday peak hour traffic volumes and weekday existing signal timing plan data were obtained through a study conducted by Vanesse Hangen Brustlin (VHB), Inc., in 2009. Transit
schedule and route data as well as passengers boarding and alighting data at transit stops were obtained from the MBTA and from field studies.

AM peak period was modeled in the simulation. During the modeled AM peak, the signals operate with coordinated-actuated control, as described in Section 2.3. With the existing signal timing plans, intersections operate on a 90 second cycle (note, however, self-organizing logic was not compared to the existing timing plan, but to an optimized plan with 80 second cycle). The LRT “C” line runs at seven minute headways during the AM peak period on weekdays.

7.1.1 Network Implementation in VISSIM

VISSIM allows using an orthographic image as a background to construct a network. The model layout was based on a map from Massachusetts’ Office of Geographic Information (MassGIS), a Statewide Resource for Geospatial Technology and Data.

In VISSIM, while non-transit vehicles enter the network based on a Poisson distribution, introducing a stochastic variation of input flow arrivals, transit vehicles enter based on the schedule provided by the user. However in reality, there is considerable variability in transit arrival time. In order to introduce a random arrival process for transit vehicles, a “dummy” transit stop was created at the beginning of each transit route. The dwell time distribution at the dummy stops followed a normal distribution with mean of 420 seconds in both directions (i.e., seven minute headway), a standard deviation of 100 seconds for inbound trains and 300 seconds for outbound trains (headway standard deviations were obtained from MBTA’s automatic vehicle identification (AVI) system). The reason for the greater variation for outbound trains is that the modeled corridor begins at the tenth for outbound direction; while it is the sixth stop for inbound direction, and headway variability tends to increase along a transit route.
In VISSIM, boarding times at each transit stop can be simulated by specifying passenger arrival rates and a unit boarding time. Passenger arrivals follow a Poisson process. Alighting rates are defined by specifying the percentage of passengers on the transit vehicle that will get off at each stop. The user also specifies a unit alighting time.

Arrival and alighting rates at each transit stop were taken from MBTA’s passenger boarding and alighting counts performed in 2006 and 2010 (some stops have 2006 counts and some have 2010 counts). Dwell time determination followed the maximum dwell time method, which uses the maximum of boarding time need and alighting time need plus a clearance time which accounts for the time needed for the transit vehicle to stop and to open/close the doors. The dwell time calculations at a stop is given as follows:

\[
\text{Dwell Time} = \text{Max} (\text{ons} \times \text{unit boarding time}, \text{offs} \times \text{unit alighting time}) + \text{Clearance Time}
\] (7.1)

where unit boarding time (i.e., boarding per person) was taken as three seconds per car, and unit alighting time (per car) and clearance time were taken as two seconds. These values were obtained through field analyses (field analyses at Harvard stop and St.Paul stop) where the dwell time and number of passengers boarding and alighting were observed.

### 7.1.1.1 Impact of Overcrowding on Dwell Time

Overcrowding on transit vehicles increases dwell time of passengers due to slower unloading and loading. To consider the impact of crowding on dwell time, the *Transit Capacity and Quality of Service Manual* (TCQSM) suggests using a half-second penalty per passenger (a bilinear function) [57]. However, Milkovits [58] showed that crowding impact increases exponentially and the crowding effect may be non-linear. To address the crowding impact, once trains reach a
load of 118 passengers, which is half of crush load, that is, the maximum passenger capacity of a transit vehicle, as given by the MBTA [59], the unit boarding time was assumed to increase non-linearly from three seconds to six seconds. However, it was assumed that crowding does not have any significant impact on unit alighting time since passengers can use multiple doors to get off a train. Therefore, the impact of crowding on alighting was not considered in dwell time impact calculation. Figure 7-2 shows unit boarding time as a function of passenger loads on trains, taken to be the arriving load minus offs plus half the number of boarding passengers.

![Unit Boarding Time vs Train Passenger Load](image)

**Figure 7-2: Change in Unit Boarding Time (Seconds per Passenger per Car) as a Function of Passenger Load on a Train**

VISSIM can keep track of passenger flows for each transit line and stop; however it does not allow the user to specify variable unit boarding and alighting time as a function of passenger load, making it impossible to model the impact of overcrowding directly. To overcome this, dummy traffic signals were modeled only for trains, located just downstream of transit stops, which turn red to stop trains for a time equal to the extra dwell time incurred by overcrowding.
7.1.1.2 Varying Demand Conditions and Development of Control Logic

Traffic counts for the morning peak show the critical intersection operating at 79% of capacity. In order to test model performance under varying demanding conditions, the volume profile for the arterial and cross streets was scaled up and down in the network. The transient demand profile is given in Figure 7-3.

Figure 7-3: Volume to Capacity Ratio for Beacon Street at the Critical Intersection over Time

Following the first 15-minute period of warm-up time, the volume profile had periods of low demand (the first and the last 15 minutes), moderate demand (79% of capacity), and heavy demand (93% of capacity) for 30 minutes, though the demand was still lower than the intersection capacity (to only test under-saturated conditions).

Signal control logic for self-organizing signals was coded in C++ and interfaced to VISSIM through its API. At the end of every time step (0.1 seconds), the simulation informs the control program of detector status, which then determines whether any signals should be
changed, and passes that information to the simulation program, which then advances vehicles for the next time step.

### 7.1.2 Determination of Coupled Intersections

The first step towards applying self-organization logic on an arterial is identifying the sets of coupled intersections that will have dynamic coordination. They should be the intersections that have limited queue storage capacity due to short block spacing. Table 7-1 lists the signalized intersections along the corridor (excluding the mid-block pedestrian crossing signals) and their spacing, and indicates the ones which were coupled together. In the end, there are three coupled zones consisting of two or three intersections: Winchester – Webster, Harvard – Pleasant – Charles, and Hawes – Carlton.

**Table 7-1: Arterial Intersection Spacing and Coupled Intersections along the Simulated Segment of Beacon Street, Brookline, Massachusetts**

<table>
<thead>
<tr>
<th>Segments</th>
<th>Segment Length (ft)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marion St – Winchester St</td>
<td>660</td>
</tr>
<tr>
<td>Winchester St&lt;sup&gt;a&lt;/sup&gt; – Webster St&lt;sup&gt;a&lt;/sup&gt;</td>
<td>280</td>
</tr>
<tr>
<td>Webster St – Harvard St</td>
<td>505</td>
</tr>
<tr>
<td>Harvard St&lt;sup&gt;b&lt;/sup&gt; – Pleasant St&lt;sup&gt;b&lt;/sup&gt;</td>
<td>350</td>
</tr>
<tr>
<td>Pleasant St&lt;sup&gt;b&lt;/sup&gt; – Charles St&lt;sup&gt;b&lt;/sup&gt;</td>
<td>280</td>
</tr>
<tr>
<td>Charles St – St.Paul St</td>
<td>530</td>
</tr>
<tr>
<td>St.Paul St – Kent St</td>
<td>850</td>
</tr>
<tr>
<td>Kent St – Hawes St</td>
<td>970</td>
</tr>
<tr>
<td>Hawes St&lt;sup&gt;c&lt;/sup&gt; – Carlton St&lt;sup&gt;c&lt;/sup&gt;</td>
<td>540</td>
</tr>
</tbody>
</table>

**Notes:** Superscripts a, b, and c indicate the three coupled zones.

In determining the sets of coupled intersections, the rule of thumb followed was to couple when intersection spacing is less than 500 ft and not to couple when spacing is greater than 600 ft. For the range 500-600 ft, discretion was used depending on the number of
intersections involved and the likely cycle lengths (which affect queue lengths). For example, even though the spacing between Charles – St.Paul and Webster – Harvard intersections is 505 and 530 ft, respectively, they were not included in the coupled zone in order to keep the coupled zone no more than three intersections (Zurich’s experience has also led them to favor coordination zones of two to three intersections). If we were to include Webster and Harvard, which has approximately 500 ft of queue storage (can store about 20 vehicles), the number of intersections in the zone would increase to five, something that would not be desirable because of the challenges with respect to achieving hard coordination of five intersections. The two main challenges in coupling five intersections are to maintain the scheduled activation time (i.e., the time when the arterial phase at the critical intersection is ready to be green) at all intersections and to provide good two-way progression along coupled intersection due to longer travel time. It should be noted that a scenario which included Webster and Harvard in the coupled zone was also tested (resulting in a coupled zone with five intersections); it resulted in higher average network delay.

7.1.3 Transit Signal Priority (TSP)

The MBTA LRT “C” line trains running along Beacon Street experience large delays and headway variation during weekday peak hours, which slows transit operations and increases passenger waiting time and crowding. As a result, the relative attractiveness of transit is reduced, which has the potential to increase automobile trips. In order to reduce the LRT delay at signalized intersections as well as improve service reliability, it was desirable to apply conditional priority, which gives priority at signalized intersections to late vehicles (i.e., transit vehicles running behind the schedule), but not to early vehicles (i.e., transit vehicles running ahead of schedule) to directly correct early and late schedule deviations [60].
With near side transit stops, TSP operations become more challenging than with far side stops due to the uncertainty of transit arrival time at an intersection. If a controller wants to trigger priority before a transit vehicle departs from a stop, dwell time estimation is required to predict the arrival time of transit vehicles at the intersection. However, high uncertainty in dwell time poses problems in accurately predicting the vehicle’s arrival time, causing poor performance of TSP operations. Therefore, with near side stops, it is typical to have detectors located downstream of the stop so that priority is requested when a transit vehicle leaves a stop. But this limits priority tactics to those with a very short lead time. For example, green extension cannot be used if transit vehicles cannot be detected in advance.

In order to test how well self-organizing signals perform under priority interruptions, all transit stops were modeled as far-side (the real traffic network includes both far side and near side stops) in order to be able to apply green extension strategy in an aggressive manner.

7.1.3.1 TSP Application for Beacon Street

TSP tactics for LRT vehicles included green extension, early green, phase rotation, and phase insertion. The maximum allowed green extension for an arriving transit vehicle was selected as 15 seconds (using check-in detectors to call for priority that were placed approximately 650 feet upstream of stop line, corresponding to approximately 15 seconds of travel time). In the case of intersection spacing less than 650 feet (for example the Webster-Harvard segment where travel time between intersections is approximately 11.5 seconds), the maximum green extension time is the travel time allowed by intersection spacing.

In giving priority to LRT vehicles, headway-based conditional priority was used, which compares the time headway between the train requesting priority and the train before it to a minimum headway threshold. Scheduled headway for both directions is 420 seconds (seven
minutes). For inbound trains (i.e., peak direction), priority was granted if the headway was greater than or equal to 315 seconds, while the headway threshold for outbound trains (i.e., non-peak direction) was 180 seconds. The motivation for the less restrictive condition for outbound trains was that the outbound direction is dominated by alightings, so that headway irregularity does not worsen crowding, and holding to reduce waiting time (by improved regularity) helps only a few boarding passengers while adding riding time to the many passengers that are already on board. Therefore, for outbound trains, the goal was not to improve service regularity, but to speed up the service as much as possible in the outbound direction. The reason for using a small headway threshold of 180 seconds instead of unconditional priority was to prevent bunching of trains. For inbound trains, however, crowding is a problem, and so service regularity was an objective of signal priority. Therefore, priority was not granted when headway was lower than 315 seconds (5.25 minutes, or 75% of scheduled headway) to retard early trains and to improve regularity.

7.1.4 Random Interruptions

Because the Beacon Street LRT is not in conflict with the arterial phases, priority interventions for LRT (e.g., green extension and early green) can actually benefit arterial traffic. Therefore, in order to further test the system’s ability to heal from interruptions that take time away from the arterial, the transit priority case in the simulation also includes granting a 20 second extension for cross-street traffic on Harvard Street, starting when the cross street gaps out, once every six minutes, as might occur from giving priority to a crossing bus route. In case of a conflicting call (i.e., LRT calls for priority at the same time that Harvard Street calls for interruption), priority was given to LRT.
### 7.1.5 Simulation Results and Analysis

In order to test the model performance, five simulation runs for each self-organizing and coordinated-actuated control were performed, with results reporting averages for the 90-minute period following the warm-up. Results from self-organizing signals (no TSP scenario and TSP with including cross street interruptions at Harvard Street intersection scenario) were compared to a coordinated-actuated operation. Coordinated-actuated control used timing parameters optimized using Synchro (optimized cycle length is 80 seconds while cycle length in the field is 90 seconds) with manual adjustments to offsets based on observing the simulation.

Simulation results showed that without transit priority, self-organizing signals reduce overall vehicular delay by 14% compared to an optimized coordinated-actuated scheme (Table 7-2). There was no appreciable change in transit delay (approximately one second increase in average delay per train per intersection compared to the optimized coordinated-actuated plan).

When conditional signal priority was applied, the self-organizing signals were able to reduce transit delay by more than 10 seconds per train per intersection, which is slightly more than 50% of the delay that occurs under coordinated-actuated logic. This is a substantial reduction, especially considering that only 69% of trains (those behind schedule) requested priority. When absolute priority was applied, average delay reduction for transit was 13.4 seconds per intersection, a reduction of approximately 70% of their non-priority delay.

At the same time, the self-organizing logic showed its self-healing properties by keeping traffic delay within 3% of the optimized fixed cycle coordination plan for both conditional and absolute priority cases. Results for the conditional signal priority case include 20 seconds interruptions every six minutes at the Harvard Street intersection, while the absolute priority case does not.
Table 7-2: Simulation Results for Beacon Street With and Without Transit Signal Priority

<table>
<thead>
<tr>
<th></th>
<th>Optimized Coordinated-Actuated</th>
<th>Self-Organizing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No TSP</td>
<td>Absolute TSP</td>
</tr>
<tr>
<td>Average Network Delay (s/vehicle)</td>
<td>68.4 (0.0%)</td>
<td>74.0 (8.2%)</td>
</tr>
<tr>
<td>Inbound Train Delay (s/train)</td>
<td>197.8</td>
<td>148.5</td>
</tr>
<tr>
<td>Inbound Train Delay Change (s/train/intersection)</td>
<td>-</td>
<td>-4.9</td>
</tr>
<tr>
<td>Outbound Train Delay (s/train)</td>
<td>206.1</td>
<td>124.5</td>
</tr>
<tr>
<td>Outbound Train Delay Change (s/train/intersection)</td>
<td>-</td>
<td>-7.8</td>
</tr>
<tr>
<td>Percent of Trains Requesting Priority (only late trains request priority)</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Inbound train headway coefficient of variation at first stop in modeled network</td>
<td>0.426</td>
<td>0.426</td>
</tr>
<tr>
<td>Inbound train headway coefficient of variation at last stop in modeled network</td>
<td>0.580</td>
<td>Not Measured</td>
</tr>
<tr>
<td>Average Cycle Length (s) during base period (v/c = 0.79)</td>
<td>80.0</td>
<td>80.0</td>
</tr>
</tbody>
</table>

1. Cross street interruptions at Harvard Street intersection was only considered under Conditional TSP scenario. Absolute TSP did not include cross street interruptions.
2. Values in parentheses indicate percentage change in average network delay compared to optimized coordinated-actuated plan without TSP case.

Equally important for transit is the large reduction in headway coefficient of variation (CV) for inbound (peak direction) service at the downstream end of the study segment. Without priority, headway CV tends to increase along a transit line as late trains, due to the increase over the usual number of passengers wanting to board, leading to crowding and longer passenger waiting times. Results show headway CV remaining almost constant along the segment when...
conditional priority is applied, rather than growing from 0.43 to 0.58 (Table 7-2) without priority under both self-organizing and coordinated-actuated control. Figure 7-4 shows headway CV for inbound trains just before arriving at each stop under different control scenarios. As trains advance in the peak direction, one can see how conditional priority improves the reliability of trains by keeping the headway CV from growing.

Figure 7-4: Headway Coefficient of Variation (CV) for Inbound Trains With and Without Conditional Transit Signal Priority

During the period with existing traffic volumes ($t = 30$ to 45 minutes, traffic demand used for Synchro optimization), the average network cycle length, which was 80 seconds under coordinated-actuated control, is only 69.2 seconds under self-organizing control, showing a
strong benefit to pedestrians. Average cycle lengths at the study intersections with self-organizing signals under varying demand conditions are presented in Table 7-3.

Table 7-3: Average Cycle Lengths by Intersection along Beacon Street under Self-Organizing Logic for each Demand Profile

<table>
<thead>
<tr>
<th>Intersection</th>
<th>Time Period (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15-30</td>
</tr>
<tr>
<td>Marion St</td>
<td>67.5</td>
</tr>
<tr>
<td>Winchester St³a</td>
<td>63.8</td>
</tr>
<tr>
<td>Webster St³a</td>
<td>64.2</td>
</tr>
<tr>
<td>Harvard St³b</td>
<td>65.5</td>
</tr>
<tr>
<td>Pleasant St³b</td>
<td>65.8</td>
</tr>
<tr>
<td>Charles St³b</td>
<td>65.7</td>
</tr>
<tr>
<td>StPaul St</td>
<td>69.5</td>
</tr>
<tr>
<td>Kent St</td>
<td>63.5</td>
</tr>
<tr>
<td>Hawes St³c</td>
<td>60.2</td>
</tr>
<tr>
<td>Carlton St³c</td>
<td>61.2</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>64.7</strong></td>
</tr>
</tbody>
</table>

Notes:
1. Cycle length for coordinated-actuated control was 80 seconds at all intersections.
2. Each superscripts (a,b,c) indicates a coupled zone.

Results indicated that self-organizing signals yielded lower average cycle length under all demand periods. It is also worth mentioning that average cycle lengths at coupled intersections are very close to each other (differences are usually within one second, and never greater than 4.1 seconds) and vary for different time periods, which verifies that the coordination of traffic within a coupled zone is dynamic and strong.
7.2 Under-Saturation Conditions – Rural Road – Test Case II

The second test bed was a segment of Rural Road, a north-south arterial in Tempe, Arizona, a suburb of Phoenix. The modeled segment includes nine signalized intersections. Figure 7-5 displays the aerial view of Rural Road as well as the study intersections.

![Figure 7-5: Study Corridor along Modeled Rural Road, Tempe, Arizona to Test Self-Organizing Signals along an Arterial That is Well-Suited to Fixed Cycle Coordination](image)

Turning movement counts at the intersections and signal timing information were obtained from Maricopa Association of Governments (MAG). Unlike Beacon Street, Rural Road
has little traffic other than automobiles (few pedestrians and little transit service), and its intersection spacing is more uniform. Except for the last segment, the spacing between intersections is greater than 1300 ft, making it well-suited to fixed cycle coordination. The last segment has an intersection spacing of 595 ft. This test allowed for a comparison of self-organizing logic to fixed cycle coordination where the latter works relatively well due to uniformly and widely spaced intersections.

Rural Road corridor was analyzed under three different scenarios. The first scenario assumed no pedestrian movements (therefore no pedestrian minimum green time constraints) and no signal priority for buses. The second scenario tested the performance of self-organizing signals with pedestrians and signal priority for arterial buses. In the third and last scenario, self-organizing signals were tested along an idealized Rural Road corridor with ideally spaced intersections that allows perfect two-way coordination. To make it ideal for two-way coordination, the geometric layout of Rural Road was slightly modified, which will be explained in details in the following sections.

In all three scenarios, in order to test the algorithms for periods with volume below capacity, a.m. peak traffic volumes were scaled down to make the v/c ratio at the critical intersection equal to 0.90. Simulation results were obtained for a 60 minute period of uniform demand following a 15 minute warm up period, with averages reported from five simulation runs. In the first two scenarios, the arterial was modeled with one coupled zone, which included E Baseline Rd and E Minton Dr intersections, with signal spacing of 595 feet. The remaining seven signals were uncoupled. The idealized Rural Road corridor was modeled without coupled zones due to well-spaced intersections (ideally spaced intersections).
7.2.1 Rural Road: No Pedestrians and No Signal Priority for Buses

Timing plans under self-organizing signals were developed without considering pedestrians and signal priority for buses. Self-organizing control logic was compared to a coordinated-actuated operation. Timings for coordinated-actuated control were optimized for the adjusted volumes (i.e., v/c ratio at the critical intersection equal to 0.90) using Synchro; they include an 80 seconds cycle length except at E Carver Rd, where the cycle length is 40 seconds.

Simulation results indicated that self-organizing signals reduced overall vehicular delay by 7.1% compared to an optimized actuated-coordinated plan (from 49.5 seconds to 46.0 seconds). This example shows that self-organizing signals can achieve performance at least comparable to that obtained using fixed cycle coordination even when the arterial has intersection spacing well suited to fixed cycle coordination. In addition, with self-organizing signals, the average cycle length (simple average over intersections) is approximately 20% lower than with coordinated-actuated control (61 versus 76 seconds), indicating a reduction in pedestrian delay and better potential for TSP. Table 7-4 provides average cycle lengths at the study intersections under self-organizing logic and the change in cycle length compared to optimized fixed cycle plan. Cycle lengths at coupled intersections are demonstrated by thick border.
Table 7-4: Average Cycle Length at the Study Intersections along Rural Road under Self-Organizing Logic

<table>
<thead>
<tr>
<th>Intersection</th>
<th>Cycle Length(s)</th>
<th>Change in Cycle Length (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E Warner Rd</td>
<td>62.6</td>
<td>-17.4</td>
</tr>
<tr>
<td>E Carver Rd</td>
<td>39.9</td>
<td>-0.1</td>
</tr>
<tr>
<td>E Elliot Rd</td>
<td>60.4</td>
<td>-19.6</td>
</tr>
<tr>
<td>E Bell De Mar Dr</td>
<td>47.0</td>
<td>-33.0</td>
</tr>
<tr>
<td>E Guadalupe Rd</td>
<td>80.0</td>
<td>-0.0</td>
</tr>
<tr>
<td>E Westchester Dr</td>
<td>48.0</td>
<td>-32.0</td>
</tr>
<tr>
<td>E South Shore Dr</td>
<td>51.0</td>
<td>-29.0</td>
</tr>
<tr>
<td>E Baseline Rd</td>
<td>78.6</td>
<td>-1.4</td>
</tr>
<tr>
<td>E Minton Dr</td>
<td>78.3</td>
<td>-1.7</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>60.6</strong></td>
<td><strong>-15.0</strong></td>
</tr>
</tbody>
</table>

At three major intersections, self-organizing signals had an average cycle length very close to the coordinated cycle length of 80 seconds, while most of the other intersections had cycle lengths between 47 and 63 seconds. Average cycle lengths at the coupled intersections (E Baseline Rd and E Minton Dr) are almost identical, indicating that they cycle synchronously. Of special interest is that the E Carver Rd intersection’s average cycle length was 39.9 seconds, comparable to its 40 seconds cycle length with double cycling under coordinated-actuated control. This demonstrates the ability of self-organizing control to find and exploit double cycling opportunities organically.

To further analyze how coordination mechanisms work under self-organizing signals, signal state changes (red and green intervals) for the critical as well as the non-critical arterial direction at four intersections on Rural Road – two major intersections (E Guadalupe Rd and E Baseline Rd) with two minor intersections in between (E Westchester Dr and E South Shore Dr) – are recorded (Figure 7-6). In the coordinated plan, all four run with an 80 s cycle. With self-
organizing control, it can be seen how the cycle length and the green interval length vary within a given intersection as well as between intersections. Diagonal solid lines show the leading edge of green bands at the progression speed in the critical (northbound) direction and diagonal dashed lines show the leading edge of the green bands in the non-critical (southbound) direction. The frequency and extent of green bands show how the self-organizing method leads to partial synchronization while allowing less busy intersections to cycle at a faster pace.

Notes: C* indicates average cycle length.

Figure 7-6: Signal State Changes and Green Bands for the Critical (Northbound, Shown by Diagonal Solid Lines) and Non-Critical (Southbound, Diagonal Dashed Lines) Arterial Direction for four Intersections on Rural Road under Self-Organizing Control Logic
7.2.2 Rural Road: With Pedestrians and Signal Priority for Arterial Buses

Rural Road was also modeled with pedestrians and signal priority for buses traveling along the arterial. No priority was considered for cross-street buses. Pedestrian signals were modeled push-button actuated (i.e., pedestrian phases are only called in case of a pedestrian detection).

Self-organizing signals were compared to an optimized coordinated-actuated operation. The optimized cycle length under coordinated-actuated control with minimum green time for pedestrians is 100 seconds at all intersections except at E Carver Rd, where it double cycles and the cycle length is 50 seconds. Priority tactics for arterial buses under coordinated-actuated operation included green extension and early green, while self-organizing signals applied green extension, early green, and phase rotation. The maximum allowed green extension for an arriving bus was selected as 15 seconds. An aggressive early green was applied for conflicting phases. Through phases were shortened to ten seconds unless pedestrian movement was actuated. Left turn phases were shortened to six seconds. With phase rotation, leading left turn was typically changed to a lagging left turn phase in order to expedite the passage of buses, which would use the lagging through phase without phase rotation. All bus stops were modeled as far-side. Note that all TSP scenarios described in the following sections along Rural Road also assumed 15 seconds of maximum green extension and far side stops.

Simulation results showed that without signal priority to arterial buses, self-organizing signals reduced average network delay by 8.9% compared to coordinated operation (Table 7-5). However, self-organizing signals resulted in higher average number of stops and average arterial delay. The results clearly indicate how fixed cycle coordination can reduce through vehicle delay along an arterial, yet at the cost of increasing cycle length at non-critical intersections, causing higher delays for non-coordinated movements, which may also lead to higher network delay.
With self-organizing control logic, average cycle length was reduced by approximately 28 seconds (about 30% reduction), which in turn reduced average pedestrian delay by more than 25%, and reduced delay to cross-traffic. Both self-organizing logic and coordinated-actuated operation without priority yielded comparable bus delay.

When priority was applied under self-organizing logic, TSP resulted in dramatic reduction in arterial bus delay compared to coordinated-actuated plan no TSP scenario. Bus delay was reduced more than 50% in the northbound and the southbound direction (approximately 12 seconds reduction per bus per intersection), while for private traffic, network delay was still 7.5% lower than under coordinated-actuated control without TSP.

**Table 7-5: Simulation Results for Rural Road With and Without Transit Signal Priority (With Actuated Pedestrian Phases)**

<table>
<thead>
<tr>
<th></th>
<th>Optimized Coordinated-Actuated</th>
<th>Self-Organizing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No TSP</td>
<td>TSP</td>
</tr>
<tr>
<td><strong>Average Network Delay (s/vehicle)</strong></td>
<td>54.4</td>
<td>56.7 (-4.6%)</td>
</tr>
<tr>
<td><strong>Average Number of Stops</strong></td>
<td>1.25</td>
<td>1.28 (2.6%)</td>
</tr>
<tr>
<td><strong>Average Arterial Delay (s/vehicle)</strong></td>
<td>114.3</td>
<td>113.8 (-0.4%)</td>
</tr>
<tr>
<td><strong>Northbound Bus Delay (s/bus)</strong></td>
<td>198.4</td>
<td>100.7 (-49.2%)</td>
</tr>
<tr>
<td><strong>Northbound Bus Delay Change (s/bus/intersection)</strong></td>
<td>-</td>
<td>-10.9</td>
</tr>
<tr>
<td><strong>Southbound Bus Delay (s/bus)</strong></td>
<td>188.2</td>
<td>114.4 (-39.2%)</td>
</tr>
<tr>
<td><strong>Southbound Bus Delay Change (s/bus/intersection)</strong></td>
<td>-</td>
<td>-8.2</td>
</tr>
<tr>
<td><strong>Average Cycle Length (s)</strong></td>
<td>92.9</td>
<td>92.9 (0.0%)</td>
</tr>
<tr>
<td><strong>Pedestrian Delay (s)</strong></td>
<td>45.1</td>
<td>44.7 (-0.9%)</td>
</tr>
</tbody>
</table>

1. Values in parentheses indicate percentage change in average network delay compared to optimized coordinated-actuated plan without TSP case.
7.2.3 Idealized Rural Road with Ideally Spaced Intersections

In order to test the performance of self-organizing control logic along an arterial with ideally spaced intersections (i.e., uniform intersection spacing, well-suited for two-way coordination), an idealized corridor was created by slightly modifying the geometric layout of Rural Road. The idealized Rural Road corridor includes seven signalized intersections where travel time on all segments between intersections is 40 seconds (intersection spacing of 0.5 miles). Figure 7-7 displays the idealized Rural Road corridor as well as the study intersections.

![Figure 7-7: Ideally Spaced Intersections along the Hypothetically Created Rural Road Corridor](image-url)
Compared to existing Rural Road corridor, E Westchester Dr and E Minton Dr were removed from the simulation model, and E Bell De Mar Dr and E South Shore Dr were moved 560 ft and 400 ft south, respectively to obtain ideal intersection spacing. Scaled down a.m. peak traffic volumes (the v/c ratio at the critical intersection equal to 0.90) were used. However, because two intersections were eliminated in the model, dummy origin and destination links were created between intersections to obtain balanced turning movement counts.

The idealized Rural Road corridor was first tested excluding pedestrians and signal priority for buses. Later scenarios included pedestrians and signal priority for buses traveling along the arterial as well as priority for cross street buses.

7.2.3.1 Hypothetical Rural Road: No Pedestrians and No Signal Priority for Buses

Self-organizing signals were compared to coordinated-actuated control and standard fully actuated control. Timings for coordinated-actuated control were optimized for the adjusted volumes using Synchro. The optimized cycle length at E Carver Rd was 40 seconds. All other intersections operated under 80 second cycle length. Actuated control was modeled with upstream detection, simultaneous gap-out and single headway detection (i.e., traditional detection for multi-lane approaches).

Simulation results indicated that even when intersection spacing is ideal that allows perfect two-way coordination under fixed cycle coordination, self-organizing signals resulted in only 1.8% increase in average network delay (Table 7-6). Fully actuated operation increased network delay by approximately 10% compared to coordinated-actuated control due to the lack of progression for through vehicles.
Table 7-6: Simulation Results for the Hypothetical Rural Road: No Pedestrian Minimum Green Constraints and No Signal Priority for Buses

<table>
<thead>
<tr>
<th></th>
<th>Optimized Coordinated-Actuated</th>
<th>Self-Organizing</th>
<th>Fully Actuated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Network Delay(s/vehicle)</td>
<td>41.5</td>
<td>42.2 (1.8%&lt;sup&gt;1&lt;/sup&gt;)</td>
<td>45.3 (9.4%&lt;sup&gt;1&lt;/sup&gt;)</td>
</tr>
<tr>
<td>Average Number of Stops</td>
<td>1.06</td>
<td>1.20 (13.2%)</td>
<td>1.22 (14.9%)</td>
</tr>
<tr>
<td>Average Arterial Delay (s/vehicle)</td>
<td>54.2</td>
<td>94.9 (75.1%)</td>
<td>134.7 (148.5%)</td>
</tr>
<tr>
<td>Average Cycle Length (s)</td>
<td>74.3</td>
<td>56.5 (-24.0%)</td>
<td>63.1 (-15.1%)</td>
</tr>
</tbody>
</table>

1. Values in parentheses indicate percentage change in average network delay compared to optimized coordinated-actuated plan without TSP case.

Compared to self-organizing signals, coordinated-actuated control reduced average number of stops (13% reduction) and arterial vehicle delay (75% reduction) substantially through longer cycle length and green time for arterial phases. However, the improved progression for arterial vehicles requires a longer cycle length (74 vs. 56 s) and causes much greater delay for non-coordinated traffic (e.g., cross-street traffic and pedestrians). The net effect on vehicles is essentially no change in overall delay, while the shorter cycle and inherent flexibility create a background far better for serving pedestrians and transit.

Results also indicated that self-organizing signals improve progression for through vehicles compared to fully actuated control by providing secondary green extension for arriving platoons. Compared to fully actuated operation, self-organizing logic reduced arterial delay from 134.7 seconds to 94.9 seconds, and also reduced overall vehicular delay by about 8%.

Finally, the reason for lower average cycle length under self-organizing signals than the actuated control is that self-organizing signals employed non-simultaneous gap-out and multi-headway detection for multi-lane approaches, which limits wasted green time and lowers average cycle length.
7.2.3.2 Idealized Rural Road: With Pedestrians and Signal Priority for Arterial Buses

The idealized Rural Road corridor was also tested considering pedestrians and signal priority for buses traveling along the arterial. No priority was considered for cross-street bus routes. Pedestrian signals were modeled as push-button actuated.

Self-organizing logic was compared to an optimized coordinated-actuated control. Signal priority tactics under coordinated-actuated operation included green extension and early green. Signal priority logic under self-organizing signals considered two scenarios. The first scenario applied green extension and early green to create exactly the same priority logic under coordinated-actuated operation, while the second one applied green extension, early green, and phase rotation to evaluate the impact of aggressive priority tactics.

Simulation results indicated that when no priority was considered, self-organizing signals reduced average network delay by 9.0% compared to the coordinated-actuated control (Table 7-7). However, similar to the previous findings, self-organizing signals resulted in a greater average number of stops (1.21 vs. 1.15), greater arterial vehicle delay (117 vs. 89 seconds), and greater northbound (critical direction) bus delay (167 vs. 126 seconds). The reason for lower bus delay with fixed cycle coordination in the northbound direction is that long cycle lengths and long green times for arterial phases offered by the fixed cycle coordination also benefit buses traveling through the corridor.
### Table 7-7: Simulation Results for the Hypothetical Rural Road: With Pedestrian Minimum Green Constraints and With Signal Priority for Buses Traveling along the Corridor

<table>
<thead>
<tr>
<th></th>
<th>Optimized Coordinated-Actuated</th>
<th>Self-Organizing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No TSP</td>
<td>TSP: GE, EG</td>
</tr>
<tr>
<td><strong>Average Network Delay (s/vehicle)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>49.9 (8.0%)</td>
<td>53.8</td>
<td>45.4 (-9.0%)</td>
</tr>
<tr>
<td><strong>Average Number of Stops</strong></td>
<td>1.15</td>
<td>1.19 (4.0%)</td>
</tr>
<tr>
<td><strong>Arterial Delay (s/vehicle)</strong></td>
<td>89.1</td>
<td>86.9 (-2.5%)</td>
</tr>
<tr>
<td><strong>Northbound Bus Delay (s/bus)</strong></td>
<td>126.1</td>
<td>83.8 (-33.5%)</td>
</tr>
<tr>
<td><strong>Northbound Bus Delay Change (s/bus/intersection)</strong></td>
<td>-</td>
<td>-6.0</td>
</tr>
<tr>
<td><strong>Southbound Bus Delay (s/bus)</strong></td>
<td>178.1</td>
<td>105.8 (-40.6%)</td>
</tr>
<tr>
<td><strong>Southbound Bus Delay Change (s/bus/intersection)</strong></td>
<td>-</td>
<td>-10.3</td>
</tr>
<tr>
<td><strong>Average Cycle Length (s)</strong></td>
<td>92.9</td>
<td>92.9 (0.0%)</td>
</tr>
<tr>
<td><strong>Pedestrian Delay (s)</strong></td>
<td>44.7</td>
<td>43.9 (-1.8%)</td>
</tr>
</tbody>
</table>

GE: Green Extension  
EG: Early Green  
PR: Phase Rotation  

1. Values in parentheses indicate percentage change in average network delay compared to optimized coordinated-actuated plan without TSP case.

Application of green extension and early green under both coordinated-actuated and self-organizing control resulted in dramatic reductions in average bus delay (more than 20% and 40% reduction for northbound and southbound buses, respectively). However, while priority under coordinated-actuated operation caused 8.0% increase in average network delay, self-organizing signals were still able to reduce average network delay by 8.2%.

Another major finding was that the application of aggressive TSP (i.e., green extension, early green, and phase rotation) with self-organizing signals reduced bus delay approximately seven seconds per intersection (38% reduction) for northbound buses and 16 seconds per intersection (63% reduction) for southbound buses. Average network delay was also reduced by 7.0% compared to coordinated-actuated scheme without TSP.
Finally, application of self-organizing logic (both with and without TSP) reduced average cycle length by approximately 32%, which in turn resulted in 25% to 30% reduction in average pedestrian delay.

7.2.3.2 Idealized Rural Road: With Pedestrians and Signal Priority for Cross Street Buses

Priority interruptions for arterial buses also benefit arterial traffic and as a result, it may not create significant disruption to signal coordination. Therefore, in order to further test the self-organizing signals’ ability to heal from interruptions, the idealized Rural Road corridor was also tested with pedestrians and TSP for cross-street buses only (i.e., no priority was applied for buses traveling along the arterial).

Priority for cross-street buses was applied at two intersections; E Elliot Rd (with intersection v/c ratio of 0.78) and E Guadalupe Rd (with intersection v/c ratio of 0.90). Self-organizing signals were compared to an optimized coordinated-actuated operation. Priority logic under coordinated operation included green extension and early green, while self-organizing signals applied green extension, early green, and phase rotation incrementally, as described in the previous section. The maximum allowed green extension for buses was assumed as 15 seconds. Through competing phases were allowed to be truncated to ten seconds in case of no pedestrian actuation and left turn phases were shortened to six seconds. At each intersection with cross street priority, bus headway was selected as ten minutes both in the eastbound direction and the westbound direction.

Simulation results showed that self-organizing signals outperformed coordinated-actuated operation under all scenarios (Table 7-8). When no priority was considered, average network delay was reduced by 9% under self-organizing control compared to coordinated operation. Fixed cycle coordination resulted in longer network cycle length (92.9 vs. 62.6) and higher delay
for cross-street traffic, which also increased the delay for buses on cross-street. Results indicated that when no TSP was applied, self-organizing signals reduced bus delay by 22.2% at E Elliot Rd and by 5.0% at E Guadalupe Rd compared to coordinated-actuated control. The reason for a larger benefit at E Elliot Rd is that it is a non-critical intersection and the cycle length at E Elliot Rd with self-organizing signals is much lower than the coordination cycle (64 vs. 100 seconds), which results in considerable less delay for cross-street vehicles. The cycle length at E Guadalupe Rd (the critical intersection) under self-organizing logic was 82.6 seconds, closer to the coordination cycle of 100 seconds. As a result, the delay reduction for buses at E Guadalupe Rd was less pronounced.

Table 7-8: Simulation Results for the Hypothetical Rural Road: With Pedestrian Minimum Green Constraints and With Signal Priority for Cross Street Buses Only

<table>
<thead>
<tr>
<th></th>
<th>Optimized Coordinated-Actuated</th>
<th>Self-Organizing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No TSP</td>
<td>TSP: GE, EG</td>
</tr>
<tr>
<td>Average Network Delay (s/vehicle)</td>
<td>50.0</td>
<td>51.4 (2.9%)</td>
</tr>
<tr>
<td>Average Number of Stops</td>
<td>1.15</td>
<td>1.17 (2.0%)</td>
</tr>
<tr>
<td>Arterial Delay (s/vehicle)</td>
<td>89.3</td>
<td>102.4 (14.7%)</td>
</tr>
<tr>
<td>E Elliot Rd Bus Delay</td>
<td>30.1</td>
<td>16.9 (-44.0%)</td>
</tr>
<tr>
<td>E Guadalupe Rd Bus Delay</td>
<td>37.2</td>
<td>13.8 (63.0%)</td>
</tr>
<tr>
<td>Average Cycle Length (s)</td>
<td>92.9</td>
<td>92.9 (0.0%)</td>
</tr>
<tr>
<td>Pedestrian Delay (s)</td>
<td>44.7</td>
<td>45.1 (0.9%)</td>
</tr>
</tbody>
</table>

GE: Green Extension  EG: Early Green  PR: Phase Rotation

1. Values in parentheses indicate percentage change in average network delay compared to optimized coordinated-actuated plan without TSP case.
Application of conventional priority tactics (i.e., green extension and early green) under self-organizing control logic resulted in more than 60% of reduction in bus delay compared to coordinated operation without TSP with no major impact to general traffic (average network delay was also reduced by 7.3%). When aggressive priority was applied (i.e., green extension, early green, and phase rotation), self-organizing signals reduced bus delay to approximately seven seconds per intersection (about 75% reduction in bus delay) without causing any significant impact to non-transit traffic.

The other important finding was the dramatic reduction in average cycle length under self-organizing control logic. Results indicated how fixed cycle coordination increases average cycle length in order to maintain coordination. With self-organizing signals under all scenarios, the average cycle length was reduced by approximately 32%, which led to more than 25% reduction in average pedestrian delay.

7.3 Oversaturation Conditions – Artificial Network – Test Case III

The first test bed used to evaluate the proposed oversaturation control logic was the artificial network used by Lieberman et al. to test their method of arterial control, RT/IMPOST [51]. The network has seven irregularly spaced intersections, spaced approximately 250 to 1200 ft apart. Figure 7-8 shows the layout of VISSIM simulation model and the hourly approach volumes for cross street at each intersection.
The simulation model was calibrated to match the saturation flow rate of 1800 vehicles per hour per lane (vphpl) used by Lieberman et al., (private communication). Simulations ran for 120 minutes after a 10 minute warm-up period. Inflow volumes for a “moderate flow” period are given in [51], for which v/c at the critical intersection is 0.76; all volumes were scaled up or down for different periods using the transient demand profile shown in Figure 7-9. It has periods of low, moderate, and heavy flow, then rises to 128% of capacity for 30 minutes before dropping to recovery periods with 76% and 38% of capacity, respectively.

Simulation results using self-organizing logic are given in Figure 7-10, where they are compared to published results for RT/IMPOST and coordinated-actuated control using timings from three standard timing packages, PASSER, TRANSYT, and Synchro [51]. Self-organizing
results were determined as the average of five replications. The coefficient of variation of vehicle delay was 0.068 during the oversaturation period (from 45-75 minutes), 0.043 during the recovery period (from 75-105 minutes), and below 0.150 in other periods when average delay was low.

Figure 7-9: Volume / Capacity Ratio at the Critical Intersection versus Time for the Test Arterial in [51]
Note: Source for all except self-organizing control is [51].

**Figure 7-10: Network Delay (vehicle-hours) by Period for Five Signal Control Models**

Both self-organizing control and RT/IMPOST perform much better during the oversaturated period and the recovery periods, a sign of better preserving and using capacity through better queue management. The self-organizing method has greater delay than RT/IMPOST during the oversaturation period, but has less delay during the recovery periods and overall. Overall, self-organizing logic reduced total network delay by more than 45% compared to PASSER, TRANSYT, and Synchro, and approximately 4% compared to RT/IMPOST. The relative parity of the self-organizing logic with RT/IMPOST shows good promise, since RT/IMPOST uses complex optimization calculations while self-organizing logic uses relatively simple first-generation rules that are amenable to refinement and continual improvement.
7.4 Oversaturation Conditions – Beacon Street and Rural Road – Test Case IV

For the Beacon Street and the existing Rural Road (i.e., not the idealized Rural Road) test beds, oversaturation tests were performed by scaling volumes up so that during one period the critical intersection’s v/c ratio exceeds capacity (Figure 7-11). Results using self-organizing logic were compared to coordinated-actuated control optimized using Synchro for the period of oversaturation. (Tests of coordinated-actuated control made using timing parameters optimized for lower flow yielded even worse results.) All results are the average of five replications.

![Figure 7-11: Volume / Capacity Ratio at the Critical Intersection versus Time for Beacon Street (solid line) and Rural Road (dashed line)](image)

Simulation results of Beacon Street indicated that the self-organizing method has far better performance in all periods, including the oversaturation and recovery periods, reflecting the relative success of the self-organizing method at preserving and utilizing capacity (Table 7-9). Cycle length for coordinated control was 120 seconds.
Table 7-9: Comparison of Total Vehicle Delay for Beacon Street by Time Period for Self-Organizing Control versus Coordinated-Actuated Control Optimized for the Period of Oversaturation (30-60 minutes)

<table>
<thead>
<tr>
<th>Time Period (min)</th>
<th>15-30</th>
<th>30-60</th>
<th>60-90</th>
<th>90-105</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delay (veh-hr), coordinated-actuated</td>
<td>72.8</td>
<td>258.3</td>
<td>223.0</td>
<td>55.6</td>
<td>609.7</td>
</tr>
<tr>
<td>Delay (veh-hr), self-organizing</td>
<td>31.7</td>
<td>185.6</td>
<td>149.9</td>
<td>20.8</td>
<td>388.0</td>
</tr>
<tr>
<td>Percentage Change</td>
<td>-56.50%</td>
<td>-28.10%</td>
<td>-32.80%</td>
<td>-62.50%</td>
<td>-36.40%</td>
</tr>
</tbody>
</table>

For Rural Road, self-organizing logic reduced total delay by 7.7%, from 342.1 veh-hrs with coordinated-actuated control to 315.8 veh-hrs. This result suggests that even when an arterial has longer intersection spacing that is amenable to fixed cycle coordination and less prone to queue interaction, fixed cycle coordination can be compromised in favor of a more flexible, decentralized logic without increasing delay.

For further insight into the operation of self-organizing logic, average cycle lengths at the Rural Road intersections are given, by period, in Table 7-10. During the period of oversaturation, at the critical intersection, the cycle length used by self-organizing control is close to that suggested for coordinated control by Synchro (100 s); but elsewhere and during other periods, the self-organizing method uses substantially shorter cycles. Similarity between some neighboring pairs of intersections reveals a considerable amount of informal synchronization. The first two intersections listed offer an instructive example. In higher demand periods, E Warner Rd has almost double the average cycle length as E Carver Rd, indicating that most of time they coordinate by having E Carver Rd double cycle relative to the E Warner Rd; but during the last, low demand period, they have nearly the same cycle length, indicating that in almost every cycle, E Carver Rd waits for E Warner Rd so that they cycle at the same rate.
Table 7-10: Average Cycle length (s) by Intersection and by Period along Rural Road

<table>
<thead>
<tr>
<th>Intersection</th>
<th>15-30</th>
<th>30-60</th>
<th>60-90</th>
<th>90-105</th>
</tr>
</thead>
<tbody>
<tr>
<td>E Warner Rd</td>
<td>56.4</td>
<td>75.8</td>
<td>59.8</td>
<td>47.9</td>
</tr>
<tr>
<td>E Carver Rd</td>
<td>33.5</td>
<td>41.6</td>
<td>48.5</td>
<td>48.2</td>
</tr>
<tr>
<td>E Elliot Rd</td>
<td>54.9</td>
<td>69.9</td>
<td>59.0</td>
<td>50.2</td>
</tr>
<tr>
<td>E Bell De Mar Dr</td>
<td>46.8</td>
<td>56.9</td>
<td>50.6</td>
<td>42.7</td>
</tr>
<tr>
<td>E Guadalupe Rd²</td>
<td>77.3</td>
<td>95.2</td>
<td>83.7</td>
<td>56.4</td>
</tr>
<tr>
<td>E Westchester Dr</td>
<td>49.3</td>
<td>57.4</td>
<td>52.7</td>
<td>39.4</td>
</tr>
<tr>
<td>E South Shore Dr</td>
<td>56.3</td>
<td>58.8</td>
<td>55.0</td>
<td>49.8</td>
</tr>
<tr>
<td>E Baseline Rd³</td>
<td>82.0</td>
<td>88.3</td>
<td>79.6</td>
<td>62.6</td>
</tr>
<tr>
<td>E Minton Dr³</td>
<td>80.6</td>
<td>86.6</td>
<td>77.5</td>
<td>65.2</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>59.7</strong></td>
<td><strong>70.1</strong></td>
<td><strong>62.9</strong></td>
<td><strong>51.4</strong></td>
</tr>
</tbody>
</table>

**Notes:**
1. Cycle length for coordinated-actuated control was 100.0 seconds at all intersections.
2. Critical intersection
3. Intersections at Baseline and Milton were coupled.

### 7.5 Oversaturation Conditions – Columbia Pike – Test Case V

The last test bed was Columbia Pike, an arterial that runs east-west through Arlington County in Northern Virginia (Figure 7-12). The modeled segment is approximately 3.9 miles long, bordered by Carlin Springs Rd on the west and Washington Boulevard on the east. The modeled corridor covers 19 signalized intersections.
Figure 7-12: Study Intersections along Modeled Columbia Pike, Arlington, Virginia

Columbia Pike is similar to Beacon Street arterial with irregularly and closely spaced intersections (signalized intersection spacing varies between approximately 315 to 1750 feet) as well as heavy pedestrian and transit activity. The primary difference is that transit service is provided by buses, which operate in mixed traffic rather than in reserved lanes. The corridor also exhibits the highest ridership of any bus corridor in Virginia, with average weekday ridership of approximately 16,000 boardings per day [61].

The simulation model of the corridor (developed in VISSIM), traffic volumes, and transit data were provided by AECOM’s Arlington office, in Virginia. AECOM modeled Columbia Pike...
Pike for the morning and evening peak periods as part of Columbia Pike Transit Initiative project. Only the morning peak was analyzed with self-organizing control logic.

Three coupled zones were used along the arterial; Carlin Springs Rd and S Jefferson intersections (520 feet apart), Four Mile Run Dr and Buchannan Street (315 feet apart), and Wayne Street and Courthouse Street (545 feet apart).

Traffic counts for the morning peak show the critical intersection operating at 94% of capacity. To test self-organizing logic’s performance under varying demanding conditions, the volume profile was scaled up and down in the network to create a transient demand profile, with v/c ratio at the critical intersection for each demand period are shown in Figure 7-13. The critical intersection has low traffic volumes in the first 30-minute period (following 15-minute of warm-up), then demand increases to near-capacity (94% of capacity), and oversaturation period is experienced for 30-minutes. Following oversaturation, recovery periods start with 94% and 56% of capacity, respectively.

![Figure 7-13: Volume to Capacity Ratio for Columbia Pike at the Critical intersection over Time](image-url)
Eight simulation runs (the calculated coefficient of variation for the simulation results with five runs were more than 10% so the number of runs were increased to eight), for self-organizing and coordinated-actuated control were run, with results reporting averages for the 150-minute period following the 15-minute warm up time. Coordinated-actuated control used timing parameters optimized using Synchro (cycle length in the field is 130 seconds, while the optimized cycle length is 116 seconds) with manual adjustments to splits and offsets based on observing the simulation. This optimization was performed based on the existing traffic volumes (i.e., with v/c ratio at the critical intersection is 94%).

7.5.1 Transit Service along Columbia Pike

Along Columbia Pike, 24 different bus routes serve the corridor in the morning peak, including cross-street buses as well as arterial bus routes, all operating in mixed traffic, with far-side, near-side, and mid-block stops. The majority of buses (approximately 60 to 65%) travel in the peak direction, that is, eastbound direction. Figure 7-14 (not drawn to scale) provides bus frequency (i.e., the number of buses per hour) in the morning peak. The George Mason intersection has the highest bus frequency in the morning peak (43 buses per hour). All intersections along Columbia Pike experience more than 10 buses per hour during the morning peak.
With that many number of buses, providing TSP plays a key role in improving transit operations (e.g., reducing passenger riding time and waiting time), which in turn can increase the relative attractiveness of transit.

### 7.5.2 Transit Signal Priority (TSP) along Columbia Pike

The TSP strategies considered include priority only for peak arterial direction buses (i.e., eastbound) and priority for both peak and non-peak arterial (i.e., westbound) direction buses. No priority was considered for cross-street bus routes. Priority was applied in an adaptive manner, in which no priority was allowed at intersections when intersection v/c ratio is over 0.90. To evaluate the benefits of aggressive TSP, a new set of simulations was performed in which intersection v/c ratio threshold to inhibit priority was set as 1.0.
7.5.2.1 Priority Logic for Peak Direction Bus Routes

For peak direction buses, applied priority strategies included early red, early green and green extension. Early red is different than the conventional TSP strategies (e.g., green extension and early green) in the sense that it forces the controller to turn the signal red for the street with the bus (i.e., “bus street”) when there is a near side stop so that green returns to the bus street by the time the bus is ready to advance [62]. This strategy not only reduces bus delay, it also prevents lane blockage (most buses stop on travel lane along Columbia Pike) by forcing the signal to turn red when there is a bus serving its passengers. The priority logic for peak direction buses in case of a near-side stop is described as follows:

- Approaching buses are detected at a user-specified distance upstream of the stopline (typically about 15 seconds of upstream of the stop line)
- If the bus street has a green signal, the controller projects the expected time that a bus will enter the intersection using the following equation:

\[
E[\text{ArrivalTime}] = t_L + d_{acc} + d_{dec} + E[\text{dwell}] + TT \tag{7.1}
\]

where \(t_L\) is the lost time due to door opening and closing (assumed as two seconds), \(d_{acc}\) and \(d_{dec}\) are the acceleration and deceleration delay, respectively, \(E[\text{dwell}]\) is the expected dwell time at the near side stop based on historical data, and \(TT\) is the travel time from the upstream detector to the stopline.

- If the 10th percentile projected arrival time is higher than the predicted green end time (Figure 7-15A), early red for the bus street is applied and green signal is terminated as soon as minimum green constraint allows. 10th percentile arrival time is calculated using
expected arrival time (equation 7.1), a historic standard deviation provided by the transit agency, and normal distribution. If standard deviation is not provided, it is assumed that standard deviation for dwell time is 20% of average dwell time.

Note that the priority logic also considers the vehicles ahead of a bus and holds the bus street green for them to clear before applying early red so that the bus can reach the near-side stop and start serving its passengers.

Following an early red decision, if the bus finishes serving its passengers before the cross-street has turned yellow, the green phase for the cross-street is terminated (subject to minimum green times). This means applying early green for the bus street in addition to early red. “Cross-street” refers any competing phases, including arterial turning phases and cross-street phases. No competing phases are skipped unless a vehicle is not detected (i.e., regular skipping logic in actuated operation).

![Diagram](image)

A: Early Red  
B: Green Extension  
C: Do Nothing  

a: Arrival time.

**Figure 7-15: Priority Logic for Peak Direction Bus Routes in case of a Near-Side Stop**
• If the 10th percentile projected arrival time is smaller than the predicted green end time, early red is not applied. In that case, if the 90th percentile arrival time is smaller than the predicted green end plus a maximum allowed green extension time (Figure 7-15B), then green for the bus street is extended. If the 90th percentile arrival time is greater than the predicted green end plus the maximum green extension time (Figure 7-15C), no alteration is made to the signal timing plan.

• When a bus is detected while traffic on another approach is being served and if the predicted bus arrival time is smaller than the earliest green start time for the bus street, an early green strategy is applied. If the predicted arrival time is higher than the earliest green start time for the bus street, no alteration is made to the signal timings. However, if the signal for the bus street is still red when a bus finishes serving its passengers, early green is still applied to allow the green signal to be returned as quickly as possible.

While applying the green extension, the controller provides a generous maximum allowed green extension (20 seconds, which was arbitrarily selected) for the bus street. However, this generous green extension is granted only when it is very likely that buses benefit from the extended green. When dwell time variability is high at a near-side stop (resulting in a high 90th percentile arrival time), the controller is less willing to provide a green extension, because buses may not clear the intersection during the extension interval due to long dwell times, which results in wasted green time and inefficient TSP.

When there is no near-side stop, the controller applies standard green extension and early green for peak direction arterial buses.
7.5.2.2 Priority Logic for Non-Peak Direction Bus Routes

Priority strategies considered for non-peak direction (westbound) bus routes include early green and green extension. Approaching buses are detected at a user-specified distance when there is no near-side bus stop. In the case of a near-side bus stop, detectors are located downstream of the bus stop; therefore priority is requested when a bus finishes serving its passengers. This type of priority is very limited as the green phases can only be extended for a few seconds (depending on how close the near-side bus stop is to the stop line), undermining the efficiency of TSP.

For example, for the non-critical direction along Columbia Pike, at three intersections, a near-side bus stop is located at the stop line, and at nine intersections, the setback distance for near-side stop is less than 225 feet, allowing maximum green extension of approximately five seconds.

7.5.1.3 Priority Logic in case of Multiple Calls

If two buses traveling in the same direction request priority, signal timings are changed to accommodate priority for the first bus. Therefore, the developed logic is not able to provide priority for two buses traveling in the same direction.

When two calls are simultaneously received from eastbound and westbound buses, if the two phases end their green together at the barrier, green extension is based on the bus that requires longer green time. For example, if a peak-direction bus needs 14 seconds of extension, while a non-peak direction bus asks for seven seconds of green extension, the green is extended by 14 seconds, allowing both buses to clear an intersection.

7.5.3 Simulation Results

Simulation results under optimized coordinated-actuated operation and self-organizing logic, both analyzed with and without TSP, are provided in Table 7-11. Under self-organizing priority
scenarios, 0.90 was used as the intersection v/c ratio threshold to inhibit priority. To evaluate aggressive TSP scenario, self-organizing signals were also tested with v/c ratio threshold of 1.0. With coordinated-actuated operation, no TSP was considered at the George Mason intersection. Because this intersection carries significant cross-street traffic, and thus has very little extra green time, any reduction in green time on the cross street to accommodate priority creates significant congestion. TSP under self-organizing logic overcomes this shortcoming by not allowing priority once an intersection v/c ratio exceeds the pre-defined threshold.

**Table 7-11: Simulation Results for Columbia Pike With and Without Transit Signal Priority**

<table>
<thead>
<tr>
<th></th>
<th>Coordinated-Actuated</th>
<th>Self-Organizing: v/c limit = 0.90</th>
<th>Self-Organizing: v/c limit = 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No TSP</td>
<td>TSP</td>
<td>No TSP</td>
</tr>
<tr>
<td><strong>Network Delay</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(s/vehicle)</td>
<td>110.7</td>
<td>129.1</td>
<td>96.9</td>
</tr>
<tr>
<td><strong>Network Bus Delay</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(s/bus)</td>
<td>262.4</td>
<td>241.9</td>
<td>290.8</td>
</tr>
<tr>
<td>(s/bus/int)</td>
<td>21.1</td>
<td>19.4</td>
<td>23.4</td>
</tr>
<tr>
<td><strong>Peak Direction</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bus Delay</td>
<td>335.8</td>
<td>308.8</td>
<td>351.8</td>
</tr>
<tr>
<td>(s/bus)</td>
<td>18.5</td>
<td>17.0 (-7.9%)</td>
<td>19.3 (4.3%)</td>
</tr>
<tr>
<td>(s/bus/int)</td>
<td>273.8</td>
<td>245.7</td>
<td>355.0</td>
</tr>
<tr>
<td><strong>Non-Peak</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direction Bus Delay</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(s/bus)</td>
<td>14.1</td>
<td>12.5 (-11.3%)</td>
<td>18.6 (31.9%)</td>
</tr>
<tr>
<td>(s/bus/int)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** Values in parentheses indicate percentage change in average network delay compared to optimized coordinated-actuated plan without TSP.
Without priority to buses, self-organizing signals reduced average network delay by 12.5% compared to optimized coordinated actuated operation. However, network bus delay increased by 2.3 seconds per intersection (10.9%) with self-organizing logic. The increase in bus delay with self-organizing signals can be attributed to the coordinated-actuated plan which offers long cycle lengths and long green times for arterial phases, which also benefits buses traveling through the corridor.

When priority was applied only for peak arterial direction buses with self-organizing logic, peak direction buses experienced 5.1 seconds reduction in delay per intersection, which resulted in a 10% reduction in bus delay (2.1 seconds per intersection) in the network. For the general traffic, peak direction priority caused hardly any impact compared to self-organizing no TSP scenario. Self-organizing signals were still able to reduce network delay by 12.2% compared to coordinated-actuated scheme. The ability to give buses priority without adding delay to general traffic can be attributed to the early red strategy, which prevents lane blockages while the signal is green.

When priority was applied in both arterial directions, TSP resulted in considerable reduction in network bus delay (18.0% or 3.8 seconds reduction per intersection) with negligible impacts to general traffic (approximately 10% reduction compared to coordinated-actuated plan). This is a significant improvement, because providing priority when transit frequency is high typically causes large delays for general traffic along a fixed-cycle signalized arterial. In contrast, simulation results indicated that priority with the coordinated-actuated plan reduced bus network delay by only 1.7 seconds per intersection (-8.1%), while increasing delay to general traffic by 16.6% compared to the optimized coordinated-actuated plan without TSP. This result confirms the findings the Dion study along Columbia Pike [63]. Their simulation results
indicated that providing five seconds of green extension as well as early green in the morning peak reduced bus delay by approximately one second per intersection (approximately 3% reduction in bus delay), but resulted in 9% increase in car delay.

Finally, priority logic with self-organizing signals was tested with the v/c ratio threshold (i.e., the threshold that inhibits priority) increased from 0.90 to 1.0. Results showed that average bus delay in the network was further reduced, resulting in a 4.1 seconds (19.4%) decrease per intersection compared to actuated-coordinated plan. In terms of impact to general traffic, self-organizing signals still performed better than actuated-coordinated plan (5.0% reduction in delay), but with more delay than when using the lower v/c threshold.

The other important finding was the dramatic reduction in average cycle length under self-organizing signals without TSP compared to the optimized coordinated-actuated plan (Table 7-12). The results are based on the entire simulation period. Columbia Pike is a great example of how fixed cycle coordination forces non-critical intersections to have long cycles in order to maintain coordination. With self-organizing signals, the average cycle length was reduced dramatically from 116.0 seconds to 68.7 seconds (-40.1%), resulting in a significant reduction in pedestrian delay. At the critical intersection (S George Mason), cycle length with self-organizing logic is close to the cycle length under coordinated-actuated operation (108.8 vs. 116.0 seconds). Neighboring minor intersections (i.e., S Thomas St and S Quincy St), however, double cycle with average cycle lengths close to 54 seconds. Elsewhere, groups of intersections coordinate with each other at cycle lengths around 65 seconds or 75 seconds. At coupled intersections, cycle lengths are close, though not identical, indicating that coupled intersections were allowed to double cycle.
Table 7-12: Average Cycle Length under Self-Organizing Signals (No TSP) at the Study Intersections along Columbia Pike

<table>
<thead>
<tr>
<th>Intersection</th>
<th>Cycle Length (s)</th>
<th>Change in Cycle Length(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carlin Springs Rd(^a)</td>
<td>100.2</td>
<td>-15.8</td>
</tr>
<tr>
<td>S Jefferson St(^a)</td>
<td>90.8</td>
<td>-25.2</td>
</tr>
<tr>
<td>S Greenbrier St</td>
<td>74.7</td>
<td>-41.3</td>
</tr>
<tr>
<td>S Dinwiddie St</td>
<td>61.2</td>
<td>-54.8</td>
</tr>
<tr>
<td>S Four Mile Run Dr(^b)</td>
<td>64.0</td>
<td>-52.0</td>
</tr>
<tr>
<td>S Buchanan St(^b)</td>
<td>65.7</td>
<td>-50.3</td>
</tr>
<tr>
<td>S Wakefield St</td>
<td>52.8</td>
<td>-63.2</td>
</tr>
<tr>
<td>S Thomas St</td>
<td>57.3</td>
<td>-58.7</td>
</tr>
<tr>
<td>S George Mason Dr</td>
<td>108.8</td>
<td>-7.2</td>
</tr>
<tr>
<td>S Quincy St</td>
<td>54.2</td>
<td>-61.8</td>
</tr>
<tr>
<td>S Monroe St</td>
<td>49.4</td>
<td>-66.6</td>
</tr>
<tr>
<td>S Glebe Rd</td>
<td>86.4</td>
<td>-29.6</td>
</tr>
<tr>
<td>S Highland St</td>
<td>55.8</td>
<td>-60.2</td>
</tr>
<tr>
<td>S Walter Reed Dr</td>
<td>74.0</td>
<td>-42.0</td>
</tr>
<tr>
<td>S Barton St</td>
<td>49.4</td>
<td>-66.6</td>
</tr>
<tr>
<td>S Wayne St(^c)</td>
<td>67.7</td>
<td>-48.3</td>
</tr>
<tr>
<td>S Courthouse Rd(^c)</td>
<td>75.2</td>
<td>-40.8</td>
</tr>
<tr>
<td>S Quinn St</td>
<td>49.8</td>
<td>-66.5</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>68.7</strong></td>
<td><strong>-47.3</strong></td>
</tr>
</tbody>
</table>

**Notes:**
1. Cycle length for coordinated-actuated control was 116.0 seconds at all intersections.
2. Each superscript (a,b,c) indicates a coupled zone.

**Figure 7-16** shows a record of signal state changes (red and green intervals) for the critical arterial direction (northbound, shown by diagonal lines) under self-organizing control without TSP scenario at three intersections on Columbia Pike – the first two intersections are coupled and the last intersection is uncoupled. At the first two coupled intersections, it is important to point out the strong coordination and how coupled intersections synchronize together in every cycle to introduce progression. Moreover, the second intersection is still allowed to double cycle, reducing delays for pedestrians and turning traffic, which cannot benefit from coordination. The third intersection, which is not coupled with the first two, offers partial
synchronization (i.e., synchronization in some cycles) that allows intersections with lower demand to cycle at a faster pace.

Figure 7-16: Signal State Changes and Green Bands for the Critical Arterial Direction for Carlin Springs Rd, S Jefferson St, and S Greenbrier St on Columbia Pike under Self-organizing Control Logic Without TSP
Chapter 8: Sensitivity Analyses of Secondary Extension Parameters for Arterial Coordination

The simulation results provided in the previous chapter indicated the success of the proposed self-organizing logic. However, some of the control parameters used in the arterial coordination logic were initially selected based on judgment rather than a scientific methodology, and, thus it is not clear whether the success of self-organizing signals heavily rely on the accuracy of those control parameters.

This chapter provides sensitivity analyses of parameters used in the secondary green extension logic. The analyses considered the parameters which affect the performance of secondary green extension logic. The tested secondary extension parameters include:

- Shape of affordable lost time per vehicle curve as a function of intersection v/c ratio
- Maximum affordable lost time per vehicle (defined as $L_{\text{max}}$, initially selected value was two seconds per vehicle)
- Intersection v/c ratio where affordable lost time per vehicle starts to decrease (defined as $X_{\text{min}}$, initially selected value was 0.50)
- The maximum length of secondary extension (defined as $S_{\text{Xmax}}$, initially selected value was 20 seconds)

In order to assess the effect of secondary extension parameters on self-organizing logic’s performance, a set of experiments were run in VISSIM and parameters were empirically tested. Over 70 simulation runs along Beacon Street and Rural Road were performed by changing one parameter at a time. The performed analyses and the evaluation of simulation results are provided in the following sections.
### 8.1 Sensitivity to the Shape of Affordable Lost Time per Vehicle Curve

When the v/c ratio increases at an intersection, affordable lost time per vehicle decreases to limit wasted green time, and the controller becomes less generous in providing secondary extension for an arriving platoon. As a result, the controller favors only platoons that contain less lost time per vehicle within a platoon. If an intersection is over capacity, affordable lost time per vehicle becomes zero and secondary extension is not allowed.

While calculating affordable lost time as a function of the intersection v/c ratio, it is assumed that affordable lost time is the allowed lost time per vehicle with respect to \( L_{\text{max}} \) that brings intersection v/c ratio to the critical v/c ratio, yielding equation (5.5) and Figure 5-4. To understand how the shape of affordable lost time per vehicle curve affects the performance of self-organizing signals, three additional curves were tested in the secondary green extension logic. Figure 8-1 shows the tested curves as well as the original curve that was used based on judgment.
Figure 8-1: Testing Different Affordable Lost Time per Vehicle Curves with respect to Intersection Volume to Capacity (v/c) Ratio

The solid line indicates the initially selected curve based on judgment, while the dashed lines given with the numbers show the three tested curves. Curve (1) assumes a concave down (i.e., function is decreasing at an increasing rate) affordable lost time per vehicle with respect to intersection v/c ratio, while curve (3) assumes a concave up function (i.e., function is decreasing at a decreasing rate), resulting in a more stringent affordable lost time per vehicle as demand approaches capacity. Curve (2) assumes a linear decline in affordable lost time.

Self-organizing logic with different affordable lost time per vehicle curves was tested along Rural Road for periods with volume below capacity (constant volume of 60-minute period following the warm-up, critical intersection operates at 90% of capacity). Results were obtained based on five simulation runs. Percentage change indicates the change in average delay within the network compared to the initially assumed affordable lost time per vehicle curve (Table 8-1).
Table 8-1: Simulation Results for Rural Road with Different Affordable Lost Time per Vehicle Curves

<table>
<thead>
<tr>
<th>Affordable Lost Time per Vehicle Curve</th>
<th>Average Delay (s/vehicle)</th>
<th>Percentage Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Curve (Based on Judgment)</td>
<td>46.0</td>
<td>0.0%</td>
</tr>
<tr>
<td>Curve 1, Concave Down</td>
<td>47.2</td>
<td>2.6%</td>
</tr>
<tr>
<td>Curve 2, Linear</td>
<td>46.5</td>
<td>1.0%</td>
</tr>
<tr>
<td>Curve 3, Concave Up</td>
<td>46.1</td>
<td>0.3%</td>
</tr>
</tbody>
</table>

Simulation results show that being more generous with allowed lost time (curve 1) as intersection v/c ratio increases resulted in the highest average delay (2.6% increase in average delay). This is intuitive, because as intersection v/c ratio increases, the slope of the affordable lost time curve should become less and less steep as v/c ratio approaches to zero, favoring concave up behavior. Therefore, the delay under the initial curve and curve (3) yielded lower average delay. Another important finding is that the shape of the affordable lost time per vehicle curve has a marginal impact on the performance of self-organizing logic (using linear curve only increased delay by 1.0%) As a result, it is concluded that the shape of the initially selected curve is reasonable to use in calculating allowed lost time per vehicle.

8.2 Sensitivity to $L_{max}$ and $X_{min}$

As explained earlier, $L_{max}$ is the maximum affordable lost time per vehicle (two seconds was used as the default value), and $X_{min}$ is the intersection v/c ratio where affordable lost time per vehicle starts to decrease (0.50 was set as default). In order to find if there is an optimal combination of the two parameters that minimizes average network delay or if a certain region exists that consistently results in lower average vehicle delay (say more generous or less
generous allowed lost time), different $L_{max}$ and $X_{min}$ values were tested along Rural Road and Beacon Street. 15 different combinations that were considered in the analysis (Figure 8-2).

![Figure 8-2: Combinations of $L_{max}$ and $X_{min}$ that were Tested along Rural Road and Beacon Street](image)

The triangular points indicate all tested combinations. $L_{max}$ values range between 1.0 and 2.0 in half-point increments (0.5), and $X_{min}$ values range from 0.3 to 0.7 in one-tenth (0.1) increments, leading to 15 different combinations. The lines show affordable lost time per vehicle values with respect to intersection v/c ratio for a specific $L_{max}$ and $X_{min}$ combination. For example, the dotted line shows affordable lost time per vehicle values when $L_{max} = 2.0$ and $X_{min} = 0.7$, while the dashed line represents the case of $L_{max} = 1.0$ and $X_{min} = 0.3$. Note that only the two extreme combinations were illustrated in Figure 8-2.
For each combination, five simulation runs for both arterials were performed, with results reporting averages for the 60-minute period following a 15-minute warm-up period. Simulation results were compared to initially selected values (default values).

Table 8-2 provides comparison of results from Rural Road. Note that the critical intersection for Rural Road operates at 90% of capacity.

Table 8-2: Comparison of Simulation Results for Rural Road under Different $L_{\text{max}}$ and $X_{\text{min}}$ Combinations

<table>
<thead>
<tr>
<th>$L_{\text{max}}$</th>
<th>$X_{\text{min}}$</th>
<th>0.30</th>
<th>0.40</th>
<th>0.50</th>
<th>0.60</th>
<th>0.70</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1.09%</td>
<td>0.00%</td>
<td>-0.43%</td>
<td>0.00%</td>
<td>0.00%</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>-0.87%</td>
<td>0.00%</td>
<td>0.87%</td>
<td>0.43%</td>
<td>2.17%</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>0.65%</td>
<td>1.09%</td>
<td>0.00%</td>
<td>-0.22%</td>
<td>1.96%</td>
<td></td>
</tr>
<tr>
<td>0.0²</td>
<td></td>
<td></td>
<td>6.09%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note:
1. Results were given in percentage changes compared to $L_{\text{max}} = 2.0$ and $X_{\text{min}} = 0.5$ combination.
2. $L_{\text{max}} = 0.0$ shows simulation result when no secondary extension is allowed.

Simulation results showed that except for the case where $X_{\text{min}} = 0.70$, the average network delay stayed within one percent of the initial result obtained using default values. Allowing higher $X_{\text{min}}$ values, however, caused approximately 2% increase in average delay. When no secondary extension logic was granted to arriving platoons, average delay increased by more than 6%, indicating how secondary extension logic can reduce delay by improving signal progression. The results also suggest that self-organizing logic is not very sensitive to the variations in the above-mentioned parameters.

However, note that the conducted analysis only considered one set of volumes and one test bed, thus it may not be accurate to draw valid conclusions. To get further insight, the same analysis was performed along Rural Road with reduced volumes, in which the volume profile in
the network was scaled down by 25% (the critical intersection operates at 67.5% of the capacity).

Simulation results are given in Table 8-3.

Table 8-3: Comparison of Simulation Results with Scaled Down Volumes for Rural Road under Different $L_{max}$ and $X_{min}$ Combinations

<table>
<thead>
<tr>
<th>$L_{max}$</th>
<th>$X_{min}$</th>
<th>0.30</th>
<th>0.40</th>
<th>0.50</th>
<th>0.60</th>
<th>0.70</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.53%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.27%</td>
<td>-0.27%</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0.27%</td>
<td>-0.27%</td>
<td>-0.27%</td>
<td>1.06%</td>
<td>0.80%</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>-0.80%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>1.86%</td>
<td>2.13%</td>
<td></td>
</tr>
<tr>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.79%</td>
<td></td>
</tr>
</tbody>
</table>

Note:
1. Results were given in percentage changes compared to $L_{max} = 2.0$ and $X_{min} = 0.5$ combination.
2. $L_{max} = 0.0$ shows simulation result when no secondary extension is allowed.

Similar results were obtained when the volume was scaled down. There was no appreciable change in average delay. Providing generous affordable lost time per vehicle generally caused higher network delays. Furthermore, inhibiting secondary extension increased delay by 4.8%. However, note that the improvement of secondary extension is less pronounced compared to higher volume case (4.8% vs. 6.1%), which can be attributed to the fact that the benefit of signal coordination starts to disappear as traffic demand gets lower.

Beacon Street was the second test bed that was considered in the analysis. The model was tested under two different traffic demands; the critical intersection operating at 79% of capacity, and the critical intersection operating at 60% of capacity. The comparison of average delay for different $L_{max}$ and $X_{min}$ values under both demand scenarios is reported in Table 8-4.
Table 8-4: Comparison of Simulation Results for Beacon Street under Different $L_{max}$ and $X_{min}$ Combinations

(a) The critical intersection’s v/c ratio is 79%

<table>
<thead>
<tr>
<th>$L_{max}$</th>
<th>$X_{min}$</th>
<th>0.30</th>
<th>0.40</th>
<th>0.50</th>
<th>0.60</th>
<th>0.70</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.30</td>
<td>2.15%</td>
<td>2.33%</td>
<td>2.15%</td>
<td>1.44%</td>
<td>0.72%</td>
</tr>
<tr>
<td>1.5</td>
<td>0.40</td>
<td>1.97%</td>
<td>2.33%</td>
<td>-0.36%</td>
<td>0.90%</td>
<td>0.54%</td>
</tr>
<tr>
<td>2.0</td>
<td>0.50</td>
<td>4.49%</td>
<td>1.08%</td>
<td>0.00%</td>
<td>1.44%</td>
<td>0.72%</td>
</tr>
<tr>
<td>0.0²</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.85%</td>
</tr>
</tbody>
</table>

(b) The critical intersection’s v/c ratio is 60%

<table>
<thead>
<tr>
<th>$L_{max}$</th>
<th>$X_{min}$</th>
<th>0.30</th>
<th>0.40</th>
<th>0.50</th>
<th>0.60</th>
<th>0.70</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.30</td>
<td>1.79%</td>
<td>1.35%</td>
<td>0.67%</td>
<td>0.00%</td>
<td>-0.67%</td>
</tr>
<tr>
<td>1.5</td>
<td>0.40</td>
<td>1.57%</td>
<td>1.35%</td>
<td>0.00%</td>
<td>-0.90%</td>
<td>-0.22%</td>
</tr>
<tr>
<td>2.0</td>
<td>0.50</td>
<td>-0.45%</td>
<td>0.45%</td>
<td>0.00%</td>
<td>0.67%</td>
<td>-0.22%</td>
</tr>
<tr>
<td>0.0²</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.48%</td>
</tr>
</tbody>
</table>

Note:
1. Results were given in percentage changes compared to $L_{max} = 2.0$ and $X_{min} = 0.5$ combination.
2. $L_{max} = 0.0$ shows simulation result when no secondary extension is allowed.

During periods of high flow (the critical intersection’s v/c ratio is 79%), initially selected values (default values) performed better than other combinations except for $L_{max} = 1.50$ and $X_{min} = 0.50$ case. When volumes were scaled down, default values performed the same or better than nine combinations, while five combinations were able to reduce average delay, though the reduction in delay was very marginal, only less than one percent.

Overall, the results from Rural Road and Beacon Street indicate that there is no optimal region or a particular combination that results in a substantial reduction in average delay compared to the default values selected based on judgment (i.e., $L_{max} = 2.0$ and $X_{min} = 0.50$).
Moreover, the results based on the default assumptions are not too far from the best ones in all scenarios. Finally, it is reasonable to conclude that the model performance is only slightly affected by the variations in $L_{\text{max}}$ and $X_{\text{min}}$.

### 8.3 Sensitivity to the Maximum Length of Secondary Extension

The results reported in the previous sections assumed 20 seconds (default) of maximum allowed secondary extension ($S_{\text{X}_{\text{max}}}$) for phases serving the critical arterial through movement. However, this assumption has not been empirically tested, and thus the selection of $S_{\text{X}_{\text{max}}}$ requires further experimentation.

To evaluate the sensitivity of $S_{\text{X}_{\text{max}}}$, self-organizing signals were also tested with 10 seconds and 30 seconds of maximum secondary green extension. The tested networks include Rural Road and Beacon Street. Only high demand periods were considered for both arterials (i.e., critical intersections operate at 90% and 79%, respectively). Table 8-5 provides average delay as well as average cycle length for modeled Rural Road and Beacon Street under different maximum secondary extension lengths.
Table 8-5: Simulation Results for Rural Road and Beacon Street under Different Maximum Secondary Extension Lengths

(a) Results for Rural Road

<table>
<thead>
<tr>
<th>Maximum Length of Secondary Extension (s)</th>
<th>Average Delay (s/vehicle)</th>
<th>Average Cycle Length (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>48.8 (6.09%&lt;sup&gt;1&lt;/sup&gt;)</td>
<td>54.3</td>
</tr>
<tr>
<td>10</td>
<td>46.9 (1.96%&lt;sup&gt;1&lt;/sup&gt;)</td>
<td>58.1</td>
</tr>
<tr>
<td>20</td>
<td>46.0 (0.00%&lt;sup&gt;1&lt;/sup&gt;)</td>
<td>60.6</td>
</tr>
<tr>
<td>30</td>
<td>45.8 (-0.43%&lt;sup&gt;1&lt;/sup&gt;)</td>
<td>62.9</td>
</tr>
</tbody>
</table>

(b) Results for Beacon Street

<table>
<thead>
<tr>
<th>Maximum Length of Secondary Extension (s)</th>
<th>Average Delay (s/vehicle)</th>
<th>Average Cycle Length (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>58.4 (4.85%&lt;sup&gt;1&lt;/sup&gt;)</td>
<td>63.5</td>
</tr>
<tr>
<td>10</td>
<td>57.4 (3.05%&lt;sup&gt;1&lt;/sup&gt;)</td>
<td>67.8</td>
</tr>
<tr>
<td>20</td>
<td>55.7 (0.00%&lt;sup&gt;1&lt;/sup&gt;)</td>
<td>67.7</td>
</tr>
<tr>
<td>30</td>
<td>Not Applicable&lt;sup&gt;2&lt;/sup&gt;</td>
<td>Not Applicable&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Note:
1. Values in parentheses indicate percentage change in average network delay compared to maximum secondary extension of 20 seconds.
2. 30 seconds of maximum secondary extension could not be tested along Beacon Street due to the close spacing of intersections, which limits the prediction horizon.

Both for Rural and Beacon arterials, applying 10 seconds of maximum secondary extension caused higher average delays, in which the delay increased approximately two percent and three percent, respectively. When 30 seconds of maximum extension was used, it slightly reduced average delay (0.43% percent reduction compared to 20 seconds extension). However, it is important to note that achieving a 30 second prediction horizon may not be practical, particularly along urban arterials with closely spaced signals, which limits far advanced detection.
Another important finding is the relation between delay and average cycle length. Results indicated that sacrificing a shorter cycle length improved progression, which in turn reduced average delay. Along Rural Road, for example, increasing the cycle length by 6.3 seconds (from 54.3 to 60.6 seconds) through granting secondary green extensions resulted in a delay reduction of 6.09%. For Beacon Street, 4.2 seconds of increase in cycle length reduced average delay by 4.85%, which clearly shows how excess capacity can be utilized to enhance progression.

Finally, fraction of cycles along Rural Road in which secondary green extension was provided for different maximum secondary extension length are given (Table 8-6).

Table 8-6: Fraction of Cycles with Secondary Green Extension when the Maximum Secondary Extension Length is 10, 20, and 30 Seconds

<table>
<thead>
<tr>
<th>Maximum Length of Secondary Extension (s)</th>
<th>Critical Intersection</th>
<th>Minor Intersections</th>
<th>All Intersections$^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.071</td>
<td>0.334</td>
<td>0.283</td>
</tr>
<tr>
<td>20</td>
<td>0.200</td>
<td>0.504</td>
<td>0.439</td>
</tr>
<tr>
<td>30</td>
<td>0.204</td>
<td>0.681</td>
<td>0.552</td>
</tr>
</tbody>
</table>

Note:
1. All intersections case does not include secondary green extensions at the first intersection where vehicles arrive randomly rather than in platoons, significantly reducing the chance of secondary extension.

Results indicated that the gain moving from 10 seconds to 20 seconds of maximum green extension is significant. By allowing only 10 seconds of maximum extension, the likelihood of secondary extension is reduced, because only platoons that are close to the intersection can be captured, de-emphasizing secondary extension logic, which in turn causes higher delay (Table 8-5). When 30 seconds of maximum green extension was used, the fraction of extended cycles was not very different compared to 20 seconds of maximum green extension case. This is
attributed to the fact that when the maximum allowed extension was 20 seconds, the controller was still able to provide secondary green extension in most cycles.
Chapter 9: Summary and Conclusions

This chapter presents the summary and conclusions on the main findings of this research. Section 9.1 draws general conclusions. The main contributions of this study are summarized in Section 9.2. Major findings are discussed in Section 9.3, and finally, in Section 9.4, suggestions are included for future research.

9.1 General Conclusions

The objective of this study was to develop signal control algorithms and rules that allow actuated signal control, which is free of any cycle length, to create a structure with green waves along an arterial. The control algorithms are built on the successful framework of actuated signal control, but with coordination rules added to make them self-organizing at the arterial level, automatically adjusting to fluctuations in demand, and organically forming green waves. The primary conclusions of this research are described as follows:

- This research has proven the potential for self-organizing arterial control based on fundamental actuation rules and communication among neighboring intersections.
- The proposed secondary green extension rules, which use excess capacity (slack time) at an intersection to improve arterial progression, create organic coordination, which leads to global coordination of traffic signals along an arterial.
- Dynamic coordination rules manage closely-spaced intersections (e.g., spacing is less than 600 ft), avoiding undesirable queue interaction (e.g., spillback and starvation).
- Control rules for managing oversaturation maximize throughput and result in less network delay with greater responsiveness.
• Self-organizing control logic achieves a flexible framework, which gives the controller the ability to recover from TSP interruptions. Self-organizing signals organically heal themselves after a signal is interrupted due to TSP, reducing the impact of TSP on private traffic.

9.2 Contributions

The main contributions of this study to the state-of-practice are as follows:

1. Multi-headway gap-out logic was developed for actuated control on multi-lane approaches. The development of multi-headway gap-out logic makes it easier to distinguish saturated flow from unsaturated flow, which limits wasted green time while ending a green phase. As a result the proposed multi-headway gap-out logic improves actuated controller’s efficiency on multi-lane approaches, reducing average delay and cycle length.

2. Self-organizing coordination rules were developed for arterial control, which allowed actuated control to be used along arterials. The proposed self-organizing arterial control logic results in less network delay with greater responsiveness to demand fluctuations.

3. Simple, rule-based control policies were developed to manage oversaturation. The developed oversaturation control policies yield comparable results compared to methods that require extensive computation or complicated mathematical models and outperform standard coordinated-actuated scheme.

4. A flexible framework that allows aggressive TSP was developed in this research. By offering a flexible framework, self-organizing signals are able to provide effective and aggressive TSP with negligible impacts to general traffic even when transit frequency is
high (e.g., priority request in every cycle), which can result in dramatic reductions in transit delay.

5. The efficiency of self-organizing signals was not sensitive to the choice of parameters used in the coordination algorithms. This eliminates the efforts required to fine-tune those control parameters in order to keep optimal performance for different applications.

9.3 Major Findings

This section describes the key findings of this dissertation.

1. The results of the analysis indicate that self-organizing control algorithms outperform optimized coordinated-actuated operation along arterials. For all test-beds considered, including arterials with irregularly and closely spaced intersections as well as arterials with uniformly spaced intersections that are well-suited to fixed cycle coordination, self-organizing algorithms yielded considerably less average delay per vehicle (reductions of 8% to 14% during under-saturated conditions). The comparison of results suggests that the reduction in average delay due to self-organizing logic is more pronounced when intersections are oversaturated (delay reductions of as high as 36% are possible).

2. Another major finding is that self-organizing signals can provide progression for through traffic while allowing intersections to operate at different cycle lengths. When intersections are not required to have the cycle length dictated by the critical intersection, cycle length is reduced significantly (more than 15% reduction in average cycle length compared to fixed cycle coordination), which in turn reduces pedestrian delay.

3. Self-organizing control logic exhibits inherent flexibility, which helps provide aggressive TSP without major disruptions to non-transit traffic. Evaluation of results indicate that a large reduction in transit delay (approximately 20% to 50% reduction in transit delay)
can be obtained using aggressive TSP tactics (e.g., generous green extension, phase skipping) under self organizing logic with no detectable impact to non-transit traffic even when transit frequency is high. In all test beds, self-organizing signals with TSP were able to keep average vehicle delay within 3% of average delay under optimized coordinated-actuated control without TSP. It was shown that, self-organizing logic with TSP still performed better (5% reduction in average delay) than the optimized coordinated-actuated control without TSP.

4. Overall, this research shows that the proposed self-organizing logic, which is built on standard actuated control, has a very promising performance along arterials, outperforming an optimized coordinated-actuated or a pre-timed signal operation.

### 9.4 Limitations and Future Directions

This section addresses the limitations of this study and provides recommendations for future research.

1. The proposed self-organizing algorithms are designed for arterial traffic signal control and cannot be applied to networks on its current state. The control logic should be expanded to include signal control logic of two-dimensional networks.

2. Factors (e.g., spillback threshold and biased factor) that are used in efficient lagging skipping and green start logic were arbitrarily selected. Further testing is required to determine if the model is sensitive to changes in these parameters.

3. The “early red” logic developed in this study to grant priority to buses when there is a near-side stop assumed that the mean and standard deviation of bus stop dwell time are given, based on historic data. A recommendation for future research is to develop a method that is able to calculate the mean and standard deviation of bus dwell time.
adaptively, allowing the controller to make more reliable dwell time prediction and eliminating the need for historic data.

4. In this study, the performance of self-organizing control algorithms was tested in a simulation environment. For future evaluation scenarios, self-organizing signals should be implemented and tested in real-world arterials/networks.
REFERENCES


