DEVELOPMENT AND EVALUATION OF A DISTRIBUTED SYSTEM FOR THE REAL-TIME CONTROL OF SIGNALIZED NETWORKS

by

François Dion

A thesis

presented to the University of Waterloo

in fulfillment for the

thesis requirement for the degree of

Doctor of Philosophy

in

Civil Engineering

Waterloo, Ontario, Canada, 1998

©François Dion 1998

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-38234-6

Canada

Borrower's Page

The University of Waterloo requires the signature of all persons photocopying this thesis. Please sign below, and give address and date.

Abstract

While past experiences show that real-time, traffic-responsive signal control has an ability to improve traffic operations in urban areas when compared to traditional fixed-time control, none of the existing real-time systems currently possess the ability to provide true optimal control in all types of networks. Two notable deficiencies exist when these systems are applied to networks in which passenger cars and transit vehicles share the right of way. First, none of these systems considers the effects on the general traffic of transit vehicles stopping in the right of way to board and discharge passengers. Second, priority of passage is often awarded to approaching transit vehicles without considering all the potential effects that such preferential treatment might have on other traffic.

The thesis describes a fully distributed signal control model that attempts to solve the above two problems. The model is a network extension of a real-time, traffic responsive signal control model for isolated intersections named SPPORT (*Signal Priority Procedure for Optimization in Real-Time*) that has been explicitly designed to consider the effects that transit vehicles might have on general traffic progression while stopped in the right of way and to provide priority to these vehicles on a conditional basis. While the network version of the model still attempts to optimize signalized intersections individually, it introduces procedures that allow the operation of adjacent intersections to be coordinated without excessively constraining the ability of the model to respond to changes in local traffic demands.

The SPPORT model features the use of a unique heuristic rule-based signal optimization procedure that allows the model to respond only to traffic events that are defined to have some importance for the signal operation. By ignoring all unimportant events, this optimization process allows a significant reduction in the number of signal-switching combinations that need to be considered to find an optimum signal control solution and make the SPPORT model more amenable to real-time control than exhaustive optimization methods.

In the model, coordination needs with downstream intersections are considered by adjusting, within the rule-based optimization process, the times at which green signal indications are required on each intersection approach following the identification of important traffic events. For each approach, adjustment are made on the basis of the projected signal timings at the main downstream intersection, the queue dissipation times along the links leading to that intersection, and the degree to which dwelling

transit vehicles stopped in the right of way interfere with the progression of general traffic on these links. Coordination with adjacent upstream intersections is achieved by examining the potential for queue spillback across these intersections, the recently implemented and projected local timings, and queuing conditions on the links joining the upstream intersections with the one being optimized. Coordination with upstream intersections is further enhanced by considering projected departures from them when predicting near-future stop line arrival patterns at each controlled intersection.

The simulation studies conducted in the thesis demonstrate the ability of the extended SPPORT model to provide efficient real-time, traffic-responsive signal control in coordinated urban networks with mixed-traffic conditions. When compared to an optimal fixed-time operation, the application of the proposed model reduced the delays incurred by all vehicle passengers at an isolated intersection by as much as 35 percent in scenarios considering transit interference on general traffic progression, transit priority treatments and peaking arrival patterns. When applied to a typical five-intersection urban arterial, reductions of up to 50 percent were observed in a performance function evaluating the stops and delays incurred by all vehicle passengers.

These results demonstrate the effectiveness of the priority rules defined within the model and of the model structure in which a series of candidate control strategies are evaluated before the best one is selected for implementation. The results also demonstrate the ability of the model to quickly respond to changing traffic conditions and automatically alter its signal control strategy when queues threaten to spill across controlled intersections. While the signal coordination procedures developed in the thesis are also found to be generally beneficial to the signal control performance, it was discovered that considering projected departures from upstream intersections can negatively affect the signal operation if the model is unable to assign different relative importance to projected traffic events in relation to current or imminent events.

Acknowledgments

I would like to express my thanks to Professors Sam Yagar and Bruce Hellinga, who both supervised this research and helped me in bringing this project to a successful end.

I gratefully acknowledge the financial support provided to me by the Natural Sciences and Engineering Research Council of Canada through its postgraduate scholarship program and the Ministry of Transportation of Quebec through the FCAR scholarship program. Generous additional funds were also provided by the Lea Associates Group through their contribution to the scholarship program of the Transportation Association of Canada. I also whish to thank the Engineering Faculty of the University of Waterloo for its added financial support at three separate occasions during the course of my program.

I send my deepest regards to my parents, who gave me the love, self-confidence and moral support to pursue my dreams.

I also give my deepest thanks to Juanita, who gave me her moral support and sacrificed so many things to allow me to complete this thesis.

To my parents

Table of Contents

AUTHOR'S DECLARATION ii

BORROWER'S PAGE iii

ABSTRACT iv

ACKNOWLEDGMENTS vi

LIST OF TABLES xii

LIST OF FIGURES xiii

LIST OF SYMBOLS xvi

1. INTRODUCTION 1

2. BACKGROUND 4

2.1. Inherent Variability of Traffic Demand 4

2.2. Traffic Signal Control Philosophies for Urban Networks 7

2.2.1. First Generation Control 9

2.2.2. Second Generation Control 13

2.2.3. Third Generation Control 13

2.3. Transit Operations in Urban Networks 14

2.3.1. Characteristics of Transit Movements 14

2.3.2. Transit Interference in Shared Right of Way 16

2.3.3. Transit Priority at Signalized Intersections 19

2.3.4. Selective Detection Technologies 23

3. EXISTING REAL-TIME SIGNAL CONTROL SYSTEMS 27

3.1. Control Principles 28

3.2. Transit Priority Features 33

3.3. Transit Interference Modeling 37

3.3. Summary of Existing Approaches to Coordinated Real-Time Control 39

4. RESEARCH OBJECTIVES **40**
4.1. Limitation of Existing Systems for Urban Networks 40
4.2. The SPPORT Model for Isolated Intersections 42
4.3. Research Goals 43

5. SIGNAL CONTROL FRAMEWORK **47**
5.1. Traffic Monitoring Scheme 48
5.1.1. Selective Detection Requirements 48
5.1.2. Estimation of Vehicle Occupancy 49
5.1.3. Traffic Detection Scheme 50
5.2. Control Architecture 52
5.3. Network Control Principles 57
5.4. Real-Time Signal Operation 61
5.5. Signal Optimization Principles 64
5.5.1. Acyclic Network Control 64
5.5.2. Objective Function 67
5.5.3. Rule-Based Switching Decision Process 71
5.5.4. Multi-Objective Optimization Process 73

6. DISCRETE-EVENT TRAFFIC SIMULATION MODEL **75**
6.1. Discrete-Event Simulation Process 76
6.2. Components of the Simulation Model 76
6.2.1. Vehicles 77
6.2.2. Segments 78
6.2.3. Links 86
6.2.4. Vehicle Activities 88
6.2.5. Signal Controllers 92

7. TRAFFIC DEMAND ESTIMATION **95**
7.1. Tracking of Current Traffic Conditions 95
7.2. Estimating of Departures from Upstream Intersections 98
7.3. Estimating Stop line Traffic Demand 102
7.4. Platoon Identification Algorithm 108

8. SIGNAL OPTIMIZATION 112
8.1. Signal Switching Rules 112
8.1.1. Rules for Queue Management 114
8.1.2. Rules for Congestion Management 116
8.1.3. Rules for Uninterrupted Traffic Progression 120
8.1.4. Rules for Transit Priority 122
8.2. Rule-Based Decision-Making Process 123
8.2.1. Rule Prioritization 125
8.2.2. Generation of Request Lists 126
8.2.3. Switching Decision Logic 129
8.2.4. Handling of Phase Sequence Restrictions 129
8.2.5. Handling of Signal Timing Constraints 130
8.3. Signal Coordination Logic 131
8.3.1. Handling of Incoming Platoons 131
8.3.2. Handling of Queue Spillbacks 132
8.3.3. Request Time Delaying Logic 132
8.4. Calculation of Ideal Request Times 133
8.4.1. Coordination with Downstream Timings 134
8.4.2. Coordination when Downstream Queues Exist 136
8.4.3. Coordination with Downstream Transit Activities 141
8.4.4. Coordination with Upstream Timings 147
8.4.5. Coordination with Upstream Queue Spillbacks 151
8.5. Signal Optimization Process 155

9. EVALUATION OF THE SPPORT MODEL 158
9.1. Special-Purpose Test Program 158
9.2. Test Scenarios 161
9.2.1. Test Networks 161
9.2.2. Traffic Demand Scenarios 163
9.2.3. Signal Control Parameters 166
9.3. Effectiveness of SPPORT Signal Control at Isolated Intersections 171
9.4. Effectiveness of SPPORT Signal Control on Urban Arterials 183
9.5. Effectiveness of Signal Coordination Logic 190
9.6. Effectiveness of Transit Priority Rules 196

9.7. Effectiveness of Multi-Objective Signal Control 201
9.8. Sensitivity to Selected Signal Control Parameters 204
9.9. Real-time Control Applicability 208

10. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS 211
10.1. Summary of Modifications to the SPPORT Model 211
10.2. Conclusions 214
10.3. Recommendations for Further Research 219

APPENDIX A. EXISTING REAL-TIME, TRAFFIC-RESPONSIVE SIGNAL CONTROL SYSTEMS 223
A.1. SCATS - Sydney Coordinated Adaptive Traffic System 224
A.2. SCOOT - Cycle, Split and Offset Optimization Technique 230
A.3. UTOPIA - Urban Traffic OPtimization by Integrated Automation 237
A.4. PRODYN - "PROgrammation DYNamique" (DYNamic PROgramming) 244

APPENDIX B. ISOLATED INTERSECTION TEST SCENARIOS 250

APPENDIX C. ARTERIAL TEST SCENARIOS 253

APPENDIX D. WEBSTER-COBBE SIGNAL OPTIMIZATION FOR ISOLATED INTERSECTIONS OPERATED IN REAL-TIME 259

APPENDIX E. SIMULATION RESULTS BETWEEN SPPORT AND FIXED-TIME CONTROL FOR ISOLATED INTERSECTION SCENARIOS 264

APPENDIX F. SIMULATION RESULTS BETWEEN SPPORT AND FIXED-TIME CONTROL FOR ARTERIAL SCENARIOS 281

APPENDIX G. COORDINATION LOGIC EVALUATION RESULTS 298

APPENDIX H. MULTI-OBJECTIVE SIGNAL OPTIMIZATION RESULTS 311

REFERENCES. 318

GLOSSARY. 328

List of Tables

Table 2.1 - Classification of Traffic Signal Control Strategies for Urban Coordinated Networks 9
Table 2.2 – Passive Transit Priority Treatments at Signalized Intersections 21
Table 2.3 – Active Transit Priority Treatments at Signalized Intersections 22
Table 2.4 - Comparison of Selective Detection Technologies 26
Table 8.1 - Request Priority Evaluation Functions 125
Table 8.2 - Evaluation of Traffic Demand Associated with each Request 153
Table 9.1 - Test Scenarios 163
Table 9.2 - Rule Control Parameters for Test Scenarios 168
Table 9.3 - Signal Timing Parameters 170
Table 9.4 - Link Control Parameters 170
Table 9.5 - Coordinated Links for Arterial Test Configuration 170
Table 9.6 - Signal Optimization Parameters 170
Table 9.7 - Performance Function Parameters for the Isolated Intersection Scenarios 172
Table 9.8 - Performance Function Parameters for the Arterial Scenarios 184

List of Figures

Figure 2.1 - Daily Variability of Five-Minute Traffic Counts in one Direction 5
Figure 2.2 - Variability of Cyclic Arrivals at a Typical Urban Intersection 6
Figure 2.3 - Notions of Cycle Time, Green Allocation, Phase Sequence and Signal Offsets 10
Figure 2.4 - Typical First Generation Traffic Signal Control with Time-of-Day Plan Selection 11
Figure 2.5 - Typical Passenger Car and Transit Vehicle Movements along Urban Streets 15
Figure 2.6 - Observed Dwell Times at a Typical Urban Transit Stop 16
Figure 2.7 - Typical Transit Loading Activities near Signalized Intersections 17
Figure 2.8 - Example of Transit Interference on General Traffic Progression along Urban Streets 18
Figure 5.1 - Sample Arrangement of Vehicle Detectors on Intersection Approaches 51
Figure 5.2 - Hierarchical Traffic Signal Control Approaches 54
Figure 5.3 - Fully Distributed Traffic Signal Control Approach 55
Figure 5.4 - Modeling of Dependent Links and Coordinated Intersections around Individual Intersections 59
Figure 5.5 - Information Exchange Between Adjacent Traffic Signal Controllers 60
Figure 5.6 - Extended Traffic Projection Horizon in Signal Optimization Process as a Result of Data Exchange between Adjacent Traffic Signal Controllers 61
Figure 5.7 - Discrete Time, Rolling Horizon Decision-Renewal Process 62
Figure 5.8 - Induced Repetitiveness of Traffic Patterns under Cyclic Traffic Signal Operation 64
Figure 5.9 - Theoretical Relationship Between Cycle Time and Delay (from McShane, 1990) 66
Figure 5.10 - Cumulative Cost along a Decision Tree 68
Figure 5.11 - Calculation of Terminal Cost at end of Decision Horizon 71
Figure 5.12 - Heuristic Rule-Based Signal-Switching Decision Process 72
Figure 5.13 - Multi-Objective Decision-Making Process 74
Figure 6.1 - Distinction Between Downstream and Opposing Segments 78
Figure 6.2 - Horizontal Queueing Behavior within Simulation Model 83
Figure 6.3 - Example of Simulated Horizontal Queue Behavior 84
Figure 6.4 - Relationship between Traffic Signal Indications and Effective Timings 85
Figure 6.5 - Modeling of Left-Turning Movements within Simulation Model 86
Figure 6.6 - Modeling of a Multiple Lane Approach in Terms of Segments and Links 87
Figure 6.7 - Modeling of Vehicle Paths along an Intersection Approach 89
Figure 6.8 - Modeling of a Vehicle Path with Two Detector Inputs 91

Figure 7.1 - Current System State Evaluation Process at Individual Intersections 96
Figure 7.2 - Generation of Projected Stop Line Arrivals using Traffic Detector Information and Projected Departures from Upstream Signalized Intersection 99
Figure 7.3 - Modeling of a Vehicle Path with a Traffic Arrival Record Point 101
Figure 7.4 - Relation between Decision Horizon and Look-Ahead Periods 103
Figure 7.5 - Use of Artificially Generated Arrivals within the Demand Estimation Process 106
Figure 7.6 - Source of Traffic Information within Decision Horizon and Look-Ahead Periods 107
Figure 7.7 - Notions of Headway and Critical Headway in Platoon Identification Algorithm 109
Figure 7.8 - Platoon Identification Example 111
Figure 8.1 - Response to Queue Spillbacks across Signalized Intersections 117
Figure 8.2 - Example of Platoon Request on Intersection Approach 121
Figure 8.3 - Response to a Transit Vehicle Approaching a Transit Stop 124
Figure 8.4 - Traffic Conditions Considered when Generating Signal Requests 126
Figure 8.5 - Active and Inactive Requests 127
Figure 8.6 - Coordination of Signal Operation with Downstream Signal Timings within Request Delaying logic 135
Figure 8.7 - Stop Line Queue Dissipation Time at the Beginning of a Green Interval 137
Figure 8.8 - Request Delaying in Response to Downstream Queuing 138
Figure 8.9 - Signal Offset Calculation Example 1 140
Figure 8.10 - Signal Offset Calculation Example 2 141
Figure 8.11 - Interfering Dwell Time on Passenger Car Traffic for Dwelling Transit Vehicles 143
Figure 8.12 - Interfering Dwell Time on Passenger Car Traffic for Approaching Transit Vehicles 144
Figure 8.13 - Examples of Request Delaying in Response to Transit Activities 146
Figure 8.14 - Effect of Signal Coordination with Upstream Signalized Intersections 148
Figure 8.15 - Calculation of Maximum Permitted Request Delay 150
Figure 8.16 - Coordination Area on Intersection Exit Links within Request Delaying Logic 152
Figure 8.17 - Signal Optimization Process 156
Figure 9.1 - Structure of Special-Purpose Test Program 159
Figure 9.2 - Isolated Intersection Test Network Configuration 161
Figure 9.3 - Arterial Test Network Configuration 162
Figure 9.4 - Temporal Variations in Traffic Demand for Peaking Demand Scenarios 164
Figure 9.5 - Variance of Simulated Fixed-Time and SPPORT Operations over 20 Replications 165
Figure 9.6 - Examination of Delay versus Cycle Time Relationship for a Simulated One-Hour Fixed-Time Operation 174

Figure 9.7 - SPPORT Effectiveness over a One-Hour Control Period for Isolated Intersection Scenarios without Transit Vehicles 175
Figure 9.8 - SPPORT Effectiveness over a One-Hour Control Period for Isolated Intersection Scenarios with Transit Vehicles 176
Figure 9.9 - SPPORT Signal Timings at an Isolated Intersection under Constant Demand while Using Five-Second Decision Intervals 178
Figure 9.10 - SPPORT Signal Timings at an Isolated Intersection under Constant Demand while Using One-Second Decision Intervals 179
Figure 9.11 - Variation of SPPORT Timings at an Isolated Intersection under Peaking Demand 180
Figure 9.12 - Variation of SPPORT Signal Timings at an Isolated Intersection in Response to Transit Priority Requests 182
Figure 9.13 - Effect of Stop Weighting Coefficient Value on Traffic Signal Performance 186
Figure 9.14 - Effectiveness of SPPORT Traffic Signal Operation over a One-Hour Control Period in Urban Arterial Scenarios with no Transit Vehicles 188
Figure 9.15 - Effectiveness of SPPORT Traffic Signal Operation over a One-Hour Control Period in Urban Arterial Scenarios with Transit Vehicles 188
Figure 9.16 - Effect of Signal Coordination Elements on Arterial Traffic Control Performance 191
Figure 9.17 - Effect of Request Delaying Logic on Arterial Traffic Control Performance 192
Figure 9.18 - Effect of Individual Elements of the Request Delaying Logic on Arterial Traffic Control Performance in Scenario without Transit Vehicles 193
Figure 9.19 - Effect of Individual Elements of the Request Delaying Logic on Traffic Control Performance in all Arterial Scenarios 194
Figure 9.20 - Effect of Transit Priority Rules for Transit Riders in Isolated Intersection Scenarios 198
Figure 9.21 - Effect of Transit Priority Rules for Non-Transit Riders in Isolated Intersection Scenarios 198
Figure 9.22 - Effect of Transit Priority Rules for Transit Riders in Arterial Scenarios 199
Figure 9.23 - Effect of Transit Priority Rules for Non-Transit Riders in Arterial Scenarios 199
Figure 9.24 - Comparative Performance of Multi-Objective Traffic Signal Operations 202
Figure 9.25 - Effect of Duration of Decision Horizon on Traffic Control Performance 205
Figure 9.26 - Effect of Duration of Look-Ahead Period on Traffic Control Performance 205
Figure 9.27 - Effect of Stop Coefficient in Performance Function on Traffic Control Performance 206
Figure 9.28 - Effect of Critical Flow Parameter on Traffic Control Performance 206
Figure 9.29 - SPPORT Traffic Signal Optimization Times on a 233 MHz Pentium II Computer 209

List of Symbols

α_R = User-specified coefficient defining the relative importance of vehicles stopped on an approach with a green signal indication at the end of the decision horizon with respect to vehicles stopped on an approach with a red signal indication ($\alpha_R \geq 0$).

$\beta_{veg\,m}$ = Gating factor on segment m ($0 \geq \beta \geq 1$).

B_k = Benefit at intersection k of extending the green signal indication on link j for one plan interval Δt.

d_i = Total delay incurred by vehicle i in the controlled system (seconds in SPPORT, vehicle-hour in TRANSYT-7F).

$d_{induced\,j}$ = Delay induced on link j by extending the red signal by Δt seconds (seconds).

$d_{saved\,j}$ = Delay saved on link j by extending the green signal indication by Δt seconds (seconds).

$f_{transit\,m}$ = Proportion of saturation flow available on segment m during transit dwell time.

h_{avg} = Average headway of vehicles in platoon (seconds).

h_{crit} = Critical headway (seconds).

h_i = Time headway between the i^{th} and i^{th-1} vehicles (seconds).

$h_{res\,m}$ = Current reserve headway on segment m (seconds).

$(h_{res\,m})_i$ = Reserve headway on segment m after arrival of i^{th} vehicle (seconds).

i = Subscript identifying individual vehicles.

j = Subscript identifying individual link.

k = Subscript identifying individual intersections

k_d = User-specified coefficient defining the relative importance of delay ($k_d \geq 0$).

k_S = User-specified coefficient defining the relative importance of stops ($k_S \geq 0$).

k_{TC} = User-specified coefficient defining the relative importance of the terminal cost ($k_{TC} \geq 0$).

k_{TT} = User-specified coefficient defining the relative importance of travel time ($k_{TT} \geq 0$).

$l_{prev\,veh\,m}$ = Size of previous vehicle having entered segment m (passenger car units).

L_m = Length of segment m (meters).

m = Subscript identifying the segments constituting a link in the simulation model.

$N_{exit\,j}$ = Number of exit links from approach link j.

N_{int} = Total number of intersections in controlled network.

N_{link} = Number of approach links to intersection k.

$N_{links\,T7F}$ = Total number of links in TRANSYT-7F networks.

$N_{up\,seg\,jm}$ = Number of segments between the upstream end of link j and segment m.

N_{sim} = Total number of vehicles entering the controlled network during a given period of time.

o_v = Coefficient representing the average occupancy or relative importance of a vehicle of type v ($o_v \geq 0$).

PI = SPPORT Performance index.

PI_{T-F} = TRANSYT-7F performance.

$q_{max\,m}$ = Maximum exit flow rate from segment m (passenger car units/second).

$q_{next\,\Delta t\,j}$ = Projected arrival on link j in the next plan interval (passenger car units).

$q_{sat\,j}$ = Saturation flow on approach link j (passenger car units/second).

$q_{sat\,m}$ = Saturation flow on segment m (passenger car units/second).

$q_{veh\,m}$ = Number of vehicles on segment m (passenger car units).

$Q_{end\,j}$ = Queue size on link j at end of decision horizon (passenger car units).

$Q_{stop\,j}$ = Number of vehicles currently queued at the stop line of link j (passenger car units).

$Q_{tot\,m}$ = Queuing capacity of segment m, taken to be zero if the segment is located upstream of the critical queue reach (passenger car units).

R_d = Request demand (passenger car units).

$R_{min\,j}$ = Minimum red interval on link j based on phasing constraints if a switch from a red to green signal display is initiated now (seconds).

RT_{del} = Delayed request time (seconds).

$RT_{del\,j}$ = Delayed request time based on traffic conditions on exit link j (seconds).

RT_0 = Initial request time (seconds).

$RT_{queue\,m}$ = Delayed request time after consideration of queuing conditions on segment m (seconds).

$\Delta RT_{queue\,m}$ = Change in ideal request time attributed to queuing conditions on segment m (seconds).

$\Delta RT_{transit}$ = Delay in request time accounting for transit interference (seconds).

$s_{avg\,j}$ = Average number of stops on link j per unit of time in TRANSYT-7F (stops/second).

s_i = Total number of stops incurred by vehicle i in the controlled system.

$s_{saved\,j}$ = Stops saved on link j by extending the green signal indication by Δt seconds (passenger car units).

t = Current time (seconds).

$t_{\Delta t}$ = Time of current decision point (seconds).

Δt = User-specified interval between decision points (typically 5 seconds).

$t_{full\,m}$ = Time that a first vehicle exited segment m after it became full (seconds).

$t_{interf\,dwell}$ = Interfering dwell time, as defined in Figures 8.11 and 8.12 (seconds).

$t_{m\,i}$ = Time vehicle i entered segment m at its upstream (seconds).

$t_{next\,m}$ = Next possible entry time at upstream end of segment m (seconds).

$t_{prev\,m}$ = Previous entry time at upstream end of segment m (seconds).

TC_k = Terminal cost for intersection k at end of traffic performance evaluation period.

TT_i = Total travel time of vehicle i in the controlled system (seconds).

$Turn_{j,j'}$ = Turning proportion from approach link j to exit link j'.

U_j = Signal display on approach link j (0 if signal is green, 1 if signal is red).

v = Subscript identifying the type of vehicle (passenger cars, transit vehicle).

V_m = Average free flow travel speed on the segment (meters/second)

w_j = Link-specific stop and delay coefficient in TRANSYT-7F (w_j = 1 for links carrying passenger cars and 40 for links carrying transit vehicles).

1. Introduction

The advent of computerized traffic signal control in the 1960's improved for many cities the possibilities to begin to deploy centrally controlled and monitored traffic signal control systems. While these early systems offered significant advantages over the previously available electromechanical devices, they still imposed a certain rigidity that restricted the opportunity for traffic responsiveness at the network level. Despite the use of computers to generate signal timings, most of these systems were still not powerful enough to allow on-line traffic signal optimization. As a result, signal timings could not be automatically adjusted to changes in traffic demand. Pre-defined timing plans still had to be used and be put on-line at predetermined times.

The emergence and rapid development of microprocessor technologies in the 1970's and 1980's opened the practical possibilities for real-time, traffic responsive signal control in urban networks. These new technologies provided sufficient computing power at a reasonable cost to allow the optimization of traffic signal timings to be carried out on-line. Since then, the interest for such type of traffic signal control has never ceased to increase. This interest is driven by the need to find more efficient ways to utilize existing network capacities in environments in which it is increasingly difficult to build new infrastructure to meet the needs of growing traffic demands. In many cities, physical and financial constraints as well as environmental concerns prevent the expansion of existing roads or the construction of new roads.

In response to the above problems, the object of traffic management policies has gradually been shifting away from the freedom of movements of individual vehicles to the fair allocation of limited network capacities to all road users. However, while past experiences show that real-time traffic signal control systems have an ability to improve traffic operations in urban areas, none of the existing systems currently possess the ability to provide true optimal traffic signal control in all types of networks. Deficiencies notably exist when these systems are applied to networks in which passenger cars and transit vehicles share the right of way. First, none of the existing systems currently considers the effect on the general traffic of transit vehicles stopping in the right of way to board and discharge passengers. While they are stopped, these vehicles can partially or completely block a traffic lane and

create a temporary bottleneck. This bottleneck may then result in insufficient use of the time allocated to the various green intervals within the signal cycle and in subsequent increased congestion levels. Second, priority of passage is often offered to transit vehicles without considering all the potential effects on other traffic. In many cases, for instance, signal timings are altered on the sole basis that an approaching transit vehicle has been detected.

In order to allocate network capacities equitably, real-time traffic signal control systems must provide proper consideration to all types of vehicles. The SPPORT model (*Signal Priority Procedure for Optimization In Real-Time*) was developed specifically to address this need. The development of this model was initiated in response to a need expressed by the Municipality of Metropolitan Toronto for providing priority of passage to streetcars along a corridor in which transit operations create significant interference to the progression of other traffic (Yagar, Han and Greenough, 1991, 1992). A first research effort resulted in the development of a Pascal computer program that could be used to control signal timings at isolated intersections controlled by a two-phase signal operation (i.e., a simple red/green signal operation) and having up to four approaches (Han and Yagar, 1991, 1992; Yagar, Han and Wang, 1992; Yagar et Han, 1994). The main novelty of this model was the use of a heuristic rule-based signal optimization procedure to generate candidate signal control strategies and select the best one for implementation. Later, Conrad (1997) and Conrad et al. (1998) expanded the model's applicability to any type of intersection configurations and to intersections where traffic is controlled by more than two phases. During this second development step, Conrad translated the model into C++ using object-oriented programming techniques, enhanced the rule-based signal optimization procedure by designing a logic allowing the model to choose which phase to go to next when a set of choices is possible, and redesigned the model's traffic simulation module. He replaced the existing macroscopic simulation model with a discrete-event microscopic model considering the progression of vehicles along a series of short, unidirectional segments, which still explicitly modeled the effects transit vehicles have on the general traffic while they are stopped at a transit stop. However, despite these enhancements, the SPPORT model remained capable of controlling only isolated intersections.

The research project described in this thesis is mainly concerned with expanding the applicability of the SPPORT model to networks of coordinated signalized intersections. The objective of the research is to develop signal optimization procedures and data communication processes that would allow the model to coordinate the operation of adjacent intersections while still maintaining sufficient flexibility of operation at individual intersections to allow quick responses to traffic demand changes and transit

priority requests. The scope of this approach is to optimize the traffic signal timings for vehicular traffic flow. Pedestrians and safety issues are not considered for simplicity reasons only.

The various aspects of the research are presented in the following nine chapters as follows:

- Chapter 2 presents background material on traffic demand variability, real-time traffic signal control, and transit priority treatments that are relevant to the current research project.
- Chapter 3 describes the existing real-time traffic signal control systems and evaluates their applicability to networks in which passenger cars and transit vehicles share the right of way.
- Considering the problems identified in Chapter 3, Chapter 4 presents the main control principles behind the development of the SPPORT model for urban networks.
- Chapter 5 presents a more technical overview of the SPPORT model for coordinated networks. It describes the traffic detection scheme, the control architecture of the model, the way the signal coordination is achieved within the chosen architecture, the real-time signal control process, and the main principles behind the signal optimization process.
- Chapter 6 describes the traffic simulation model used by SPPORT to estimate the demand placed on each intersection and to evaluate the candidate control strategies produced by its signal optimizers.
- Chapter 7 presents how the traffic demand from each intersection approach is estimated.
- Chapter 8 describes the signal optimization process used by SPPORT to generate near-optimal signal timings in a coordinated network environment.
- Chapter 9 evaluates the ability of the proposed model to control traffic at single isolated intersections and along urban arterials.
- Chapter 10, finally, presents some general conclusions and provides recommendations for future research regarding the SPPORT model.

2. Background

Real-time traffic signal control systems have the ability to automatically adjust the timing parameters of traffic signals in response to the latest observed traffic conditions. These advanced control systems are required to dynamically respond to traffic demand that vary over time within a given day, and from one day to the next. Further more, these systems provide an opportunity to give priority of passage to transit vehicles at signalized intersection without unduly affecting other traffic.

This chapter describes the challenges posed by real-time traffic signal control and generally describes the past efforts that have been made to develop efficient real-time traffic signal control systems. The first section describes and illustrates the variability in traffic demands. The following section presents the evolution of traffic signal control philosophies that has taken place over the years as a result of the search for signal control systems that would be more responsive to changes in current traffic conditions. The last section identifies and describes the additional complexities associated with providing priority to transit vehicles in networks where transit vehicles and passenger cars share the right of way.

2.1. Inherent Variability of Traffic Demand

Urban traffic demand is a temporally varying process. Figure 2.1 illustrates the daily variation in traffic as observed in an extensive traffic survey conducted in 1973 by the Metropolitan Toronto Department of Road and Traffic (McShane and Crowley, 1976). The figure shows a compilation of five-minute traffic counts on an urban arterial leading traffic towards downtown Toronto. The solid line in the figure indicates the average level of traffic in each count interval over the 77-weekday collection period while the shaded area represents the region into which 95 percent of all the counts fell.

The first observation that can be made on the basis of Figure 2.1 is the degree to which the average traffic demand fluctuates within a given weekday. In this example, traffic demand peaks sharply in the morning, is reduced after the morning rush, and increases again during the late afternoon and early evening before declining to its daily minimum. Within a period of two hours in the morning, traffic demand rapidly increases from less than 100 vehicles per hour (vph) to about 950 vph. During the mid-

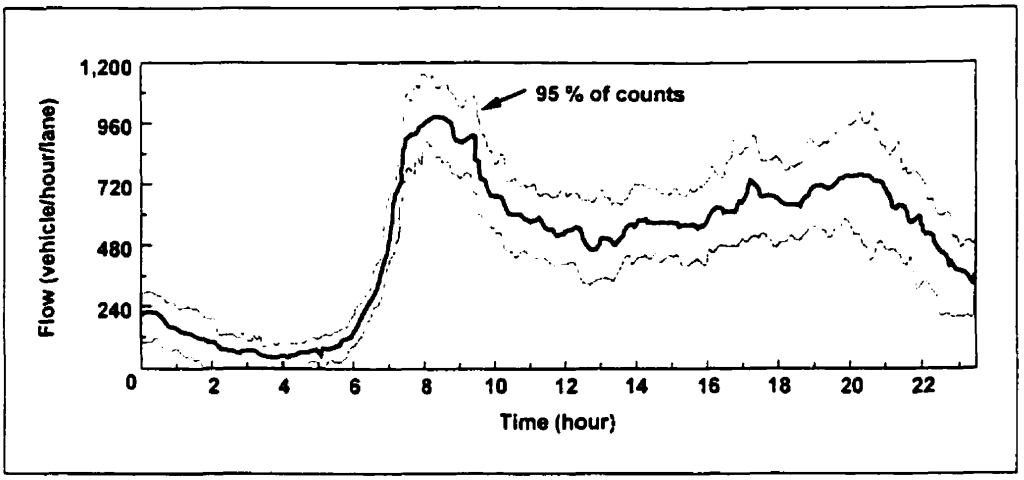


Figure 2.1 - Daily Variability of Five-Minute Traffic Counts in one Direction
(Redrawn from: McShane and Crowley, 1976)

day period, demand drops by about one third to 600 vph before increasing again to 800 vph during the evening peak period. The magnitude of these variations makes it impossible to design a single traffic signal control policy that would be optimal throughout the entire day. During high-demand periods, there is typically a need to provide longer green intervals on intersection approaches to serve all incoming traffic vehicles. However, during periods of low-demand, these longer intervals usually result in longer delays and additional stops. In order to maintain a certain degree of optimality in traffic signal control, different control policies must therefore be prepared for each major period of the day. In this case, distinct control policies would be required for the morning peak, mid-day, afternoon peak, and night control periods.

A second observation that can be made is the variability of traffic demand from one day to the next. An analysis of the diagram of Figure 2.1 reveals that the boundaries of the shaded area are located at about plus and minus 25 percent of the corresponding average five-minute traffic count. This indicates a high degree of variation from one day to the next, and consequently a high degree of uncertainty in predicting future five-minute traffic counts.

The full extend of the short-term traffic variability is illustrated in Figure 2.2. The three diagrams shown illustrate the variability of vehicle arrivals at a typical urban intersection where the cycle time is 80 seconds. The two diagrams show the observed 80-second counts for two different weekdays one week apart. In this case, it can be seen that the number of arrivals over successive 80-second periods is far from being constant. While there are on average 6.7 vehicles crossing the intersection stop line

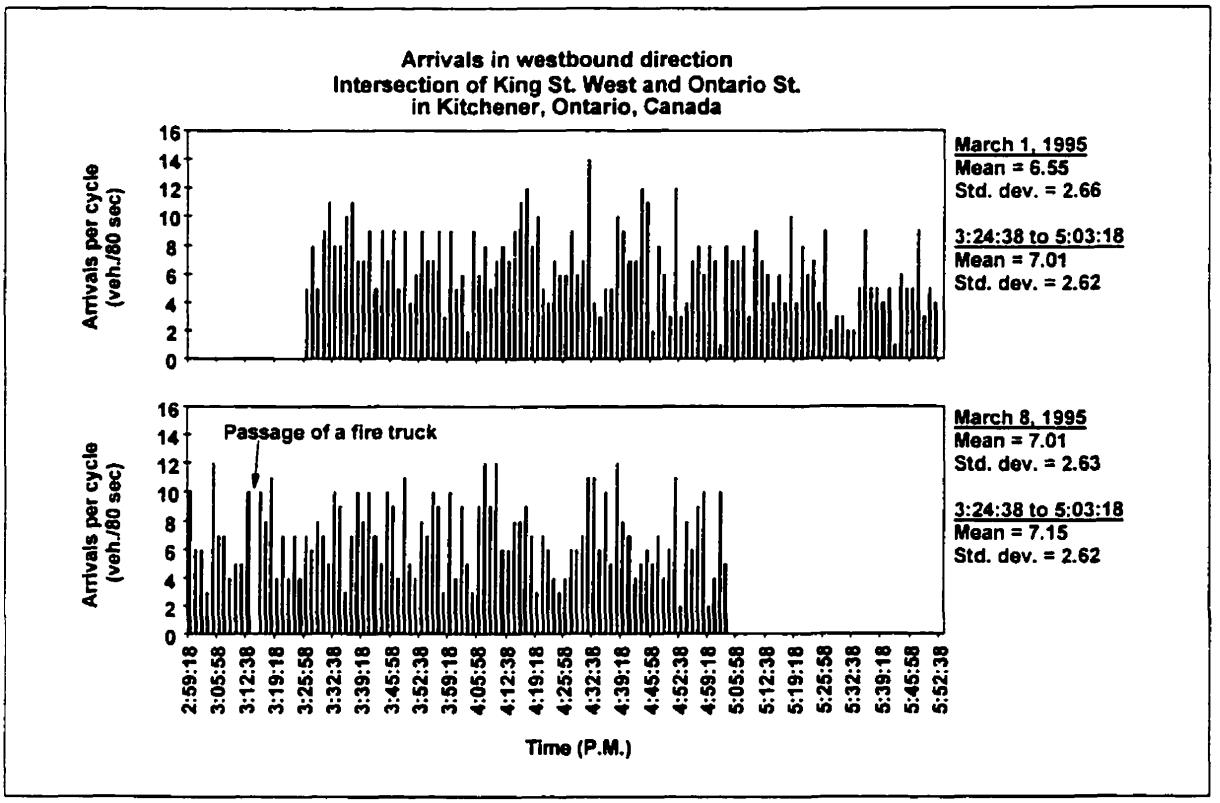


Figure 2.2 - Variability of Cyclic Arrivals at a Typical Urban Intersection

every signal cycle, the actual number of vehicles entering the intersection during a given cycle of signal operation varies between two and 14 vehicles.

A runs test and a correlation plot analysis (Law and Kelton, 1991) of the observed counts indicate that the number of arrivals per signal cycle can be considered as a independent random variable. This leads to the conclusion that short-term variability is essentially stochastic. On the other hand, the average time-of-day variations shown in the example Figure 2.1 are essentially deterministic in nature, as they are the products of scheduled urban activities. For instance, the peak in demand in the early part of the morning is mainly produced by the fact that a majority of persons choose to leave their home to go to their workplace at about the same time every day. Similarly, the simultaneous movements of people returning home in the late afternoon, or engaging in shopping and other leisure activities during the evening, explain the increase in traffic demand regularly observed in the late afternoon and during the evening. Such daily variations are not easily observable in the diagrams of Figure 2.2 due to the relatively short time intervals over which the stop line arrivals were observed.

In addition to the time-of-day and day-to-day variability discussed in the previous paragraphs, traffic variability can be further increased by the occurrence of special traffic events. One example of such increased variability can be observed in the second diagram of Figure 2.2. In this diagram, it is observed that no vehicle crossed the intersection where the data were recorded during a full cycle of signal operation at about 3:15 p.m. This is attributable to the passage of an emergency vehicle that completely halted traffic flow around the intersection for a few seconds. While no other special events were observed during the recording, other temporary changes in traffic demand patterns could also have been caused by nearby traffic incidents, temporary road or lane closures, etc.

The above analysis indicates that traffic demand is a time-varying stochastic process. If there were no random effects, some variability would still be observed in traffic demands, but this variability could be predicted with very good accuracy on a time-of-day, day-to-day or week-to-week basis. However, due to the presence of random effects, there is always a certain degree of uncertainty in traffic predictions. As a result, it appears desirable to allow traffic signal systems to automatically adjust the signal timings to traffic demands prevailing around each intersection. In such case, the adjustment capabilities of the traffic signal control system would ensure that the timings implemented at any given time truly matches the current needs.

2.2. Traffic Signal Control Philosophies for Urban Networks

Following the introduction of computer-based traffic signal control in the 1960's, numerous experiments were conducted for the purpose of developing signal control strategies more responsive to changes in traffic demands. One of the most comprehensive studies was the Urban Traffic Control System (UTCS) experiment carried out in Washington D.C. by the U.S. Federal Highway Administration throughout the 1970's (MacGowan and Fullerton, 1979-1980). The main purpose of this project was to develop and test a variety of advanced network control concepts and strategies. For comparison purposes, the strategies and concepts that were tested were divided into three generations, where each generation reflected a step in the development of traffic responsive signal control systems.

The following subsections use the same classification to retrace the evolution of traffic responsive signal control systems. In conjunction with the information shown in Table 2.1, these subsections successively present the first, second and third generations of traffic signal control philosophies.

Table 2.1 - Classification of Traffic Signal Control Strategies for Urban Coordinated Networks

Control Strategy	Sub-Generation	Characteristics			
		Cycle time	Traffic Prediction	Control Principles	Degree of Responsiveness
First Generation	Fixed-time (1-GC)	Fixed within each area	No	• Use of pre-stored timing plans calculated off-line based on historical traffic data. • Plan in operation selected by time-of-day or manual activation.	• Insensitive to changes in traffic demand.
	Plan Selection (1-GC/Traffic Responsive)	Fixed within each area	No	• Use of pre-stored signal timing plans calculated off-line based on historical traffic data. • Plan in operation selected from a library based on recently measured traffic conditions (traffic volume, detector occupancy).	• Use of pre-defined plans. • Timing plan selected according to recent traffic conditions. • Choice limited by the number of pre-defined plans in the library.
	Local Adaptation (1-GC/Traffic Responsive)	Fixed within each area	No	• Use of pre-stored timing plans calculated off-line based on historical traffic data. • Use of traffic actuation principles to adjust timings at individual intersections (green interval extension or early end based on gap size between successive detected vehicles, phase skipping when no vehicle is detected).	• Use fixed-time plan, but can make temporary alterations each signal cycle (within given limits) to adjust these plans to existing demand.
	Plan Generation (1.5-GC)	Fixed within each area	No	• Use of pre-stored timing plans calculated off-line based on historical traffic data. • Timing plans generated automatically off-line when traffic conditions warrant it and implemented after operator approval.	• Automatically generates fixed-time plans using recent historical traffic demand. • 15-minute minimum interval between changes.
Second Generation	Real-time signal control (2-GC)	Fixed within variable group of intersec-tions	Yes, but historically based	• On-line optimization of timing plans based on predicted trends in traffic conditions. • Stepwise transitions among timing plans. • Optimization repeated every five minutes.	• Generates the next plan to implement based on predicted changes in demand. • Prediction of future traffic demand as a function of historical volumes and speeds. • Ten-minute interval between plan modifications.
Third Generation	Fully demand-responsive, real-time control (3-GC)	Variable in time and among different intersec-tions	Yes	• On-line optimization of signal timings with very short sampling between updates. • Optimization repeated every one to two minutes, more often in some systems. • Cycle time, green allocation and signal offset, if used, are allowed to change continuously among intersections.	• Generates signal timings based on predicted future arrivals. • Future arrivals predicted by using current detector data or by smoothing them with past measurements. • System constantly in transition as it responds to changes as they occur. • 5 to 120 s between signal modifications.

2.2.1. First Generation Control

First generation control strategies use a catalogue of fixed timing plans to control a network of intersections. Each plan in the catalogue is calculated off-line based on historical traffic data before being stored in a computer's database. The plan controlling the traffic system at a given time can then selected on the basis of time of day, by direct operator intervention, or by matching a plan from the existing catalogue best suited to recently measured traffic conditions (usually traffic volumes and detector occupancies).

Signal timing plan generation for first generation control can be done using any available off-line signal optimization method. In practice, two main optimization methods are used. Methods of the first type, known as direct signal optimization methods, primarily attempt to minimize a performance function considering stops, delays and any other relevant parameters. These methods are best exemplified by the TRANSYT-7F traffic signal optimization model (Robertson, 1969; Wallace and Courage, 1992). Methods of the second type, known as progression methods, attempt to align the green intervals along successive intersections on a given route in such way that vehicles can travel across as many intersections as possible without being stopped by a red signal indication. These optimization methods indirectly attempt to minimize stops and delays by performing a geometrical analysis of traffic progression patterns. They are best exemplified by models such as PASSER II (Messer et al., 1974, Chang et al., 1985; Chang and Messer, 1990; Malakapalli and Messer, 1993), MAXBAND (Little et al., 1981; Chang et al., 1989; Chaudhary et al., 1991), MULTIBAND (Gartner et al., 1990, 1991), and PASSER IV (Chaudhary and Messer, 1993).

In all cases, the objective of the signal optimization is to determine a set of optimal values for the following four parameters at each intersection (see Figure 2.3):

- The cycle time or period needed to complete the full sequence of signal indications. In first generation control systems, all intersections within a given control area are usually required to operate under the same cycle time to ensure the repetitiveness of timing patterns between adjacent intersections.
- The green allocation, which divides the total available green time within the cycle time at an intersection to the various phases in use at that intersection.
- The signal offset, or moment when the main green interval (or any other defined point within the signal cycle) should start at the intersection in relation to a reference system time.
- The phase sequence, or order in which the phases are to be displayed within the signal cycle.

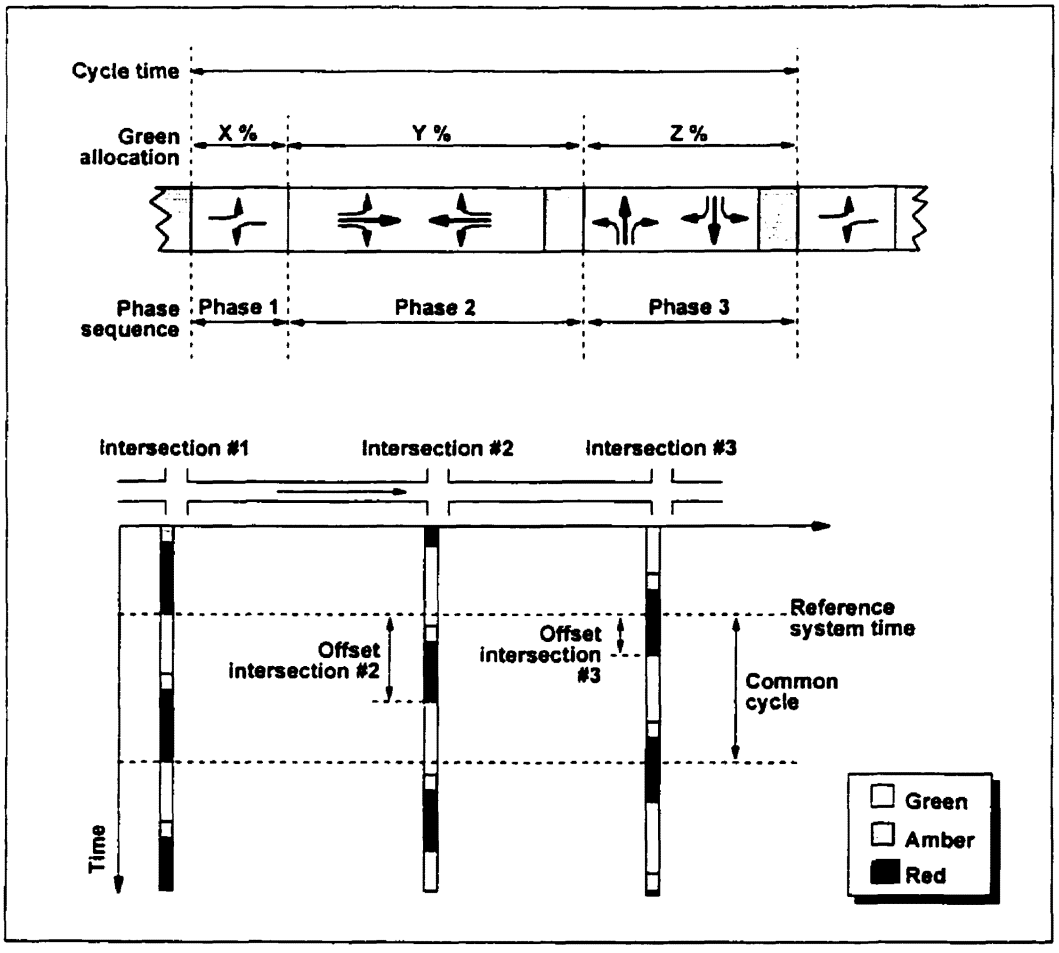


Figure 2.3 - Notions of Cycle Time, Green Allocation, Phase Sequence and Signal Offsets

Figure 2.4 illustrates the result of a typical first generation control strategy in which plan selection is done on a time-of-day basis. In this example, six different plans are implemented during the day based on major observed historical traffic variations between peak and off-peak periods.

At the beginning of each new control period, there is usually a transition period during which the new timing parameters are gradually implemented. Sudden changes from one plan to another are not made to avoid creating significant disturbances in traffic patterns. Within each control period, the plan in operation is designed to match the projected maximum traffic demand expected to occur. Maximum, and not average, demands are used to ensure that sufficient traffic capacity will be available at each intersection to serve all the vehicles that could arrive during any signal cycle. Since random fluctuations cannot all be predicted, the demand levels used in the design of the various signal timing plans are also sometimes artificially inflated to create a certain margin of safety within the traffic signal control process.

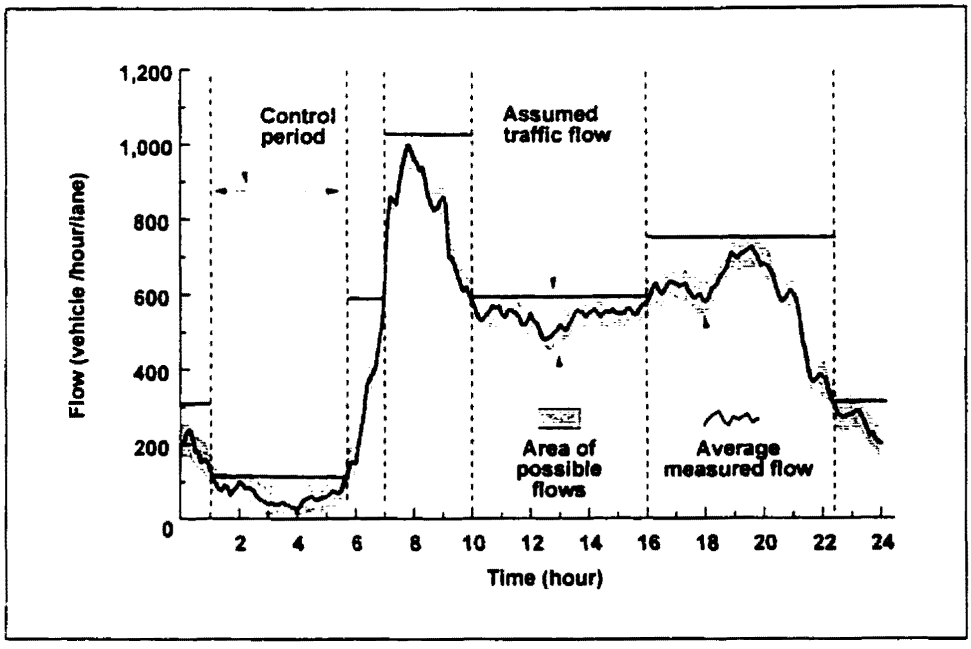


Figure 2.4 - Typical First Generation Traffic Signal Control with Time-of-Day Plan Selection

In the traffic responsive mode, the number of predefined timing plans stored in the catalogue and the time required to update the signal timing plans limit the frequency at which the plans in operation can be changed. In this case, plan updates can typically be done every 15 minutes. Some systems also possess a critical intersection feature enabling the duration of green intervals to be adjusted to current demands by traffic actuation. The principle of traffic actuation is as follows: once a minimum level has elapsed, the green signal display is extended until a preset maximum interval is reached or until the interval between two successive arrivals on all approaches on which traffic flow is monitored exceeds a certain threshold. An example of a first generation, traffic responsive control system is SCATS (Sims, 1978; Sims and Dobinson, 1979; Fehon and Moore, 1982; Lowrie, 1982; Sims and Finlay, 1984; Lowrie, 1990; Wood, 1993).

Hardware limitations initially dictated that first generation signal control systems be designed to rely only on a finite number of timing plans to control traffic in a given area. When these systems were first developed, limitation in computer technology constrained the number of timing plans that could be processed or stored in memory. While this limitation has now largely been removed, practical reasons currently explain why many traffic agencies still use first generation signal control principles, and more particularly a time-of-day operation.

First, while it is theoretically possible to design timing plans for all frequently occurring situations, it can be argued that the number of situations that might arise in practice is too great to allow the calculation of a signal timing plan for each one of them beforehand. Consequently, the attention is focused only on the main control situations. Second, an extensive amount of information needs to be collected to perform signal optimizations and traffic agencies' limited staff resources are often not enough to allow the compilation of all the data required to generate a large number of timing plans. This point is particularly true if the agency does not possess a centralized traffic monitoring system. In such case, staff must be sent to each intersection to collect the required data. Third, the number of detectors required to implement traffic responsive signal control systems make the systems costly to implement and expensive to maintain. In regions that that experience very low temperatures and rely on inductive loop detectors installed in the pavement, the difficulties in keeping the loop operation may even be extreme (low temperature pavement cracking and shearing of the loop wires). Finally, at the urban transportation system level, the predictability of travel times imposed by fixed-time network control systems is often considered a valuable tool in maintaining demand and supply equilibrium in traffic operations. In this case, the resulting stability of traffic patterns notably allow the traffic signal control systems to perform more readily a system-wide optimization of all timing parameters.

A remaining limitation of first generation signal control systems is their inability to automatically update the database of signal timing plans, as changes in traffic demands would warrant it. As previously indicated, the amount of information that needs to be collected to generate new signal timing plans is often too great to justify the frequent updating of signal timing plans. As a result, the timing plans used by first generation signal control systems are often only re-optimized when citizen complaints reach the point of forcing something to be done (NCHRP, 1992). In Toronto, for instance, the timing plans in operation in the various parts of the city are typically reviewed and upgraded once every five or more years, with the reviewing process usually initiated only when a complaint is received (Kelman et al., 1993). For that city, it was estimated that a staff of 30 full-time employees would be required to update the signal timings annually and that the use of so many employees for such a specific task would be well beyond the available resources. The use of plans that are often several years old may result in increased stops and delays for vehicle passengers, as the signal timing plans may no longer be optimal. A study reported by Hunt et al. (1981) suggests for instance that the loss of optimality that results from the aging of first generation traffic signal timing plans could translate into a four to five percent increase in delays every year when compared to the use of up-to-date plans. For five-year old plans, this could translate into increases exceeding 20 percent, which is far from being negligible.

This particular problem was solved with the introduction of first generation signal control systems possessing plan generation capabilities. Like other first generation systems, these systems use a library of pre-defined timing plans to control a network of intersections. In this case, however, the plans stored in the library are automatically updated by the system when new traffic conditions warrant it. The new plans are generated off-line using historical and recently measured traffic data and then subjected to operator approval before being implemented or stored in the database.

2.2.2. Second Generation Control

Second generation strategies move away from the use of pre-stored timing plans. Traffic signal control systems belonging to this category compute and implement timing plans on-line based on average surveillance data and projected future traffic conditions. Unlike first generation systems, there is no limitation on the number of timing plans that can be generated. New plans are calculated every five minutes. This allows second generation systems to better follow changes in traffic demands, especially during transient periods between peak and off-peak periods. However, plan changes cannot be initiated at intervals of less than 10 minutes. This constraint is imposed to avoid large transition disturbances. Similar to first generation systems, second generation systems also force groups of intersections to operate under the same cycle time. However, the intersection grouping is allowed to vary to provide increased flexibility in adjusting the signal timings to traffic conditions prevailing in the controlled network. A typical example of an advanced second generation signal control system is SCOOT (Hunt et al., 1981; Robertson, 1986; Wood, 1993; Bowen et al., 1994; Bowen and Bretherton, 1996; Bretherton, 1996).

2.2.3. Third Generation Control

Third generation control strategies implement fully responsive on-line traffic signal control. Similar to second generation strategies, these strategies compute traffic signal timing plans in real-time based on current and predicted traffic conditions. In the UTCS experiment, third generation control strategies included only traffic signal control systems that could automatically re-optimize traffic signal timings every two to three minutes. However, this generation is now often extended to include systems that can perform signal optimizations as often as at every five seconds. Examples of third generation systems are UTOPIA (Donati et al., 1984; Mauro and Di Tranto, 1989; Mauro, 1991; Davidson and Di Taranto, 1992; Wood, 1993) and PRODYN (Henry et al., 1983; Henry and Farges, 1984; Barrière et al., 1986; Henry and Farges, 1989; Kessaci et al., 1989; Wood, 1993).

Unlike previous generations, third generation strategies allow the cycle time to vary from one intersection to the other, as well as from one moment to the next at the same intersection. Green allocation and signal offset can also be freely adjusted at individual intersections based on the needs of local traffic demands. In most recent systems, traffic signal control is even accomplished on a phase-by-phase basis, without any explicit reference to the notion of cycle time, green allocation and offset. In this case, the main control decisions are when to end the current phase and which phase to go to next.

2.3. Transit Operations in Urban Networks

Transit vehicles and passenger cars share the right of way in many cities. This mixed operation is the source of many flow disruptions, as both types of vehicles do not share the same travel requirements. While transit vehicles must stop at regular intervals along a given route to board and discharge passengers, passenger cars usually attempt to travel the same route without being stopped. The goal of this section is to clearly define the transit operations to be addressed in this research, and more specifically, the aspect of these operations that might affect the progression of general traffic. It successively presents:

- The characteristics of transit movements along urban streets,
- The effects of transit vehicles blocking one or more traffic lane while they are stopped to board and discharge passengers,
- The various types of treatments that can be considered to give priority of passage to transit vehicles at signalized intersections.
- The technologies that currently can be used to selectively detect transit vehicles within a group of mixed vehicles.

2.3.1. Characteristics of Transit Movements

The need for transit vehicles to stop at regular intervals along a given route to board and discharge passengers results in a general inability for these vehicles to stay in the main traffic stream. It also results in an inability to benefit from the general signal coordination, as traffic signal timings are often designed to accommodate the general traffic only. This problem is illustrated in the example of Figure 2.5, in which a transit vehicle is shown to leave an upstream intersection within a platoon of vehicles. However, this vehicle falls behind the platoon when it stops at mid-block. When it reaches the downstream intersection, it arrives just too late to cross that intersection during the same green interval

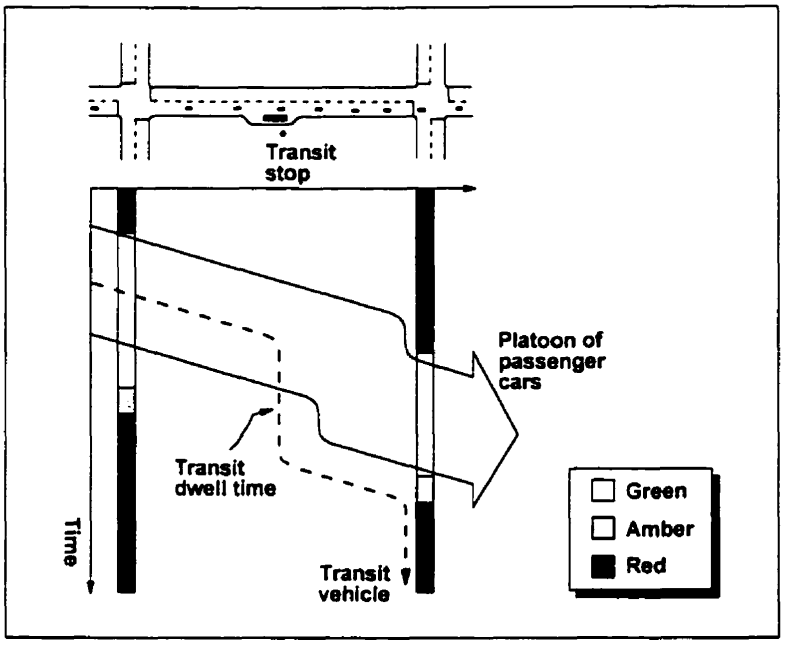


Figure 2.5 - Typical Passenger Car and Transit Vehicle Movements along Urban Streets

as the platoon. Therefore, the transit vehicle is forced to stop a second time on the approach and to wait for almost the entire duration of the red interval before being able to resume its progression.

In addition to this inability to stay in the main traffic stream and to benefit from signal coordination, transit travel times along urban streets are subject to dwell time variability in addition to the normal travel time variability experienced by passenger cars. As a result of fluctuations in the number of passengers loaded and unloaded at each stop, transit dwell times are not constant. Figure 2.6 illustrates the distribution of dwell times at a typical transit stop. These observations were made on February 23 and 24, 1995 at a transit stop on King Street West in downtown Kitchener between Ontario St. and Queen St. These data first indicate that 42.6 percent of all buses did not stop to board and discharge passengers. They also indicate that while a mean dwell time of 9.8 seconds was observed for all buses that stopped, individual dwell times varied over a wide range - going from six seconds to 16 seconds. The requirement that transit vehicles stop at regular intervals and the fact that dwell times at these stops vary considerably makes it difficult to predict the degree of progression each transit vehicle will experience. Consequently, it is very difficult to predict with great accuracy the average travel time of transit vehicles along urban streets.

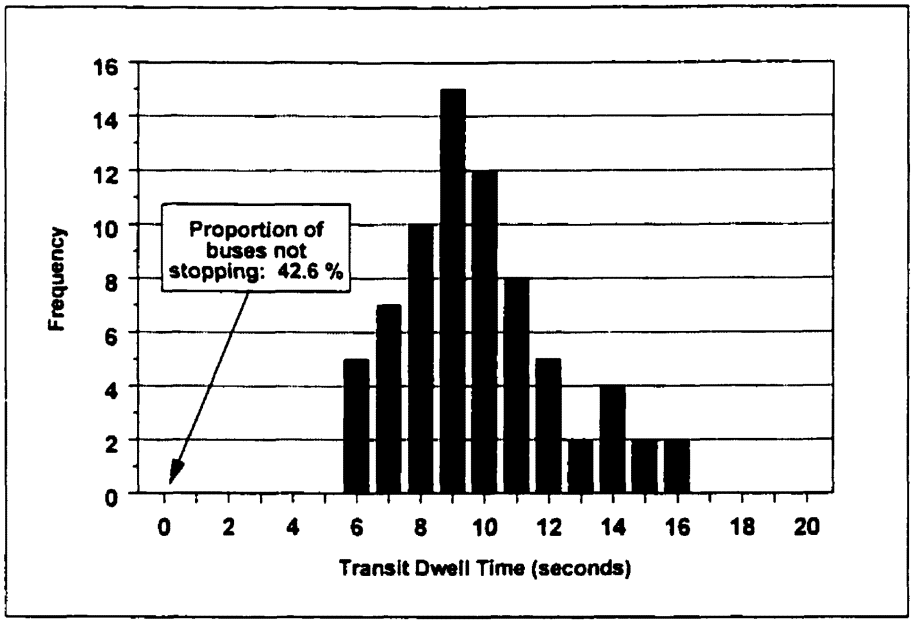


Figure 2.6 - Observed Dwell Times at a Typical Urban Transit Stop

2.3.2. Transit Interference in Shared Right of Way

Transit activities carried out in the right of way cause interference to the general traffic. The level of interference depends on several factors, such as the number and location of transit stops, the time required to board and discharge passengers at each stop, and the volume of transit vehicles. However, the most important factors are the facilities built along the streets to allow transit vehicles to load and unload passengers as well as the number of traffic lanes in the direction of travel (Jacques, 1993):

- The effect of a stopped bus is generally small where stops are located in a bus bay or in a no-parking zone along the side of the street (Figures 2.7a and 2.7b). In such cases, deceleration and acceleration of buses in the curb lane are the main sources of disruption. However, the overall effect of these disruptions is usually small enough to be neglected.
- Buses stopping within a traffic lane may cause more severe disruptions. If there were only one lane in the traveled direction, any stopped bus would block the whole approach and hold up all other traffic behind it (Figure 2.7c). If there is more than one lane, then only the curb lane would be blocked (Figure 2.7d). In such cases, the number of vehicles able to pass the stopped bus would depend on how traffic spreads out among the lanes behind the vehicle and on drivers' behavior around the temporary obstacle.

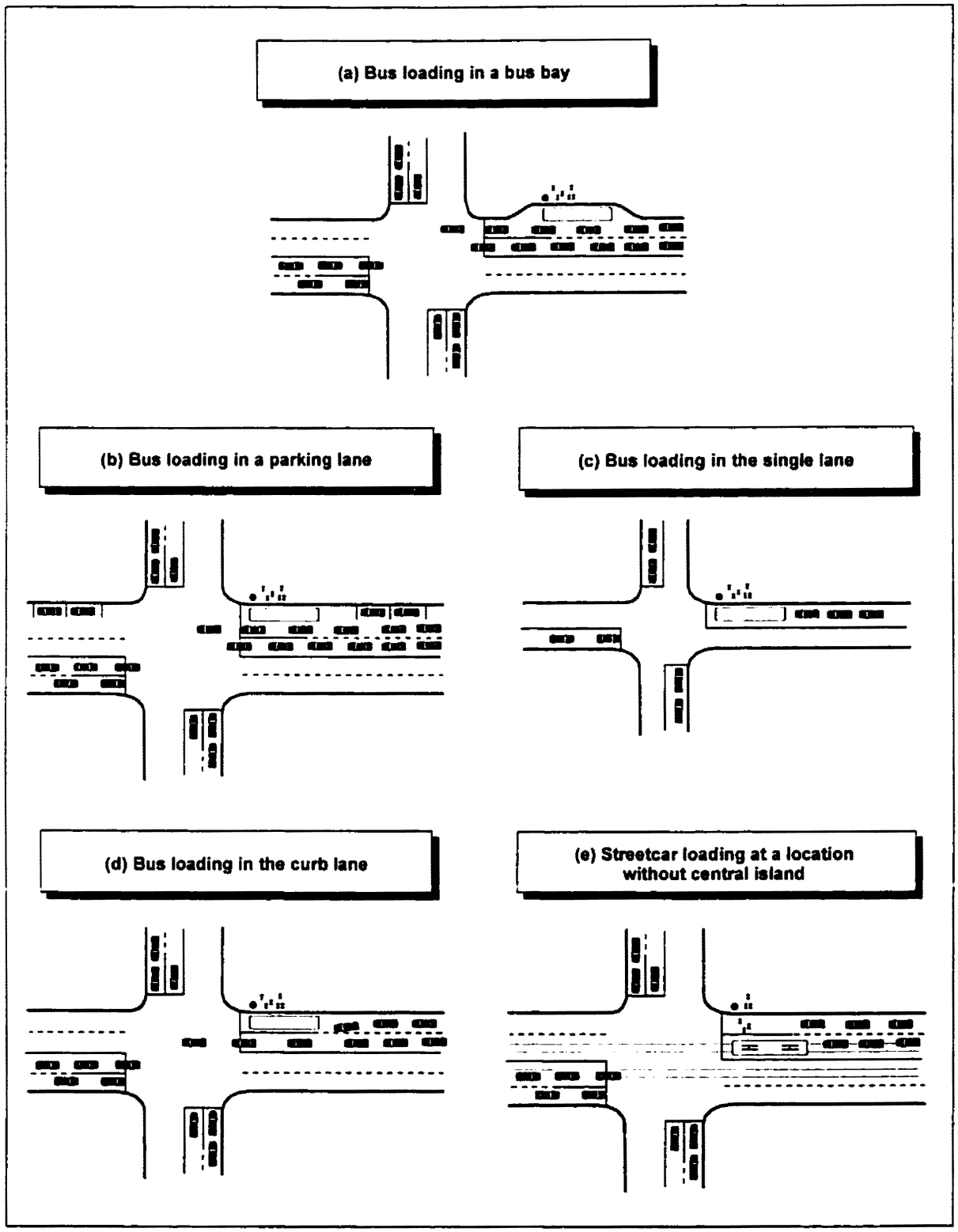


Figure 2.7 - Typical Transit Loading Activities near Signalized Intersections

- While other alternatives are possible, streetcars usually run in the median lane of urban streets and must therefore stop in that lane to board and discharge passengers. If a central island is provided, streetcars' passengers can use this place as a refuge and a stopped streetcar will then only affect the vehicles traveling in the median lane. However, if the street does not have a central island, all traffic behind the streetcar in all traffic lanes must stop to ensure that passengers safely reach the side of the street (Figure 2.7e).

The main effect of these disruptions is to delay the progression of passenger cars in the general traffic stream. This delaying effect may then negatively affect the traffic signal operation. However, these considerations do not mean that transit vehicles are an undesirable element of the traffic stream. The word interference is only used in this thesis to reflect the effects that these vehicles might have on the progression of passenger cars and trucks while stopped in the right of way. For instance, consider the example of Figure 2.8, in which a transit vehicle is traveling along a one-lane approach and must stop at mid-block to board and discharge passengers. In the first case, illustrated in the left-hand side of the figure, a bus bay is provided and the bus is able to pull out of the roadway in front of the transit stop to board and discharge passengers. As a result, there is no interference caused to other traffic and all vehicles are able to cross the downstream intersection before the end of the green interval. In the second case, no bus bay exists and the bus must stop in the right of way, causing a complete blockage

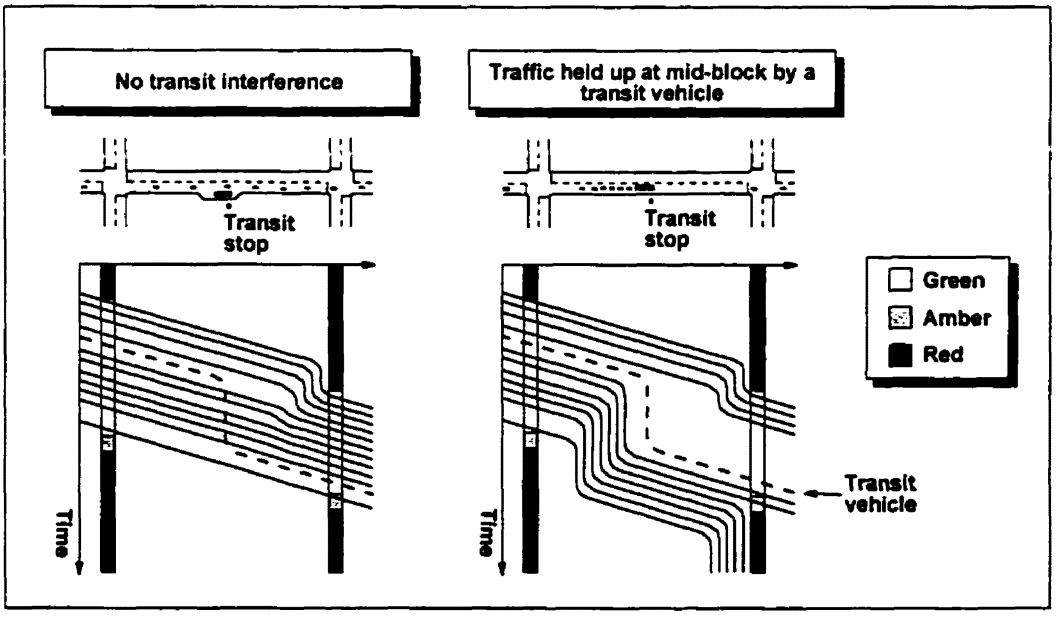


Figure 2.8 - Example of Transit Interference on General Traffic Progression along Urban Streets

of the approach for the entire dwell time. In this case, a significant portion of the traffic is held up behind the transit vehicle and is delayed to such extend that only a few vehicles are then able to reach the downstream intersection before the end of the green interval.

2.3.3. Transit Priority at Signalized Intersections

Delay at traffic signals is one of the largest components of transit delay on arterial streets. For example, Evans and Skiles (1970) found that the delays incurred by buses at signalized intersections comprise between 10 and 20 percent of their overall trip time and nearly 50 percent of the total delay that each vehicle experiences. This delay is particularly significant when performance measures are compiled on a person basis rather than a simple vehicle basis. For instance, consider that an average bus holds 60 persons and an average passenger car 1.5 person. In this case, each bus being stopped as a result of the operation of traffic signals will result in the same total amount of person delay as if 40 passenger cars were being stopped.

In many cases, the delays caused to transit vehicles by the operation of traffic signals at urban intersections can be reduced by providing preferential treatment to these vehicles. The main immediate benefit of such treatment would be a reduction in the number of stops and the amount of delay incurred by transit riders along transit routes. A potential long-term benefit also includes a shift in mode choice, as drivers are enticed from using their passenger cars to using transit vehicles. The reduction in passenger car use would also result in a reduction of fuel consumption and emissions. General improvements in transit system performance through travel time reductions and better adherence to schedules may also be observed (Sunkari et al., 1995).

On the opposite side, providing priority to transit vehicles at signalized intersections may increase stops and delays to cross street traffic. Such increase would be caused by the allocation of more green time within a given signal cycle time to the approaches on which transit vehicles are traveling, causing less green time to be available to serve traffic on the other approaches. In some cases, these cross-street stop and delay increases may even completely offset the benefits of providing priority of passage to transit vehicles (Chang et al., 1995). Consequently, transit priority treatments should only be implemented after all possible trade-offs have been evaluated and when it has been demonstrated that overall benefits can be obtained, on an overall person basis or vehicle basis.

As an example, the studies conducted by Vincent et al. (1978) and Richardson and Ogden (1979) report overall benefits from the application of transit priority strategies. On the other hand, the evaluations conducted by Reedy and Ashworth (1978) and Al-Sahili and Taylor (1996) indicate that the use of transit priority may also result in an overall increase in vehicle and person travel time and delays. In the later case, Al-Sahili and Taylor found that the overall benefits gained by the transit vehicles from the various simulated priority treatments along Washtenaw Avenue in Ann Arbor, Michigan, were not sufficient to counter the delay increases caused to other vehicles in the traffic stream as a result of increased disruptions in traffic progression patterns. Significant increases in cross-street delays were also observed by Cornwel et al. (1986) in a traffic responsive signal control system. In this case, the authors tested vehicle-responsive priority strategies within the SCATS real-time traffic signal control system in Melbourne, Australia. In their field experiment, they found that while travel times for transit vehicles and passenger cars were respectively reduced by ten and seven percent along the main direction of travel, increases in travel time totaling as much as 13 percent were observed on some cross-streets. However, despite this adverse effect, the authors judged that the results were not sufficient in concluding that the tested strategy was unsuccessful. As a final example, Benevelli et al. (1983) also found that the benefits of coordinating traffic signals along an arterial decrease with the implementation of preemption strategies at a greater number of intersections, and that such decrease could eventually result in overall vehicle delay increases.

Tables 2.2 and 2.3 summarizes the various treatments that have been designed over the years to reduce the delaying effects of traffic signal operations on transit operations. These treatments are categorized as being either passive (Table 2.2) or active (Table 2.3) based on how the needs of transit vehicles are considered.

Passive priority treatments do not explicitly recognize the actual present of transit vehicles on intersection approaches. They use historical average traffic patterns to deduce typical transit vehicle arrival times at each intersection and then use these average times to generate signal timings that will favor their particular movements. These treatments are not traffic responsive since they do not rely on the detection on approaching transit vehicles. Passive priority treatments introduce a permanent bias in the signal operation in favor or transit operation. Consequently, it is only on average that transit vehicles benefit from these treatments. Individual vehicles arriving at an intersection within the predicted average will generally be able to benefit from the transit-weighted traffic signal operation, but not necessarily vehicles arriving too late or too early. In addition, while the transit-weighted timings may often also benefit the general traffic traveling on the prioritized approaches, they may create some

disadvantage for the vehicles traveling on other links. This problem is particularly true when there is currently no need for the implementation of transit priority treatments. As a result, while the operation some signalized networks may benefit from the implementation of transit priority treatments, the negative effects on general traffic may completely offset the benefits in other networks.

Active priority treatments improve on passive strategies by altering the signal timings only when such action is required. Instead of using historical data to predict average transit behavior, priority treatments are granted only when an approaching transit vehicle is detected. One general concern about active transit priority is that sudden changes in signal settings to accommodate a transit vehicle on an

Table 2.2 – Passive Transit Priority Treatments at Signalized Intersections

General characteristics	Possible actions
• Do not explicitly recognize the presence of a bus. • Predetermined timing plans permanently favoring the approaches on which transit vehicles usually travel.	**Adjustment of cycle time** • Reduction of cycle time to reduce the delays incurred by transit vehicles having to stop at the intersection. Since this treatment reduces the capacity of an intersection, care should be taken not to reduce the cycle time up to a point resulting in traffic congestion. **Phase splitting** • Splitting a priority movement into multiple phases within a given signal cycle. This can reduce transit delays without necessary reducing the cycle time, but may break progression along an arterial. **Permanent special transit phase** • Short phase permanently inserted into the normal signal operation allowing exclusive movement through the intersection in such way that transit vehicles can make their movements while all other traffic is stopped. **Transit-oriented area-wide timing plans** • Priority through preferential progression schemes. This is accomplished by setting the offsets between adjacent intersections using transit average travel times rather than passenger car travel times. Converting transit vehicles into passenger car equivalents may also be used to justify the allocation of more green time to the approaches carrying transit vehicles. **Vehicle metering** • Flow regulation through a network by limiting the number of vehicles into the system. Transit vehicles benefit from metering if they are allowed to bypass the metered traffic signals by using special reserved lanes, special signal phases, or alternate routes through non-metered traffic signals. **Permanent queue jumps** • Installation of special reserved lanes allowing transit vehicles to bypass the queues of vehicles that form at the stop line of signalized intersections.

Table 2.3 – Active Transit Priority Treatments at Signalized Intersections

General characteristics	Possible actions
Control principle ▪ Activation of a new traffic signal timing pattern overriding the existing one after the detection of an approaching transit vehicle. **Unconditional priority** ▪ Describes the provision of signal priority each time it is requested by a transit vehicle. **Conditional priority** ▪ Considers other factors in determining if priority will be granted to approaching vehicles: vehicle occupancy, cross-street queue length, current traffic conditions, time since last priority, effect on coordination, point in signal cycle at which the bus is detected, etc.	**Phase extension** • Extension of the green signal indication for the priority movement until a maximum duration is reached. **Phase recall** • Advance of the start of the green interval by prematurely ending other non-transit phases. This may be constrained by providing a minimum green interval for the phases to be prematurely terminated. **Phase skipping** ▪ Temporary omission of one or more non-priority phases from the normal phase sequence to facilitate the provision of the transit priority phase. This treatment should be used only when traffic demand on the skipped phase(s) is low. **Phase flipping** ▪ Change in the order in which phases are displayed within the signal cycle. The phase sequence is temporary rearranged to display a green signal indication on the approach on which a transit vehicle is approaching when this vehicle is projected to reach the stop line. **Phase repetition** ▪ Repetition of a phase, entirely or in part, after its normal activation within a given signal cycle. **Special transit phase** • Short phase inserted into the normal signal operation allowing exclusive movement through the intersection in such way that transit vehicles can make their movements while all other traffic is stopped.

approach may significantly disrupt progression patterns on other approaches. Signal coordination and transit priority objectives are often contradictory. The former attempts to accommodate the general traffic going through sets of intersections, while the later attempts to facilitate the progression of individual vehicles at specific intersections. A trade-off must be made between signal coordination and transit priority objectives when active schemes are used. However, since each signal control option has stop and delay implications, this trade-off is theoretically easy to determine when stops and delays can be estimated by the signal control system.

Active priority treatments are further divided into conditional and unconditional treatments. In conditional priority treatments, incoming transit vehicles only receive priority if the required signal changes will not reduce the overall performance of the intersections. In unconditional priority treatments, preferential treatments are granted to any incoming transit vehicles regardless of the consequences on the performance of the intersection.

Conditional active priority is theoretically better than unconditional priority. Initiating signal changes based only on the detection of transit vehicles might not be harmful at intersections where there is a relatively low volume of transit vehicles, but can be in all other cases. For example, if green interval extensions and phase early recalls are granted frequently to accommodate transit vehicles traveling on a given approach, excessive delays and queues may then result on other approaches as a consequence of the availability of shorter green intervals on these approaches, the implementation of signal switches when platoons are being served, etc. The generation of longer queues on some approaches may even make it more difficult for the traffic signal control system to accommodate transit vehicles arriving later in the control period by increasing the amount of time required to serve the queues. This is especially true if the queues are located on the same approach as the transit vehicle requesting priority of passage and must be cleared before that vehicle can cross the intersection.

After having considered many of the above elements, Wood (1993) concludes in his review of existing urban traffic signal control systems that research is still needed into the most effective form of priority. He notably indicates that research is still needed toward the development of signal control strategies that can offer priority to transit vehicles without the need for sudden forced changes of signal settings when these vehicles are approaching an intersection. This implies tracking vehicles' movements and predicting their intersection arrival time at least tens of seconds before their actual arrival time so that preferential treatments can be considered and planned when the signal timings are first generated.

2.3.4. Selective Detection Technologies

The implementation of transit priority treatments in mixed-traffic environments requires the use of a selective traffic detection system capable of differentiating transit vehicles from the general traffic. While there are numerous technologies that can be employed to identify vehicle types, only a few types have been implemented in the field for signal priority systems for transit vehicles. Concerns relating to the stability of the detection zone and the reliability of the detection process, as well as implementation and maintenance costs, have been major deterrents in many cases (Metro Toronto Transportation et al., 1996).

According to a traffic signal priority study conducted by the Municipality of Metro Toronto, the Toronto Transit Commission and the Ministry of Transportation of Ontario (Metro Toronto Transportation et al., 1996), the following technologies were available in 1992 for performing selective transit detection:

- Ultra sonic detectors
- Video image processing
- Conventional induction loops
- Loop detector signature processing
- Passive transponder
- Active transponder
- Infrared and microwave
- Optical emitters

Ultra sonic detectors are installed above a lane and are fine-tuned to detect only vehicles above a specific height. A major drawback of this technology is that it will provide priority to all vehicles satisfying the minimum present height.

Video image processing involves the use of roadside television cameras to digitally compare the video signal of approaching transit vehicles with bus templates. Priority is provided if the signature of an approaching vehicle matches a template. While there has been applications in video image processing for general real-time traffic detection (Michalopoulos, 1993; Lialias, 1998), the technology is not currently being used in applications of transit signal priority systems as one of its main drawbacks is its very high cost (Metro Toronto Transportation et al., 1996).

Conventional inductive loops are typically formed by putting two to three turns of electrical wire in slots cut into the pavement in the middle of the traffic lane for which vehicle detection is required. Vehicle presence is then determined by monitoring the changes in the loop's electric current induced by the metallic body of passing vehicles. With this technology, transit priority calls can be achieved by installing loops the size of a bus on intersection approaches and tuning them to place priority calls only when a vehicle of sufficient length passes over them. Similarly, multiple loops actuated in a specific order with a minimum period of simultaneous actuation can be used to request priority. The main advantage of this technology is that loops are relatively inexpensive to install. However, there still exists the problem of differentiating trucks from transit vehicles. As a result to the high probability of false detection, conventional inductive loops are usually only used with simple priority algorithms and at locations where the negative effects of false priority requests are minimal.

Loop detector signature processing classifies vehicles by examining the changes in the loop inductance as each vehicle passes over it. The signature of each vehicle is analyzed and priority is provided if

there is a sufficient match to a predefined signature. Currently available recognition equipment is capable of distinguishing vehicles among several categories, including several types of transit vehicles (Cheng et al., 1979), and is relatively inexpensive.

Passive transponders are on-board equipment attached to a bus, which interact with inductive loops put into the pavement and pole mounted or overhead antenna. When the transponder receives a signal from an emitter, it modifies the signal and reflects it back to the antenna. The modified signal contains a unique code identifying the vehicle, allowing it to be selectively identified. As a result of the many potential applications of this technology and its high degree of effectiveness, there is considerable ongoing development regarding its use in selective detection systems.

Active transponders, like passive transponders, are mounted to a bus. However, unlike passive transponders, active transponders transmit a continuous signal. Selective detection then occurs when a wayside antenna receives the signal.

Infrared and microwave selective detection systems also require the use of on-board equipment. This equipment is either activated by the vehicle operator or continually transmits a signal. Similar to passive and active transponders, selective detection occurs when the signal is received by a pole-mounted receiver and transmitted to the signal controller. Unfortunately, there is little information currently available about the effectiveness of this technology.

Optical emitters involve the use of light beams transmitted from transit vehicles. Selective detection occurs when a receiver detects the light beam. There are numerous variations of this technology. Some of these variants have the capability to identify specific buses though the use of coded retro-reflective plates mounted on buses. As this is an optically based system, natural lighting and prevailing weather conditions may affect the stability of the detection zone.

Table 2.4 summarizes the results of the comparative evaluation of selective detection systems conducted in 1992 by the Municipality of Metropolitan Toronto, the Toronto Transit Commission and the Ministry of Transportation of Ontario (Metro Toronto Transportation et al., 1996). In this evaluation, effectiveness refers to the ability of the selective detection system to satisfy the functional requirements of a transit priority system. The risk identifies the extent to which the hardware and/or software have been previously employed and shown to be effective in similar applications and environmental conditions. Flexibility is related to the ability of each alternative to be employed in

Table 2.4 - Comparison of Selective Detection Technologies

(Source: Metro Toronto Transportation et al., 1996)

Technology	Installation Cost	Availability	Effectiveness	Maintenance Cost	Risk	Flexibility
Ultra sonic detectors	Low	Available	Low	N/A	Low	Low
Video image processing	Very high	Available	Moderate	N/A	High	Low
Conventional induction loops	Low	Available	Low	N/A	Low	Low
Loop detector with one-signature processing	Moderate	Available	Low	N/A	Moderate	Moderate
Loop detector with three-signature processing	Moderate	Near Future	Moderate	N/A	Moderate	Moderate
Passive transponder	Moderate to high	Available	High	N/A	Low	High
Active transponder	Moderate to high	Available	High	N/A	Low	High
Infrared and microwave	High	Available	Low	N/A	Low	Low
Optical emitters	High	Available	Moderate	N/A	Low N/A	Moderate

Note: N/A indicates that the information is not available

unusual locations and support sophisticated priority algorithms. Based on this evaluation, a selective detection system based on the use of active or passive transponder was judged the most effective and flexible. While such systems are more costly than systems employing less effective technologies, they are comparable to the moderately effective options and have the advantage of already being successively used to provide priority in various operating environments.

To conclude this chapter, it can be observed that numerous efforts have been made to try to overcome the various challenges posed by real-time traffic signal control in urban networks. In order to provide automatic adjustment capabilities to traffic demand variations three different general signal control philosophies have been developed over the years. Other efforts have been directed to the additional complexities associated with providing priority of passage to transit vehicles at signalized intersections in networks in which passenger cars and transit vehicles share the right of way. While the discussion conducted in this chapter generally focused on traffic signal control principles, the next chapter will provide a more detailed and critical look at the major real-time traffic signal control systems that have been developed over the past twenty years.

3. Existing Real-Time Signal Control Systems

Over the past two decades, four real-time, traffic responsive signal control systems have been developed with the explicit objective of controlling traffic in urban signalized networks. These systems are:

- SCATS (Sims and Dobinson, 1979; Fehon and Moore, 1982; Lowrie, 1982, 1990),
- SCOOT (Hunt et al., 1981; Robertson, 1986; Wood, 1993; Martin and Hockaday, 1995; Bretherton, 1996),
- UTOPIA (Donati et al., 1984; Mauro and Di Taranto, 1989; Mauro, 1990; Davidson and Di Taranto, 1992; Wood, 1993), and
- PRODYN (Henry and Farges, 1983, 1984, 1989; Barrière et al., 1986; Kessaci, 1988; Kessaci et al. 1989, Wood, 1993).

Other systems have also been proposed, such as the five following prototypes developed in the United States under the RT-TRACS project:

- OPAC (Gartner, 1982, 1983, 1989; Gartner et al., 1983, 1990);
- RHODES (Head et al., 1992; Dell'Olmo and Mirchandani, 1995; Head, 1995);
- ISAC (Owen et al., 1997);
- Maryland/Pittsburg Prototype (Owen et al., 1997); and
- University of Minnesota Prototype (Owen et al., 1997).

The goal of this chapter is to briefly describe the efforts that have been put in developing efficient real-time traffic signal control systems for urban networks. The chapter begins by describing the control philosophy behind each one of the four existing real-time systems mentioned above. This description is followed by an analysis of their ability to provide priority of passage to transit vehicles at signalized intersections and to consider transit interference on other traffic while generating new traffic signal timing strategies. For a more detailed description of the features of SCATS, SCOOT, UTOPIA and PRODYN, the reader is invited to consult the Appendix A.

3.1. Control Principles

3.1.1. SCATS

SCATS is the first real-time, traffic responsive signal control system to have been explicitly designed for the control of urban networks. This system, which can be viewed as an advanced first-generation signal control system, was initially developed in the early 1970's for the City of Sydney, Australia, by the Roads and Traffic Authority of New South Wales. Since then, it has been implemented in more than 30 cities around the world, including one North American installation in Oakland County, near Detroit, Michigan (Michalopoulos, 1993).

SCATS is a two-level hierarchical control system having the capability of responding to both time-of-day and cycle-to-cycle traffic demand fluctuations. At the upper level of the hierarchy, a central computer is responsible for calculating signal timings for each intersection in all coordinated areas based on the average traffic conditions prevailing in each area. Automatic response to traffic demand variations is achieved by instructing the central computer to incrementally adjust the signal timings once every signal cycle, i.e., typically every 40 to 120 seconds. At the lower level, individual signal controllers equipped with microprocessors are given the ability to further modify the timings they receive from the central computer to adjust them to the actual traffic conditions around each intersection.

The main control objective of the SCATS system is to minimize the number of stops and the amount of delay incurred by vehicles in the network under control, and more particularly the number of stops incurred along major routes. This strategy is reflected in the fact that the primary timing decisions are taken by the central computer. Establishing area-wide progression patterns is easier when all data are available at a single location. When the network is operating near saturation, this objective may also be biased to maximize the number of vehicles that can flow through the system in a given unit of time.

At the upper control level, coordinated traffic signal control is first achieved by dividing the network under control into a large number of comparatively small sub-systems containing no more than ten intersections. As far as possible, the sub-systems are chosen to be traffic entities that can run without relation to each other under a range of traffic conditions. As traffic conditions require it, adjacent sub-systems may be linked together or separated to form larger or smaller coordinated groups. Sub-systems are merged when it is established that improved coordination could be achieved by forcing them to

operate under the same cycle time. Similarly, sub-systems are separated when using different cycle times within each sub-system could improve traffic performance.

For each sub-system, the central computer generates signal timings for each intersection in the area using plan generation and plan selection algorithms. First, the cycle time and green allocation are determined for the critical intersection. This intersection is defined by the user and remains the same regardless of changes in traffic conditions. It is usually the one having the highest traffic load in the group. The cycle time for the intersection is calculated so as to maintain its degree of saturation below a preset value. The green allocation is calculated with the objective of maintaining an equal degree of saturation on all competing approaches. Once an optimal cycle time has been determined, it is imposed on all other intersections in the sub-system. The green allocation of other intersections is then selected from a library of predefined plans by a matching process that attempts to select plans that are compatible with the splits calculated at the critical intersection. Offsets are also selected from a library of predefined plans, with the objective of minimizing the number of stops incurred by vehicles along the direction carrying the highest average demand.

To ensure a certain degree of stability in signal operation, restrictions are imposed on the amount of change allowed in the timing parameters between each signal optimization. For example, changes in the cycle time are restricted to a maximum of nine seconds, while changes in the green allocation cannot exceed four percent of the current cycle time. Also, while new optimum offsets are calculated once every signal cycle for all intersections, a change at a particular intersection is only initiated when at least three of the five previous calculations have suggested a change.

At the lower level of control, signal controllers use vehicle-actuation logic to adjust the duration of green intervals to the actual demand. Green intervals are extended beyond their minimum duration until the observed gap between two successive arrivals at the stop line exceeds a preset value or until a maximum extension is reached. Phases for which there is currently no demand can also be entirely skipped. The only exception to this control scheme is that one phase, usually the main-street green interval, cannot be omitted or ended earlier. This constraint is implemented to preserve the cycle time determined by the regional computer and avoid compromising the progression scheme established through offset selection by the central computer. As a result, this green interval receives any unused green time resulting from the earlier termination of other phases.

To monitor traffic conditions around each intersection, SCATS relies on the use of inductive loop detectors installed at or near the stop line on every approach. In a typical installation, the detectors are located where vehicles would normally start to queue. These detectors allow the system to directly monitor the interval between successive arrivals at the intersection and to count the number of vehicles entering the intersection from each approach during any given phase. However, they do not allow the detection of incoming vehicles before their actual arrival at the intersection. To perform such detection, the detectors would have to be installed at the some distance upstream from the traffic signals. As a result, the traffic demands in the current and future signal cycles can only be estimated by extrapolating previously recorded traffic counts.

3.1.2. SCOOT

SCOOT is the second real-time traffic signal control system for urban networks to have been successfully developed. The Transportation and Road Research Laboratory of the United Kingdom designed this system, which can be viewed as a second-generation signal control system. It was first implemented in 1982 and is now being used in more than 165 cities around the world, including Toronto, Red Deer and Halifax in Canada. As of 1998, there were also three implementation sites in the United States, in Oxnard and Anaheim, California, and Montgomery County in Maryland.

SCOOT is a fully centralized traffic signal control system. In this system, a central computer performs all timing calculations. Contrary to SCATS, there are no signal timing adjustments made by local controllers, except for providing priority of passage to transit vehicles (see the description of transit priority features in Section 3.3). This design choice again reflects the system's main optimization objective, which is to globally minimize the stops and delays incurred by vehicles inside each coordinated area. As indicated in the previous section, global optimization is easier to perform when all data are available at a single location.

The main control philosophy of SCOOT is to react to changes in observed average traffic demands by making frequent, but small, adjustments to the cycle time, green allocation and offset of every controlled intersection. For each coordinated area, the central computer evaluates every five minutes if the common cycle time should be changed to keep the degree of saturation of the most heavily loaded intersection at or below 90 percent. To maintain some stability in the operation of the network, changes to the cycle time are limited to a maximum of eight seconds per optimization. A few second before each scheduled phase change, the signal optimizer also evaluates if the current phase should be

terminated immediately, as scheduled, or later. The optimizer implements at each intersection whichever alteration will minimize the maximum estimated degree of saturation on any approach to that intersection. In order to avoid large transition disturbances, changes in the green allocation of each intersection are limited to eight seconds. Once during each signal cycle, the optimizer also assesses whether altering the offset of each intersection by up to eight seconds can reduce the stops and delays around each intersection.

Contrary to SCATS, SCOOT detects incoming traffic at some distance upstream from the stop line. In typical installations, the detectors are located on each approach at the exit of the upstream intersection. This location provides the system with maximum advance information about future vehicle arrivals without the need to consider turning movements at the upstream intersection. It also allows SCOOT to detect queues that are about to spill across the upstream intersection and to determine whether special action should be taken to clear the queues.

In SCOOT, the data collected by the traffic detectors are stored in the central computer in the form of cyclic flow profiles. These profiles are histograms indicating how the flow rate varied at each detection station within one signal cycle. In order to limit the sensitivity to random fluctuations in the estimated profiles from one cycle to the other, recent traffic counts do not directly replace the previously estimated flow rate in each count interval. Instead, recent counts are merged with previously observed data using moving average principles. As a result, SCOOT flow profiles do not truly represent the actual demand to be placed on an intersection. Rather, they represent the average demand that was placed on each intersection in the previous few cycles.

3.1.3. UTOPIA

The next system to have been developed is UTOPIA, which is often regarded as the first true third-generation system for urban networks. FIAT Research Center, ITALTEL and MIZAR Automazatioze designed this system for the City of Turin in Italy. Its development started in 1976 and resulted in a first implementation in 1984. Since then, a modified version of the system named AUT has also been installed in Gothenburg, Sweden (Burton et al., 1993; Peterson, 1994).

The aims of the system are to provide good traffic signal coordination, improve the flow of vehicles within each coordinated area, give priority to transit vehicles running along selected routes, and permit high flexibility in signal settings at individual intersections. Similar to other systems, the main goal of

the signal optimization is to minimize the passenger car delays. However, this minimization is constrained by the need to provide priority of passage to transit vehicles at controlled intersections. Unlike the previously described systems, which were primarily designed to improve general traffic conditions, UTOPIA was explicitly designed with the objective of providing transit priority.

To achieve the above objectives, UTOPIA places a strong emphasis on decentralization of timing decisions. This system divides the network control problem into a supervisory network coordination problem and a series of individual intersection optimization problems. At the network level, signal timings are generated every six minutes for the next 30 minutes of signal operation based on historical and predicted traffic demands. The goal of this optimization is to obtain smooth traffic flow patterns throughout the controlled area. At the intersection level, individual signal controllers have the ability to completely override the timings provided by the network control level. Each controller seeks its own local optimum based on currently observed traffic conditions, future traffic and signal control information provided by adjacent controllers, and constraints provided by the network control level. At this level, timings for the next 120 seconds of signal operation are updated every six seconds. For the first 30 seconds, the goal of the optimization is to minimize a cost function considering the stops and delays incurred by all types of vehicles, as well as any excess queuing. For the remaining 90 seconds, timings are generated by also taking into account the probability of not providing priority to transit vehicles, as well as any deviation from the timings established six seconds earlier and the network reference plan. The signal control achieved at this level is without any explicit reference to the concepts of cycle time, green allocation and signal offset. Control is made on a phase-by-phase basis, with the main decisions being when to end the current phase and which phase to go to next.

Similar to SCOOT, UTOPIA attempts to predict the demand that will be placed on each intersection in the near future. At the network level, the system predicts the most important routes that will be taken by the general traffic in next 30 to 60 minutes using a two-part traffic assignment model. The first part of the model operates on a day-to-day basis. Its role is to update the attributes of a predefined set of origin and destination points and to identify day-by-day similarities on all predefined routes. The second part of the model operates in real-time. This part counts the vehicles detected at each origin and then makes predictions of traffic volumes on each defined route based on these counts and historical data. At the intersection level, near-future arrivals are predicted by projecting the stop line arrival time of vehicles detected at the upstream end of each approach. In order to increase the horizon over which future arrivals can be predicted UTOPIA also instructs adjacent controllers to exchange pertinent traffic and signal information.

3.1.4. PRODYN

PRODYN is the only other third-generation signal control system for urban network currently in operation. This system was developed in France by the Centre d'Etudes et de Recherche de Toulouse and was first tested around 1988. This system is unique in that it adopts a fully distributed approach to control urban signalized networks. In this case, the local signal controllers do all timing calculations. The role of the central computer is limited to monitor signal operations and to perform data management tasks.

In PRODYN, the main goal of the signal optimization is to minimize a performance function considering the total amount of delay incurred by all types of vehicles. This function is minimized using dynamic programming techniques. Changes in traffic demands are considered by instructing the signal controllers to update every five seconds their timing strategy for the next 75 seconds of signal operation. As in UTOPIA, signal control is also provided on a phase-by-phase basis, with no explicit reference to the notions of cycle time, green allocation and signal offset.

Traffic detection within PRODYN is made using two or three detectors per approach. The first detector is located at the upstream end of each approach and is used to obtain advance information about future vehicle arrivals. The second detector is located 50 metres upstream of the stop line. On long links, a third set can also be installed at 200 metres from the stop line. These detectors are used to correct projected stop line arrival times and queuing estimates. To ensure adequate coordination between adjacent intersections, each controller is also instructed to send forecasts of vehicle departures from the intersection under its control to downstream controllers. For approaches for which there are no nearby adjacent controllers, a moving average estimation technique is used to predict traffic demand in the later part of the decision horizon.

3.2. Transit Priority Features

3.2.1. UTOPIA

In UTOPIA, transit priority is provided on an active basis. Upon detection, transit vehicles traveling along selected routes are given absolute priority of passage at the following signalized intersections. An absolute priority is given in that there is no direct consideration for the potential impacts on other traffic of the proposed signal changes to accommodate incoming transit vehicles. As indicated earlier,

the main control objective of this system is to minimize delay to passenger car traffic, subject to any delay necessary to accommodate transit vehicles. At each intersection along transit routes, the signal timings are generally set to avoid delaying approaching transit vehicles. The only exception to this rule is when there are conflicts between two or more transit vehicles at an intersection.

To monitor transit movements within the controlled network, UTOPIA first relied on the use of special detectors placed at strategic points along transit routes. In Turin, UTOPIA's only implementation site, detectors were installed immediately upstream and downstream of every transit stop, as well as at the entrance and exit of every signalized intersection along transit routes (Donati et al., 1984). Each one of these detectors was able to recognize specific identification codes send by transponders mounted on transit vehicles. This allowed them to selectively detect transit vehicles in mixed-traffic environments. Each time a transit vehicle was detected, UTOPIA's central computer was instructed to predict the arrival time of the detected vehicle at successive downstream intersections over the next eight minutes. The results of this prediction were then passed to the corresponding signal controllers along transit routes and updated every time the vehicle would pass over another detector.

Improvements in this detection scheme were later made when the City of Turin put an automatic transit vehicle location system into operation (Mauro and Di Taranto, 1989). Under the new system, predictions of transit arrival times at signalized intersections are directly provided to signal control system by the city's transit vehicle location system. UTOPIA's signal controllers now only remain in charge of awarding priority to incoming transit vehicles.

3.2.2. SCOOT

Transit priority features in SCOOT were initially limited to passive principles. In earlier versions of the system, transit priority could only be achieved indirectly by biasing the split and offset optimizations to favor links with high transit volumes, or by providing fixed offsets based on observed average transit travel times between successive intersections. Since there was no direct detection of transit vehicles, it was therefore only on average that these vehicles would truly receive priority of passage at controlled intersections. Active priority features have been only added recently (Bretherton, 1996), giving SCOOT the capability to accommodate detected incoming transit vehicles by either extending the current phase or causing a specific phase to occur earlier.

Similar to UTOPIA, SCOOT can detect transit vehicles using either selective transit vehicle detectors or an automatic vehicle location system. In London, England, for instance, selective detection is achieved by mounting transponders on transit vehicles. In Southampton, transit information provided by an independent vehicle location system using dead reckoning to locate transit vehicles in the controlled network. In all cases, while it is generally recommended to detect approaching transit vehicles as far as possible from the intersections where priority is to be considered, detection cannot take place before any transit stop. The reasons for this restriction is that SCOOT does not attempt to predict the amount of time transit vehicles spend loading and unloading passengers at each transit stop (Bowen and Bretherton, 1996). Any detection made upstream of transit stops would therefore not result in accurate prediction of intersection arrival times.

Each time a transit vehicle is detected, the information is sent to the local controller and the central computer. At this point, a decision is made as to whether to extend the current phase or to force a given phase to occur earlier by shortening the current phase and any other intermediate phases. Decisions to call a given phase at an earlier time are always taken by the central computer, while either the local controller or the central computer can award green signal extensions. In the last case, it is generally preferred that the local controllers take the decisions. This eliminates three to four seconds of communication delay between the local controller and the central computer and allows extensions to be granted to vehicles arriving just a few seconds before the scheduled end of a phase. Once an extension or recall has been implemented, the system must pass through a period of recovery to realign the timings with the normal SCOOT optimization.

To ensure that extensions awarded locally will not excessively disrupt area-wide traffic progression patterns, controllers can only award extensions if they have permission to do so from the central computer. This permission is reevaluated every second by the central computer based on the level demand placed on each intersection. On a general basis, the central computer only allows local controllers to grant extensions when the degree of saturation of the intersection under their control is below a threshold value. In other words, local controllers can only award extensions when there is sufficient spare capacity at the intersection.

Active transit priority is generally offered on an unconditional basis only. Changes to the signal timings are made with the sole objective of accommodating detected incoming transit vehicles, without directly evaluating the consequences of these changes on other traffic. This is especially true when local controllers award green interval extensions. As indicated in Section 3.1.2, the central computer

normally performs the signal optimization in SCOOT. Consequently, any extension awarded by the signal controllers can be considered as a temporary override of the timings produces by the central computer. This is especially true if one considers that a recovery period usually follows the extension to realign the timings at the intersection where priority was granted with normal SCOOT timings. As a result, even if the central computer imposes some constraints on the ability of local controllers to award green interval extensions, there is nothing to guarantee that all awarded extensions will not have any significant negative impact on general traffic performance around the intersections under consideration.

3.2.3. SCATS

SCATS also provides priority on both passive and active bases. Passive priority is achieved by defining transit routes as priority itineraries within the network and by allocating a favorable bias to them, permanently, by time-of-day, or by direct operator command. Active priority features allow the system to switch to a particular phase or extend the current green signal indication after transit vehicles have been detected on selected approaches. In the last case, however, the ability to implement signal changes is restricted by the need to preserve the existing cycle time at each intersection. In this case, any time gained by a particular phase through green signal extensions or used to implement a special transit phase must be taken from other phases. It results from this priority scheme that transit vehicles only receive preferential treatment at specific intersections when there is enough spare green time to move around.

To monitor transit activities, SCATS uses two traffic detectors per approach. The first detector is placed at some distance from the intersection to provide advance information about incoming transit vehicles. The second detector is installed at the stop line and is used to hold priority requests until vehicles detected by the upstream detector have entered the intersection. Similar to previously described systems, selective detection technology is used to provide transit detection in a mixed-traffic environment. In Melbourne, for instance, transponders are mounted on trams to allow loop detectors to discriminate them from other traffic (Cornwell et al., 1986).

Similar to UTOPIA and SCOOT, SCATS only provides active priority on an unconditional basis. In this system, signal changes designed to accommodate transit vehicles are implemented without pre-evaluating their effects on general traffic. Priority requests are issued and granted on a simple call basis. The system relies on its adaptive capabilities, specifically its ability to adjust the green allocation

at each intersection response to the demand on each approach, to restore the balance of traffic demand on each approach after the implementation of transit priority measures.

3.2.4. PRODYN

PRODYN does not currently provide priority to transit vehicles. However, according to Wood (1993), the system's developers are considering the addition of such facilities.

3.3. Transit Interference Modeling

3.3.1. UTOPIA

In UTOPIA, different models project the movements of passenger cars and public transit vehicles within the controlled networks. In the first implementation of the system, three different prediction models were used. The first one was a macroscopic simulation model used to predict general traffic movements in the controlled network. The second was a microscopic model performing a detailed simulation of private traffic behavior around signalized intersections. The third model predicted the movements of transit vehicles along transit routes considering dwell times and delays at signalized intersections. Since different models were used to predict the movements of passenger cars and transit vehicles, the interaction between both types of vehicles is not accurately simulated. The accurate modeling of this interaction is now even more difficult since the transit vehicle movements are now predicted in Turin by an automatic vehicle location system operated independently from UTOPIA.

3.3.2. SCOOT

SCOOT directly predicts the movements of both passenger cars and transit vehicles along intersection approaches. However, this system does not attempt to measure or predict the amount of time transit vehicles spend at a transit stop to board and discharge passengers. SCOOT simulates the progression of transit vehicles along urban links assuming these vehicles do not have to stop between any detection point and the intersection stop line. In other words, it is assumed that transit vehicles behave like passenger cars. The only interaction with other traffic that is considered is the delay caused to transit vehicles by other vehicles queued at the intersection stop line. As a result of this modeling choice,

approaching transit vehicles can only be detected downstream of any transit stop and transit interference on other traffic during dwell times cannot be considered.

The installation of traffic detectors immediately downstream of transit stops to monitor flow patterns during transit dwell times cannot significantly improve the system's ability to react to disruptions caused by transit vehicles. As indicated earlier, SCOOT generates signal timings using cyclic flow profiles that are generated by merging recent traffic counts with past-calculated flow averages. Thus, disruptions caused by dwelling transit vehicles will always be averaged out with traffic observations from previous signal cycles. Unless the disruptions occur every cycle or so, their real impact on traffic behavior will always be masked in traffic demand estimates. Furthermore, as a result of the averaging process, any recorded disruption will also later affect projected demand estimates over a certain number of cycles.

3.3.3. SCATS

Like UTOPIA and SCOOT, SCATS does not attempt to model transit interference on other traffic. In this case, however, the reason for the absence of such modeling is found in the way the demand placed on each intersection is estimated. In UTOPIA and SCOOT, traffic demand is estimated from detectors installed at the upstream end of each approach to the signalized intersections. For these systems, traffic prediction must be used to convert vehicle detection into stop line arrivals. In SCATS, however, traffic is directly detected at the stop line. The system uses stop line detectors to monitor the rate at which vehicles are flowing across the intersection from each approach during the various signal phases. As a result, there is no need to project private traffic behavior along intersection approaches between an upstream detection location and the stop line. In this case, only transit vehicles are the objects of upstream detection in systems in which active detection is required.

3.3.4. PRODYN

In PRODYN, only passenger cars are currently modeled. Significant discrepancies may therefore exist between the stop line arrival patterns estimated by the model and the actual arrival patterns if control is performed on streets where transit vehicles and passenger cars share the right of way. This is particularly true where transit stops are located near the intersection and more particularly if they are located downstream of the traffic detector closest to the stop line.

3.3. Summary of Existing Approaches to Coordinated Real-Time Control

In summary, it appears that very different approaches have been proposed to achieve coordinated real-time traffic signal control in urban signalized networks. For instance, SCATS and SCOOT are both designed to maintain traffic signal control at a central computer and to adjust signal timings at individual intersections with the constraint of maintaining a common cycle time at all controlled intersections within a given area. This design was chosen to promote area-wide coordination. Both models are also designed to react to changes in traffic demands by slowly varying the cycle time in operation in a given coordinated area, as well as the green allocation and offset of individual intersections, by up to a few seconds in each signal cycle. However, SCATS was also designed with the added ability of allowing greater temporary changes at individual intersections through the use of traffic-actuation logic.

On the other hand, UTOPIA and PRODYN are designed to respond more quickly to changes in traffic demands. In these systems, the traditional concepts of cycle time, green allocation and signal offset are not used directly. Signal optimization is done locally, in a reactive manner, by directly calculating the best time to switch from one phase to another. Both models place a strong emphasis on decentralization of signal timing decisions. The timing decisions are made at the intersection level by the local signal controllers. In both cases, signal coordination between adjacent intersections is promoted through information exchanges between signal controllers. There is no network-wide signal optimization in PRODYN, while UTOPIA further enhances signal coordination by setting up timing constraints for the control of individual intersections based on area-wide traffic signal optimizations.

The review that has been conducted in the chapter also reveals that none of the existing systems currently considers transit interference on other traffic in networks in which passenger cars and transit vehicles share the right of way. In PRODYN, transit vehicles are completely ignored. In SCATS, SCOOT and UTOPIA, transit vehicles are selectively detected from the mixed traffic stream, but it is assumed that transit activities carried out between the intersections do not affect the progression of general traffic. It is also observed that transit priority features are not currently offered in PRODYN and that the three other systems can only provide priority on a passive or unconditional active basis. As it will be explained in the next chapter, these deficiencies create a need for the development of a new real-time traffic signal control model for coordinated urban signalized networks.

4. Research Objectives

It has long been appreciated that significant benefits in the management of urban signalized networks could be obtained by allowing traffic signals to be responsive to variations in traffic demands. With respect to that objective, real-time, traffic responsive signal control systems for urban networks have proven their ability to reduce the stops and delays incurred by vehicles or vehicle passengers. However, while these systems have demonstrated benefits, their implementation has also resulted in the identification of a number of remaining limitations. A detailed analysis of their control features notably reveals that they are still not capable of providing truly optimal traffic signal control in all types of networks. In particular, system limitations are evident when applying these systems to networks in which passenger cars and transit vehicles share the right of way.

The goal of this chapter is to clearly define the research project described in this thesis. In that goal, the chapter first defines the main limitations of the existing and proposed traffic signal control systems prompting the need to develop new signal control strategies for coordinated urban networks. The chapter then introduces the SPPORT model, which has been explicitly developed for the control of isolated intersections and which is used within this thesis as a basis for the development a new real-time, traffic responsive signal control model for urban networks. The last part of the chapter presents the goals and main criteria that should be followed in developing a new signal control system for urban networks.

4.1. Limitation of Existing Systems for Urban Networks

A detailed review of the existing and proposed real-time traffic signal control systems for urban networks first reveals that none of these systems currently considers transit interference on other traffic on streets on which passenger cars and transit vehicles share the right of way. In some systems, such as PRODYN, transit vehicles are simply ignored. It is assumed that all detected vehicles are passenger cars. In other systems, such as SCATS, SCOOT and UTOPIA, transit vehicles are selectively detected from the mixed traffic stream, but it is assumed that transit activities carried out between signalized intersections do not significantly interfere with the progression of general traffic. These systems

generally hypothesize that transit vehicles either travel on exclusive lanes between signalized intersections or can otherwise entirely pull out of the right of way in front of every transit stop.

In reality, transit vehicles have often no other choice than to stop in the right of way to board and discharge passengers, causing a temporary closure of the curb lane. This disruption is most severe on streets on which streetcars are running in the median lane, such as in downtown Toronto. On such streets, traffic could be completely stopped behind the stopped streetcar for the entire dwell time. As a result, there could be significant discrepancies between the actual arrival patterns at signalized intersections and the simulated patterns upon which the signal timings are based. Such discrepancies could then lead signal control systems to implement timing strategies that do not necessarily meet the needs of the current demands.

Modeling transit interference on other traffic would not be as important if the disruptions caused by dwelling transit vehicles would only marginally affect traffic progression between signalized intersections, or if these disruptions would not occur frequently. Many traffic signal control systems can implement timings that do not perfectly meet the needs of traffic demands once every 20 to 30 signal cycles. This is particularly true for systems having the ability to quickly react to changes in traffic demands. If corrective measures are promptly implemented following a sub-optimal decision, the overall impact of not adequately serving traffic during one or two signal cycles may not significantly affect the general performance of the system over the entire period of 20 or 30 signal cycles.

Unfortunately, transit interference on other traffic is not uncommon. Because of their potential disrupting effect on other traffic, transit activities carried out in the right of way and causing temporary lane blockage should not be considered as unpredictable random events and neglected in the signal optimization process. To do so would go against one of the basic principles of real-time traffic responsive signal control system, which is to automatically respond to changes in traffic demands. In many cases the disruptions caused by transit vehicles can be anticipated and pre-estimated. As a result of this predictable nature, transit activities affecting the progression of traffic should always be considered in signal optimizations.

A second problem is related to how priority of passage is given to transit vehicles at signalized intersections. While transit priority is generally considered to be an efficient way of minimizing the stops and delays incurred by transit riders at signalized intersections, it could also increase the stops and

delays incurred by vehicles traveling on the streets crossing the prioritized route. For instance, severe queues and excessive delays may be caused on some approaches if last-minute changes in signal displays are frequently implemented to accommodate incoming transit vehicles on other approaches. In many cases, these added stops and delays might completely offset the benefits provided to the transit riders. They might also make it more difficult to accommodate transit vehicles arriving later by reducing the possibilities of freely moving the available green time around.

Presently, none of the existing real-time traffic signal control systems for urban networks can be said to properly consider the impacts of the transit priority treatments they implement. SCATS, SCOOT and UTOPIA, the only systems currently providing priority, can only consider passive or unconditional active priority treatments. Passive priority treatments can be advantageously used where transit volumes are fairly high and repetitive, but do not react to the actual presence of transit vehicles. Priority is purely based on the historical behavior of transit vehicles. As a result, it is only on average that transit vehicles benefit from the priority treatments. Active priority strategies are conceptually more efficient in the fact that they only results in signal changes when required. However, none of the existing systems appear to evaluate beforehand the consequences of altering the signal timings to accommodate incoming transit vehicles. Green interval extensions and phase recalls are often implemented with the sole objective of avoiding delays to approaching transit vehicles. A common assumption seems to be that providing priority to transit vehicles automatically yields overall benefits; however, this assumption is not true for all cases.

4.2. The SPPORT Model for Isolated Intersections

In the previous section, two major limitations of existing real-time traffic signal control systems for urban networks were identified. The first limitation is concerned with the modeling of transit interference on streets on which passenger cars and transit vehicles share the right of way. The second limitation is associated with the way priority is given to approaching transit vehicles.

While working on a project aimed at implementing effective real-time transit priority to Toronto's Queen Street Corridor, Yagar, Han and Greenough (1991) found similar problems in control strategies designed to handle traffic at isolated intersections. In their project, they reviewed the real-time traffic signal control models proposed by Miller (1963), Bang (1976), De Groot (1981), Gartner (1982b), Lin et al. (1987), Vincent et al. (1986, 1988), Bell et al. (1989), and Heydecker (1990).

In an attempt to solve the above problems, they develop a new real-time traffic signal control model for isolated intersections entitled SPPORT, which stands for Signal Priority Procedure for Optimization in Real-Time (Yagar, Han and Greenough, 1991, 1992; Han and Yagar, 1991, 1992; Yagar, Han and Wang, 1992; Yagar et Han, 1994). The two main unique characteristics of this model are:

- The use of a discrete-event microscopic traffic simulator that explicitly models transit interference on other traffic during dwell times when predicting future traffic behavior around controlled intersections.
- The use of a heuristic rule-based signal optimization process allowing the generation of several candidate signal timing strategies before choosing for implementation the one yielding the best performance. By instructing the signal control system to consider offering priority to transit vehicles in some strategies and no priority in others, it thus becomes possible to evaluate the impacts of the proposed priority treatments and to decide whether or not preferential treatments should be implemented.

Simulated tests reported by Yagar and Han (1994) indicates that this model may be advantageously used to control traffic in real-time at isolated intersections where passenger cars and transit vehicles share the right of way. For example, a first test in which SPPORT was instructed not to provide priority to detected transit vehicles resulted in a two percent decrease in the total amount of delay incurred by all vehicles' passengers when compared to the use of optimum fixed timings. When transit priority features were activated, the total decrease in person delay reached five percent. In another test, the installation of traffic detectors just downstream of the stop line to indicate when transit vehicles enter the intersection and no longer require priority treatment produced a ten-percent delay reduction. Alternatively, the total reduction reached 14 percent when the signal control system was informed of the exact time transit vehicles would have finished loading passengers. Other tests performed with an updated version of the model and reported by Conrad et al. (1998) indicate reductions in a performance index combining stops and delays ranging between 8 and 30 percent when the SPPORT model is compared to an optimized fixed-time traffic signal operation.

4.3. Research Goals

The main objective of the research project described in this thesis is to develop efficient real-time, traffic responsive signal control strategies for urban networks solving the current limitations of the

existing signal control systems with respect to their ability to consider transit-related issues in mixed-traffic environments.

To achieve this goal, it is proposed to extend to the network level the applicability of the SPPORT model for isolated intersections initially developed by Yagar et al. (1991). As indicated in the previous section, the model already implements solutions to the two main limitations criticized in existing systems for urban networks. In addition, the model has already proven in simulation studies its ability to efficiently control traffic in real-time around isolated intersections.

While it is technically possible to control groups of adjacent intersections using traffic signal optimization software designed for isolated intersections, such practice would not necessarily lead to optimal traffic control on a network basis. Under isolated control, each signal controller makes timing decisions without directly considering the impacts of its decisions on adjacent intersections. In such case, a minimum amount of coordination would still be achieved due to the fact that observed arrival patterns at a given intersection usually reflect the control strategies implemented at upstream intersections. Coordination would mainly be achieved through the natural tendency of traffic responsive signal control systems to display a green signal indication when platoons formed at upstream intersection are expected to reach the controlled intersection. However, without an explicit set of coordination rules, there is nothing to prevent signal controllers from implementing signal timings contrary to what the upstream and downstream controllers were expecting when they generated their own timing strategy. Experiences conducted by Barrière et al. (1986) with the PRODYN model notably showed that the stops and delays incurred by vehicles in a network could be significantly reduced when traffic responsive signal controllers are forced share traffic information and partly coordinate their operation with their neighbors. In this experiment, the delays incurred by vehicles were increased by 5.0 percent on a four-intersection arterial when isolated control was used, and reduced by 30.7 percent with coordinated control.

In more precise terms, the main goal of the research project is to develop traffic signal control strategies that would allow the SPPORT model to coordinate the operation of groups of adjacent intersections. In order to achieve this goal, the four following prescriptions defined by Gartner (1982) and based on the failure of the UTCS experiment conducted in Washington D.C. by the U.S. Federal Highway Administration in the 1970's (MacGowan and Fullerton, 1979-1980) should be followed:

- The system must be designed to provide better performance than off-line signal optimization methods. Although this objective may seem self-evident, it has not always been recognized

in the past. In some cases, this objective was superseded by less relevant criteria, such as providing good platoon progression along major arterials or variable cycle times.

- The system must implement signal control methods suited to the objective of quickly responding to traffic demand variations. As demonstrated in the UTCS experiment, effective traffic responsiveness is not necessarily achieved by implementing off-line methods at an increased frequency. Signal optimization methods capable of generating efficient timing plans in a few seconds are required, as well as a method to ensure that adequate coordination is maintained between intersections.
- The system must be truly demand-responsive, i.e., must adapt to actual traffic conditions and not to historical or predicted values that may be far off from the reality. As indicated by Gartner et al. (1995), deficiencies in traffic prediction algorithms might also be responsible for some of the failure of the UTCS experiment. This implies detecting traffic upstream of signalized intersection and predicting with reasonable accuracy the time at which vehicles are expected to reach each intersection, especially on streets on which transit activities carried out in the right of way may significantly affect general traffic progression.
- The system should not be arbitrarily restricted to control periods of a specified duration but should be capable of updating plans at any time, at any location. This again calls for the use of efficient signal optimization methods capable of generating new signal timing strategies in a few seconds. It also calls for the ability to freely adjust signal timings at individual intersections within a reasonable set of constraints designed to maintain adequate network signal coordination. This might even involves departing from the traditional notions of cycle time, green allocation and signal offset.

In conclusion, it must also be pointed out that the main focus of the research project is to look at the traffic signal optimization problem from the point of view of vehicular traffic only. Pedestrian issues are not considered for design purposes only. While pedestrians are in many urban areas a major traffic component, their consideration in the signal optimization increases the complexity of the problem to solve, often because of the fact that pedestrian requirements are different than vehicular traffic requirements. For instance, consider the case of a minor street crossing a major arterial. In this case, due to expected traffic distributions, the major street usually requires a long minimum green time, while the minor street requires a much shorter one. Pedestrians, however, cross the major street, which is usually wider, during the minor street vehicular phase, and the minor street during the major street vehicular phase. Consequently, pedestrian requirements are in this case for a longer green during the minor street phase and for a shorter green during the major street phase, which is the opposite of the

requirements for the vehicular traffic. According to McShane et al. (1998), the minimum green times for vehicular traffic rarely, if ever, safely accommodate pedestrian crossing times. As a result, most traffic responsive systems must include a pedestrian push button and an actuated pedestrian phase, which both add unpredictable elements in the signal control system.

In the thesis, the various safety aspects of the traffic signal control problem are also implicitly considered. It is assumed that all safety concerns regarding the flows of vehicles and pedestrians are appropriately addressed through the implementation of signal cycle structures, phase compositions, minimum green intervals, amber intervals and all-red periods that all meet established guidelines such as those described in the Canadian Capacity Guide for Signalized Intersections (ITE, 1995) or the Highway Capacity Manual (TRB, 1994).

5. Signal Control Framework

A general problem faced when designing a traffic responsive signal control system is to develop a system possessing the necessary flexibility to react to the expected variations in traffic demands. In many cases, however, adding flexibility to the control of individual intersections with the objective of better reacting to changes in local traffic demands may contradict the general objective of providing a high degree of progression through sets of coordinated intersections.

At the network or strategic level it is desirable to set the signal timings so that platoons of vehicles released at one intersection will have a high probability of going across the next few intersections without being stopped. However, at the local intersection or tactical level, the main priority is usually to provide signal timings that will explicitly minimize the stops and delays incurred by vehicles around each intersection. As a result, planning traffic movements on an area basis often involves constraining the operation of individual intersections so that they follow the strategic plan, while providing optimum local control often implies destroying existing progression schemes in order to implement short-term tactical decisions. Consequently, one of the major problems faced in the current research project is to design a network signal control system that provides a reasonable balance between local and network signal control strategies in the presence of highly variable demands in which there is still some degree of predictability.

This chapter describes the basic conceptual elements of the proposed SPPORT model for coordinated networks that have been developed to address this problem. It successively presents:

- the model's traffic monitoring system,
- the model's control architecture,
- the method by which the operation of adjacent intersections can be coordinated within the chosen architecture,
- the real-time process by which signal timings are regularly adjusted in response to prevailing traffic conditions, and
- the basic principles of the model's signal optimization process.

5.1. Traffic Monitoring Scheme

Real-time traffic signal control is based on the ability of traffic signal control systems to monitor traffic conditions around controlled intersections and to automatically respond to detected changes in traffic demands. Consequently, the placement, information collected by, and reliability of traffic detectors have a significant impact on the efficiency of the traffic signal operation in all real-time, traffic responsive systems.

This section presents the basic principles of the traffic detection scheme to be used by the proposed SPPORT model to monitor traffic conditions around each intersection. In particular, the need to selectively detect transit vehicles in mixed-traffic environments, the need to estimate vehicle occupancy to estimate person-based performance measures, and the influence of the location of traffic detectors on the signal control strategy are examined.

5.1.1. Selective Detection Requirements

As indicated in the previous chapter, the project described in this thesis is concerned with the development of a real-time, traffic responsive signal control system capable of providing priority to transit vehicles at signalized intersections and to consider transit interference on other traffic. The main effect of this objective is to require that the SPPORT model differentiates transit vehicles from other vehicles when detecting traffic at given locations.

In the proposed model for urban network, it is assumed that traffic detectors are all capable of selectively detecting transit vehicles in an environment in which passenger cars and transit vehicles share the right of way. Each time a vehicle passes over a detector, it is assumed that the detector is able to record to the detection time and determine whether the vehicle is a passenger car or a transit vehicle. For simplicity, it is also currently assumed that the selective detection of other types of vehicles, such as emergency vehicles or trucks, is not required.

The above assumptions are realistic, as any one of the selective detection technologies described in Section 2.3.4 could be used to provide the required information. For design purposes, it is assumed that the selective detection of transit vehicles in mixed-traffic environments will be achieved using inductive loops coupled with either active or passive transponders mounted on transit vehicles. This

choice is largely based on the fact that transponders are currently the most widely used selective detection technology. For example, transponders are used in networks controlled by UTOPIA, SCATS and SCOOT to allow the signal control system to discriminate between transit vehicles and passenger cars. In addition, while the use of transponders is a relatively expensive alternative, this technology still provides the greatest flexibility and highest efficiency according to a study conducted by the Municipality of Metropolitan Toronto (see Table 2.4 in Chapter 2).

5.1.2. Estimation of Vehicle Occupancy

In order to estimate performance measures on a person basis, knowledge of vehicle occupancies is required. As it will be seen later, such knowledge is a critical element of the control philosophy of the proposed SPPORT model for coordinated signalized networks.

In the real world, transit vehicle and passenger car occupancies may significantly vary over time. These variations are not only the consequence of major traffic demand variations, but are also a reflection of different person demand patterns related to various trip purposes and conditions. For example, average vehicle occupancy may be higher in traffic going to a sport stadium before the beginning of a sport game than it is during the off-peak period in the middle of the day.

For design purposes, it is currently assumed in the SPPORT model that all passenger cars and transit vehicles have a constant occupancy rate. This allows the model to automatically assign an occupancy rate to every detected vehicle using a simple look-up table. For passenger cars, an average occupancy rate of 1.5 person per vehicles is currently assumed, while a rate of 60 persons per vehicle is assumed for all transit vehicles.

To consider changes in vehicle occupancy, different occupancy tables reflecting the average occupancy of vehicles at various moments during a typical day of operation could be defined and introduced in the SPPORT model for use at the appropriate times. The SPPORT model is also currently designed to consider occupancy information provided by an external source. In the model, each vehicle that is introduced in the simulation model used to perform the traffic projections and to estimate the performance measures of candidate signal timing plans is characterized by its own occupancy. Introducing real-time occupancy estimates would then be a question of building the proper interface between the vehicle detection system being used and the SPPORT model. If such interface were built,

it would then be possible for the model to evaluate traffic demands and candidate signal control strategies using real-world occupancy estimates.

5.1.3. Traffic Detection Scheme

The placement of traffic detectors around each intersection is of critical importance to the successful implementation of real-time traffic signal control system. To be of maximum use, detectors must be located so that they can reliably detect incoming traffic on each approach. In addition, they must be located as far upstream from the intersection as possible in order to give enough time to the signal control system to analyze the new traffic condition, generate new signal timing plans, and implement any required signal change.

Based on the above considerations, locating traffic detectors at the exit of the upstream intersection appears to be an ideal solution. Such a location provides the soonest information about future stop line vehicle arrivals without the need to predict the effect of the upstream intersection's signal timings on the approaching flow patterns. This location also has the benefit of allowing the detection of queues that are just about to spill across the upstream intersection and block the cross-street traffic.

However, it must also be noted that the reliability of correctly predicting the time at which vehicles will arrive at the stop line decreases as the distance between the detector and the stop line increases. To illustrate this point, consider an approach on which traffic is detected 500 metres upstream of the stop line. If vehicles travel at an average speed of 50 km/h, there will be a 36-second lapse of time between the moment a vehicle is detected and the moment the vehicle reaches the intersection stop line. During these 36 seconds, many events, often unpredictable, may occur. Individual drivers' behavior might also differ from the general assumptions made in the traffic projection process. As a result, predicted stop line arrivals will likely not correspond to the actual arrival times. If the discrepancies are significant, the signal control system may generate and implement signal timings that do not match the needs of the current demand.

To resolve this accuracy problem, the SPPORT model has been designed to consider traffic information from more than one detection station per approach (Yagar and Han, 1994). In the example of Figure 5.1, two detection stations are implemented on each approach:

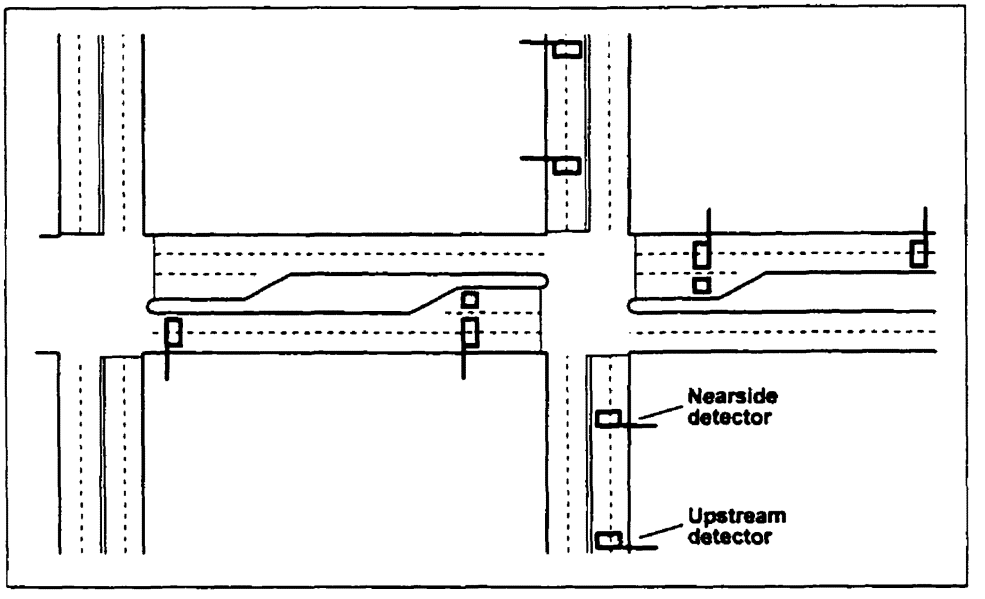


Figure 5.1 - Sample Arrangement of Vehicle Detectors on Intersection Approaches

- An upstream detection station, which is located as far as possible from the stop line. In networks in which the intersections are sufficiently close to justify signal coordination, the detectors are installed just downstream of the upstream intersection. According to McShane et al. (1998), a common practice is to coordinate traffic signals that are less than 800 metres apart on major streets and highways. Where the intersections are too far apart for signal coordination purposes, the location of the upstream detectors is a function of the need for advance traffic information and the cost of the communication system.
- A nearside detection station, which is typically located at about 150 metres upstream of the stop line or where traffic would split into specific through and left-turning streams on approaches on which a left turn bay is present.

The basic principle behind this multi-detector design is to allow the projected stop line arrival time of individual vehicles to be corrected as these vehicles approach the intersection. To illustrate, consider that the upstream and nearside detectors are installed 500 and 150 metres respectively from the stop line, and that the average travel speed on the approach to the intersection shown in Figure 5.1 is 50 km/h per hour. A vehicle leaving the upstream intersection at the left-hand side of the figure will be first detected while it is still 36 seconds away from the intersection. A new stop line arrival time will later be estimated when the vehicle reaches the nearside detector, i.e., when it will be about 10 seconds away from the intersection.

The use of two detector stations per approach produces a sequence of specific future vehicle arrivals in which the expected arrivals within the next 10 seconds is a more accurate traffic prediction than the expected arrivals between the next 10 and 36 seconds. While neither the short-term nor the long-term predictions are completely accurate, they provide sufficient accuracy in the short-term to determine if a signal change is currently required and sufficient accuracy of future demands to permit some valuable longer-term planning.

In all cases, the use of a minimum of two detectors per approach is recommended. While the use of additional detectors might increase the overall accuracy of the traffic detection scheme by providing additional points at which vehicle arrival times are known with exactitude, such use would also increase the total installation and operating cost of the system. Consequently, the decision to install more, or less, detectors should be based on an analysis of all the potential benefits and costs. For instance, the cost of installing additional detectors immediately downstream of a transit stop might be completely recuperated in stops and delay savings for transit riders through the added ability of monitoring the exact time at which transit vehicles finish boarding and discharging passengers and the resulting improved stop line arrival predictions. Like in SCATS, stop line detectors could be also be used to monitor the rate at which vehicles are entering the intersection and help determining whether a given indication should be extended or ended. In all cases, the exact location and number of traffic detectors to use on each intersection approach will depend on the particularity of each site. As a rule, detectors should be placed immediately downstream of the upstream intersection and between 50 to 150 metres from the stop line, or where the approaching flow divides itself into through and turning movements.

5.2. Control Architecture

To be applicable to real-time traffic control, solutions to network signal optimization problems must be obtained within a short time span. Short optimization times are required to ensure that signal control systems truly react to actual traffic demands and not to traffic demands that have already passed through the intersection. For instance, to provide effective transit priority on an active basis it is required that the time needed to generate new signal timings and implement any required changes be shorter than the time transit vehicles normally take to travel from the most upstream detection point to the intersection stop line.

The major problem faced by real-time traffic signal control systems in fulfilling this requirement is that the time required by a central computer system to simulate traffic and optimize the signal timings in real-time for typical large-scale urban signalized networks is still too long for real-time applications. First, some communication time is required in a centralized system to send to the central computer the traffic data collected by the traffic detectors, and to return to the signal controllers the signal timing decisions taken by the central computer. For instance, a communication delay of three to four seconds is required in the SCOOT system installed in London, England, to exchange data between the local signal controllers and the central computer (Bretherton, 1996). Second, the CPU time required by a single computer for optimizing a network of signalized intersections increases with the number of intersections to consider in the network. This increase is a function of the need to consider more complex optimization problems with every increase in the number of controlled intersections. Finally, the complexity of the optimization problem is also a function of the variability of the traffic demand to control.

For example, under constant demands, the repetitive behavior of traffic patterns at individual intersections might help in planning signal control strategies at adjacent intersections. However, under variable demands, there might not be any repetitive patterns to help in the signal optimization. As a result, last-minute changes at a given intersection to respond to unexpected changes in traffic conditions might significantly change traffic patterns at downstream intersections and cause the controllers of these intersections to also change their signal control strategy. Such changes in signal strategy are then reflected back to the intersection where a first detected change in traffic conditions triggered the re-optimization of signal timings, causing the control loop to start again. As a result, the number of timing combinations that need to be considered to find an optimal solution might be too large in many cases for a single computer to evaluate all of them within a reasonable time.

In an attempt to reduce the complexity of network signal optimization problems and allow solutions to these problems to be obtained within a short time span, real-time traffic signal control systems often decompose the initial control problem into a series of sub-problems of manageable complexity and size. Over the years, two main decomposition approaches have been proposed. The first is the hierarchical approach and the second is the fully distributed approach.

In hierarchical approaches, the network optimization problem is transformed into a multi-level control problem with distinct optimization objectives at each level. The most common hierarchical approach

consists of a two-level control structure in which network objectives are considered at the upper level and local objectives at the lower level. In such a structure, a central computer typically produces a first estimate of the optimum timing parameters for each intersection in the network under control by projecting area-wide traffic conditions over the next several minutes. These initial timings are then passed down from the upper level to the lower level where the corresponding signal controllers adjusts the initial timings to suit the traffic conditions prevailing around each intersection.

Depending on which level performs the main optimization, hierarchical systems are said to be either mainly centralized or mainly distributed (see Figure 5.2). In centralized systems, the signal optimization is essentially performed at the network level. Adjustments are still made by local controllers, but these adjustments are usually constrained to preserve the signal coordination patterns established by the central computer. SCATS is an example of such a centralized hierarchical system. In decentralized systems, the controllers perform the main signal optimization. In this case, the central computer only plays an advisory role. It produces signal constraints for individual intersections reflecting network coordination objectives and proposes candidate timings to individual controllers. However, the controllers usually make the final timing decisions. UTOPIA is an example of such a decentralized hierarchical system.

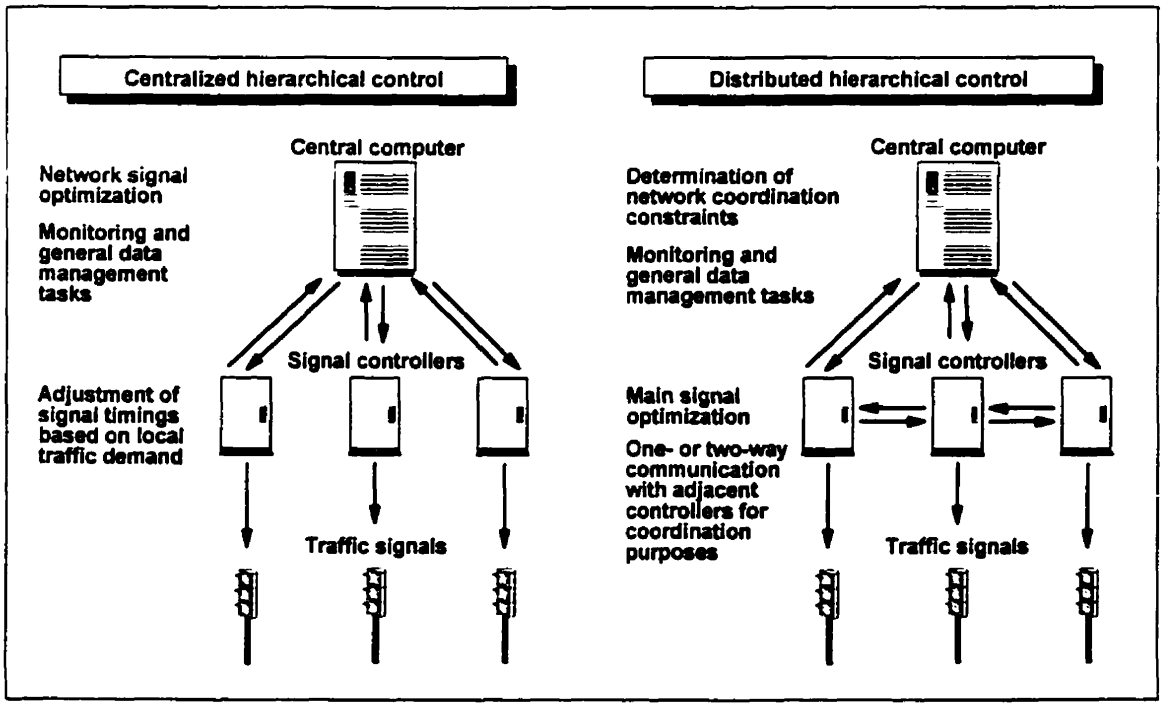


Figure 5.2 - Hierarchical Traffic Signal Control Approaches

In the fully distributed approach, the network control problem is simply broken up into a series of smaller problems that can be solved in parallel by independent processors (see Figure 5.3). There is no upper or lower control level. There is also no explicit global network optimization. Coordination between adjacent sub-system is achieved by strategically modeling the linkage that exists between each sub-system and its immediate neighbors. For example, in the PRODYN system each intersection is treated as an independent optimization problem. This allows the signal timing calculations to be entirely distributed among the local controllers and minimizes the total time required to perform a network optimization by allowing all intersections to be simultaneously optimized in parallel. It also makes the optimization problem independent of the size of the network.

On the basis of the characteristics of the various control approaches, it appears advantageous to follow the hierarchical decomposition approach as this decomposition divides the network control problem according to specific optimization objectives. Specifically, this approach permits long-term proactive control to be considered at the network level, and short-term reactive control to be considered at the intersection level. However, developing an efficient hierarchical signal control structure also necessitates finding ways to resolve the conflicts that often appear between the two control levels. For example, often the adjustments proposed by the signal controllers to better meet the needs of the local

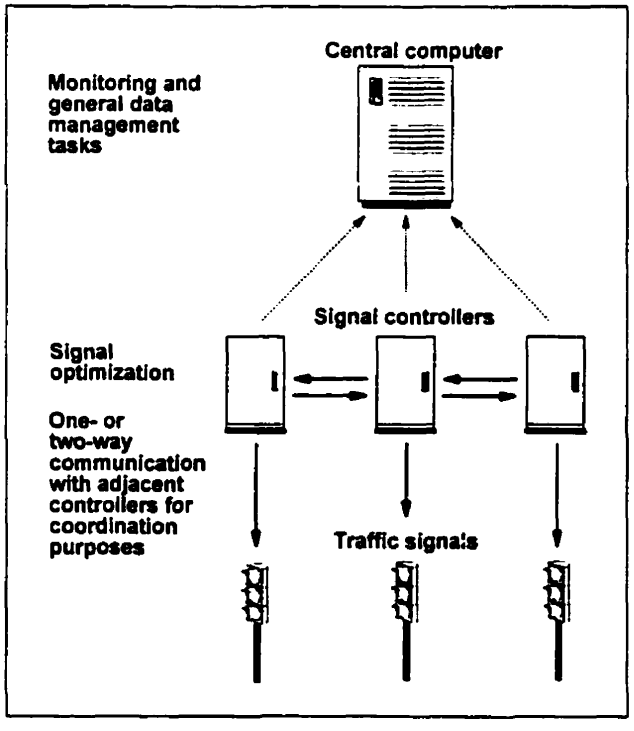


Figure 5.3 - Fully Distributed Traffic Signal Control Approach

demand conflict with the area-wide progression scheme recommended by the network control level. In such cases, a decision must be made as to which timing scheme to adopt and which to reject. To-date, no satisfactory solution to the problem of conflicting local and network recommendations has been found. Signal controllers are either allowed to completely override the proposed network timings, as in UTOPIA, or subjected to constraints determined by the network control level and based on predicted average flows that may not necessarily correspond to the prevailing demand.

One major element of consideration in the choice of an appropriate real-time traffic signal control structure is the variability of the demand to control. As indicated earlier, planning and maintaining efficient area-wide progression patterns is very difficult in networks in which there are large and frequent variations in traffic demands and/or saturation flows. A noted example is on Toronto's Queen Street corridor (Jacques, 1993). In this two-arterial corridor encompassing 13 intersections, streetcars loading in the right of way can completely block traffic near each intersection for 10 to 30 seconds every three minutes. By the time a vehicle would reach the last intersection along the controlled corridor, local traffic conditions at that intersection could be very different from those initially predicted. Because of the severity and frequency of the traffic disruptions, attempting to plan either one- or two-way progression on any of the arterials of the corridor is almost futile. Providing priority to transit vehicles would likely make the situation even worse by allowing sudden signal changes to take place to accommodate approaching transit vehicles. Given the difficulty of generating efficient area-wide timing plans, a hierarchical control structure does not appear to be suitable for this corridor. Achieving control on an intersection-to-intersection basis (i.e., a fully distributed approach), likely would be a more efficient control approach.

The literature does not provide a clear indication as to which control approach provides better performance in the field. However, several studies have been conducted in which different control strategies have been compared under laboratory conditions. Using simulation, Barrière et al. (1986) compared hierarchical and distributed models of the PRODYN real-time traffic signal control system and concluded that a single-level, fully distributed model would perform almost as well as a two-level, hierarchical model, but with significantly fewer calculations involved. These conclusions resulted in the rejection of the hierarchical approach that had been at the center of the PRODYN's research efforts up to that time (Farges and Henry, 1988).

While studying the operation of the PRODYN model in periods of congestion, Kessaci (1988) later found that improvements could be made in the operation of the model by defining a hierarchical control structure that would provide constraints to the individual signal controllers based on network control objectives. The problem he found with PRODYN is the inability of the model's fully decentralized architecture to consider downstream queue limitations in urban networks (Kessaci et al., 1989). To solve this problem, he suggested the implementation of an upper control level that would determine the minimum and maximum duration of candidate phases with the objective of ensuring that traffic fluidity is maintained at the network level without overly constraining the operation of the model, but only in periods of congestion.

While the work of Kessaci seem to indicate that a hierarchical control structure appears to be the best solution for the PRODYN model, it is not necessarily the case for the SPPORT model. In the case of PRODYN, no consideration was made of traffic conditions downstream of each individual intersection. Signal coordination was only attempted with upstream intersections. In SPPORT, the rule-based optimization logic used by the model allows the introduction of signal-switching rules considering downstream traffic conditions. If such rules are introduced in the signal optimization process, it may then not be necessary to define a hierarchical control structure to provide efficient real-time, traffic responsive signal control in periods of congestion or any other control period.

To test this assumption, the current research will focus on the development of a version of the SPPORT model that could operate in fully distributed control architecture. This would provide the model with the highest possible degree of flexibility in the operation of traffic signals at individual intersection, and therefore, the best response level to variable demands. However, there is currently no definitive evaluation on whether a fully distributed approach is better, or worse, than a hierarchical structure. If it is later found that a hierarchical structure would provide a better control environment, the SPPORT model could be converted for use in such structure. In case of such conversion, the current research effort would then be viewed as the step in the general development of the model during which the control logic used by local controllers at the lower lever of the hierarchy would have been developed.

5.3. Network Control Principles

By dividing network control into a series of independent intersection optimization problems that can be solved in parallel by individual signal controllers, the proposed SPPORT model for urban networks

loses the ability to proactively plan and promote traffic movements across large groups of intersections. In this case, coordination only remains possible between adjacent intersections, if the linkage between these intersections is properly modeled in each signal optimization problem. As it was indicated in Section 5.1.3, a common practice is to attempt to coordinate within a network signalized intersections that are less than 800 meters apart (McShane et al., 1998). It was also observed that platoons of vehicles released from a signalized intersection often maintain their grouping for well over 300 meters past the intersection (TRB, 1985).

Regardless of the type of control architecture being used, a certain degree of coordination can always be achieved with upstream intersections since the observed arrival patterns at a given intersection usually strongly reflect the timing strategies implemented at upstream intersections. Any signal optimization algorithm attempting to respond to projected arrival patterns will implicitly try to coordinate its operation with upstream intersections.

In the proposed SPPORT model for urban networks, signal coordination between adjacent intersections is also promoted by allowing signal controllers to evaluate the effects of candidate signal timing strategies not only on links leading to the intersection being optimized but also on links leading to neighboring downstream intersections. This is achieved by providing each signal controller with a modeling of dependent links and surrounding intersections similar to the mini-network shown in the left-hand side of Figure 5.4.

The mini-network of Figure 5.4 allows controllers to simulate the effect of each candidate signal timing strategy on both upstream and downstream traffic conditions. This allows them to discriminate timing plans that might appear ideal for incoming traffic but that could produce poor offsets with downstream adjacent intersections. In many cases, poor offsets with downstream intersections may result in vehicle stops and delays on exit links from an intersection that may exceed the benefits of arranging favorable progression with upstream intersections. In the SPPORT models for isolated intersections, only the links leading to the intersection being optimized were modeled (links shown with a solid line in Figure 5.4). Candidate timings were generated and evaluated only on the basis of their effect on upstream traffic conditions. This structure limited the evaluation of candidate timings, often leading to the selection of sub-optimal solutions. For example, the impact of a queue spilling back across the controlled intersection on the choice of the optimal timing could not be considered.

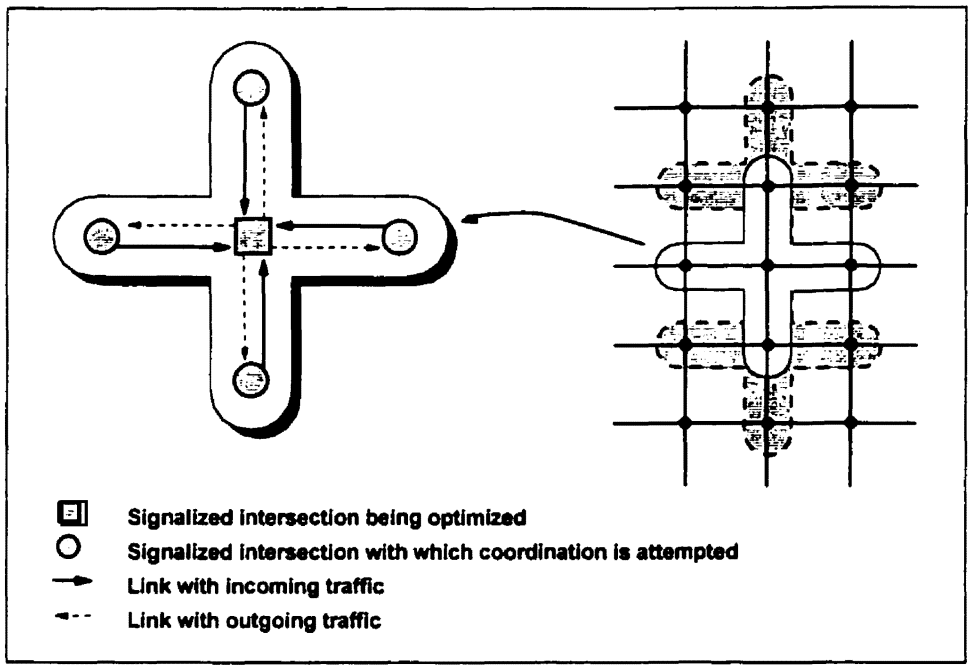


Figure 5.4 - Modeling of Dependent Links and Coordinated Intersections around Individual Intersections

The mini-network of Figure 5.4 can also be viewed as an elementary block for the coordinated control of large urban networks. For instance, if each intersection in the grid network illustrated in the right-hand side of the figure is optimized in turn, then each intersection would successively be the subject of the optimization and an intersection with which signal coordination is attempted. Due to the overlapping of control areas, any timing decision taken at one intersection will gradually propagate its effects across the entire network. In this way, the entire network becomes interconnected and coordinated even if all intersections are individually optimized. Global optimization cannot be guaranteed, but if all intersections are operated with timings that minimize stops and delays within each mini-network, it is hypothesized that the entire network would operate close to the global optimal conditions.

Similar to what is currently done in PRODYN and UTOPIA, signal coordination in the SPPORT model for urban networks is further enhanced by instructing each signal controller to provide pertinent traffic information to its immediate neighbors. In SPPORT, all signal controllers are instructed to send their latest projected timings to both upstream and downstream neighboring controllers, and to send a list of projected vehicle departure times from the intersection under their control to downstream controllers. As illustrated in Figure 5.5, individual signal controllers on one-way streets will receive the projected

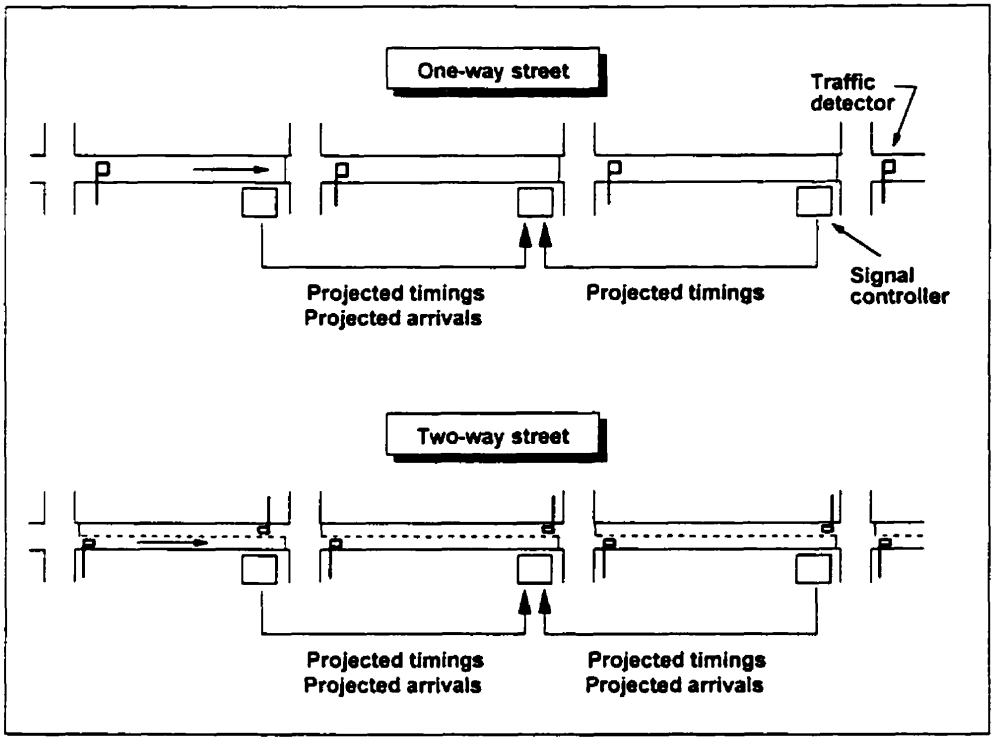


Figure 5.5 - Information Exchange Between Adjacent Traffic Signal Controllers

timings of both upstream and downstream intersections, as well as a list of projected arrivals from upstream intersections. On two-way streets, the projected arrivals are obtained from both upstream and downstream controllers.

All information transmitted from one controller to the next is used to increase the horizon over which future arrivals can be projected at each intersection. As illustrated in the diagram of Figure 5.6, traffic detectors installed at the upstream end of intersection approaches can only provide estimates of future vehicle arrivals at the downstream intersection over a limited time horizon. When projected departures from upstream intersections are made available, it then becomes possible to project traffic conditions over a much longer period. This information also increases the ability of signal controllers to pre-plan their operation by enabling them to evaluate the consequences of their signal-switching decisions over longer periods using projected traffic arrivals that reflect implemented and proposed signal timing strategies at the upstream intersections.

However, by instructing signal controllers to consider information received from neighboring intersections, there is a risk that some nervousness may be introduced in the signal control system. The problem lies in the fact that the information each controller receives reflects the optimum traffic signal

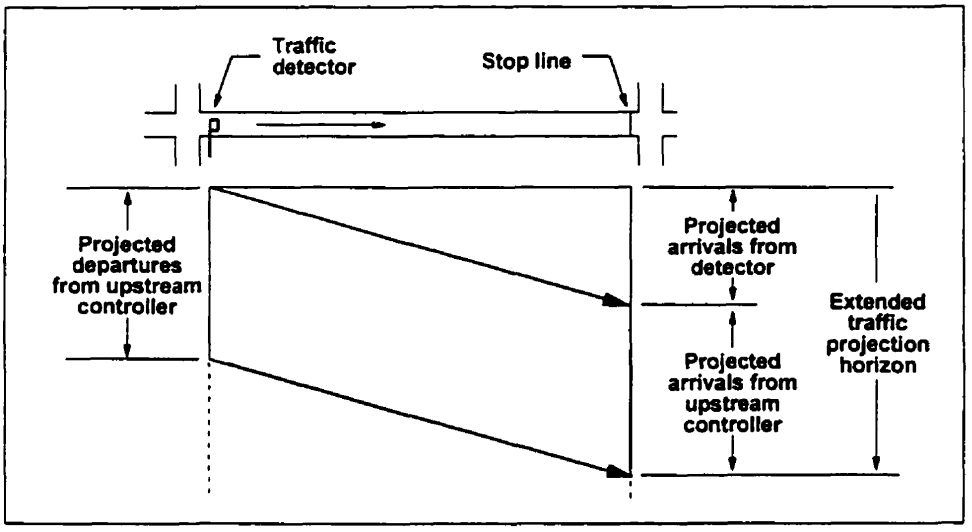


Figure 5.6 - Extended Traffic Projection Horizon in Signal Optimization Process as a Result of Data Exchange between Adjacent Traffic Signal Controllers

control strategies that neighboring controllers were considering to implement at the end of the previous optimization step. By the time a controller finishes generating a new signal timing plan based on the latest information received, the projected timing plans at adjacent intersections might have also changed. A feedback effect may then push the signal controllers to constantly oscillate between two sets of signal timings. While such effect might not be avoided, as it will be seen in Chapter 9, its overall effects on vehicular traffic should be constrained by the ability of the SPPORT model to automatically adjust its signal timing strategy to newly detected traffic conditions.

The process by which departures from individual signalized intersections are estimated and transmitted to downstream adjacent intersections will be explained in more detail in Chapter 7.

5.4. Real-Time Signal Operation

As with most traffic responsive signal control systems, the proposed SPPORT model for coordinated urban networks relies heavily on projected traffic information to generate switching decisions. As with all projections, the accuracy of the projection is not guaranteed. Unpredictable events such as drivers accessing or exiting mid-block streets and drive ramps may cause changes in traffic patterns. Furthermore, the mean value (and the distribution of individual values about this mean) of many

important simulation parameters are very difficult to estimate precisely. Among these are the chosen paths and the preferred traveling speeds of individual drivers.

Based on the sources of error likely to influence projected vehicle arrival times, it appears essential to provide the SPPORT model with means of correcting itself as traffic conditions change. This ability has been implemented in the form of a discrete time, rolling horizon decision-renewal process. This process, which is illustrated in Figure 5.7, operates as follows:

- At time t, the decision horizon is divided into a number of decision intervals having a duration of specified by the user;
- Signal timings are generated over the entire duration of the decision horizon using available traffic information;
- Within the decision horizon, signal switches can only occur at fixed intervals;
- At the end of the optimization, only the first few seconds of the newly generated timings are committed to field implementation;
- At the beginning of the next optimization stage, at time t+n, the start of the decision horizon is rolled over by n seconds to the beginning of the next decision interval and the optimization process starts again.

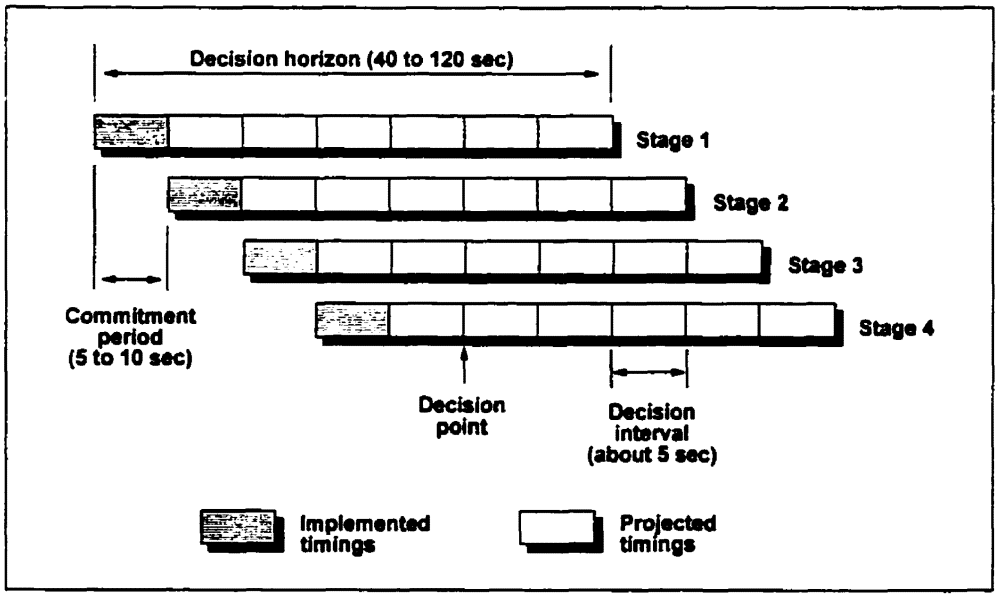


Figure 5.7 - Discrete Time, Rolling Horizon Decision-Renewal Process

The use of a discrete time, rolling horizon approach for real-time traffic signal control purposes was initially suggested by Robertson and Bretherton (1974). Gartner (1983) later further developed this approach for use in the OPAC real-time traffic signal control model. The first real-time traffic signal control system for urban networks using the discrete time rolling horizon approach to be fully developed and implemented is the UTOPIA system in Turin, Italy. This system provides real-time traffic signal control at the intersection level using a 120-second decision horizon, three-second decision intervals, and a six-second commitment period. The PRODYN system, which was developed after UTOPIA, also uses a rolling horizon approach. In this case, however, the decision horizon is 75-second long while the decision intervals and the commitment period are both five-seconds long.

In SPPORT, the duration of the decision horizon, decision intervals and commitment period are all user-defined. Any reasonable value can be used in the signal optimization process, provided that a few simple rules are respected. First, the duration of the decision horizon should not exceed the period over which future traffic conditions can be projected. If this rule is violated, the timings proposed in the later part of the decision horizon may not reflect the needs of traffic demands. In addition, the duration of the decision horizon should be long enough to allow optimal signal-switching decisions to be made in the commitment period. The duration of the commitment period should also be greater than the time required to perform a signal optimization and implement any required change in the field. Currently the values used for the decision horizon, decision intervals and commitment periods in SPPORT are 60, five and five seconds respectively (the impact of these selections will be examined more fully in Chapter 9).

The main advantage of using a rolling horizon approach is that it allows a constant renewal of signal timings based on latest traffic information. By regenerating the signal timings every few seconds, SPPORT signal controllers have the ability to quickly adjust the timings to unpredicted changes in traffic demand and to correct any mistakes that might have previously been made based on incomplete or inaccurate traffic information. Even though the signal timing decisions made in the later part of the decision horizon are not implemented, they are not wasted. As optimum projected timing plans are passed to adjacent controllers, these timings allow signal controllers to better plan future timing strategies, and consequently, to better coordinate their operation with surrounding intersections. The timing decisions taken in the later part of the decision horizon are also used to simulate future traffic conditions over the whole duration of the decision horizon. This allows the compilation of stop and delay estimates that are useful in selecting the optimal candidate control strategy.

5.5. Signal Optimization Principles

This section provides general insights on the principles that are followed by SPPORT to generate efficient real-time traffic signal control strategies for urban networks with mixed-traffic. Four major elements related to the signal optimization process are presented: the application of acyclic signal control principles, the objective function against which candidate timings plans will be evaluated, the ruled-based signal-switching decision process and the multi-objective optimization process.

5.5.1. Acyclic Network Control

Two types of signal operation can be used to control traffic in urban signalized networks: cyclic operation and acyclic operation. Cyclic operation relies on the traditional concepts of cycle time, green allocation and signal offset defined in section 2.2.1 and illustrated in Figure 2.3. This type of operation is based on the requirement that all signalized intersections within pre-defined coordinated groups operate under the same cycle time. As illustrated in Figure 5.8, this requirement is imposed on the

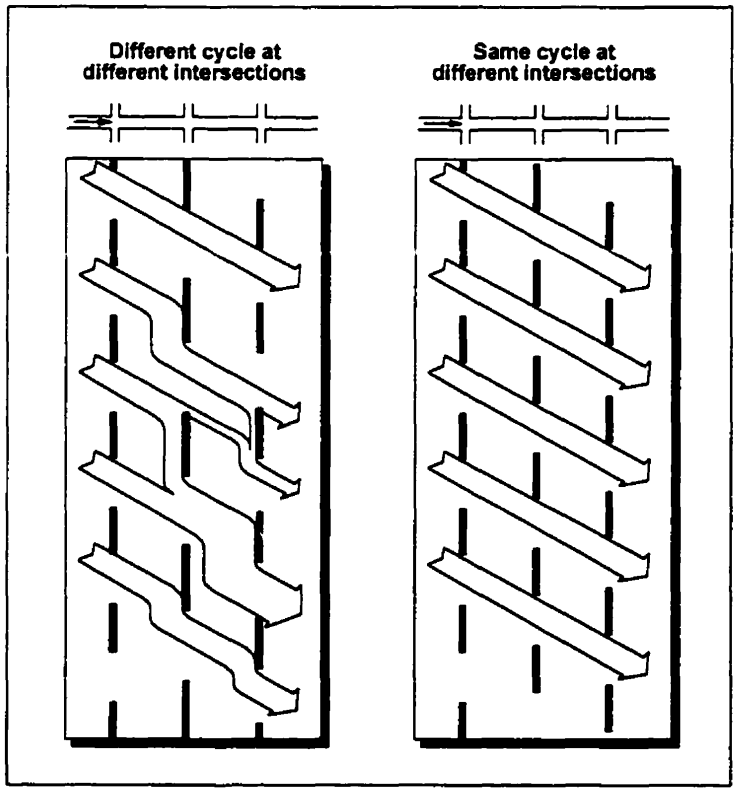


Figure 5.8 - Induced Repetitiveness of Traffic Patterns under Cyclic Traffic Signal Operation

signal operation to ensure the repetitiveness of signal timings, and consequently traffic patterns, over time. In acyclic operation, the concepts of cycle time, green allocation and signal offset are discarded. Signal operation is performed on a phase-by-phase basis, with the main control decisions being when to end the current phase and to which phase to go to next. There is in this case no common cycle time imposed on the operation of groups of intersections. Each intersection is allowed to operate under its own cycle time and to vary the duration of the signal cycle over time to best suit local traffic conditions.

The basic philosophy behind cyclic signal operation is to promote long-term area-wide signal coordination through the establishment of repetitive traffic patterns. As it can be deduced by comparing the two time-space diagrams of Figure 5.8, attempting to coordinate large groups of intersections is much easier in the presence of repetitive traffic patterns than under variable demands. Under repetitive demands, traffic engineers only have to analyze traffic arrival and departure patterns at individual intersections over a single signal cycle to efficiently plan traffic progression across sets of intersections for the entire duration of a given control period. Under variable demands, however, the analysis cannot be contained to a typical signal cycle. Signal timings must be uniquely determined for each second of the control period, which could be computationally demanding.

The basic philosophy of acyclic signal control is to improve traffic performance around individual intersections by relinquishing the restrictive characteristics of the fixed common cycle time operation. This improvement is achieved by giving signal controllers the ability to implement signal timings that truly match the needs of current local traffic demands. While imposing a common cycle time to groups of intersections eases the tasks of planning traffic movements on an area basis it also imposes a penalty on the performance of the network. First, due to the need to maintain a common cycle time, restrictions are put on the ability to adjust signal timings to local traffic conditions. For instance, in many systems transit priority requests can only be granted if the required change will not cause a change in the implemented common cycle time. In addition, since traffic demand usually varies from one intersection to the other, different intersections will usually possess different optimum cycle times. By arbitrarily requiring that all intersections in a group adopt the longest individual cycle time that must be implemented at any intersection, many intersections are then forced to operated under a cycle time that is longer than their optimum. As illustrated by the theoretical relationship between delay and cycle time shown in Figure 5.9, this use of longer cycle times might cause higher levels of delays for vehicles, mostly through the implementation of longer red intervals. In some networks, the benefits of

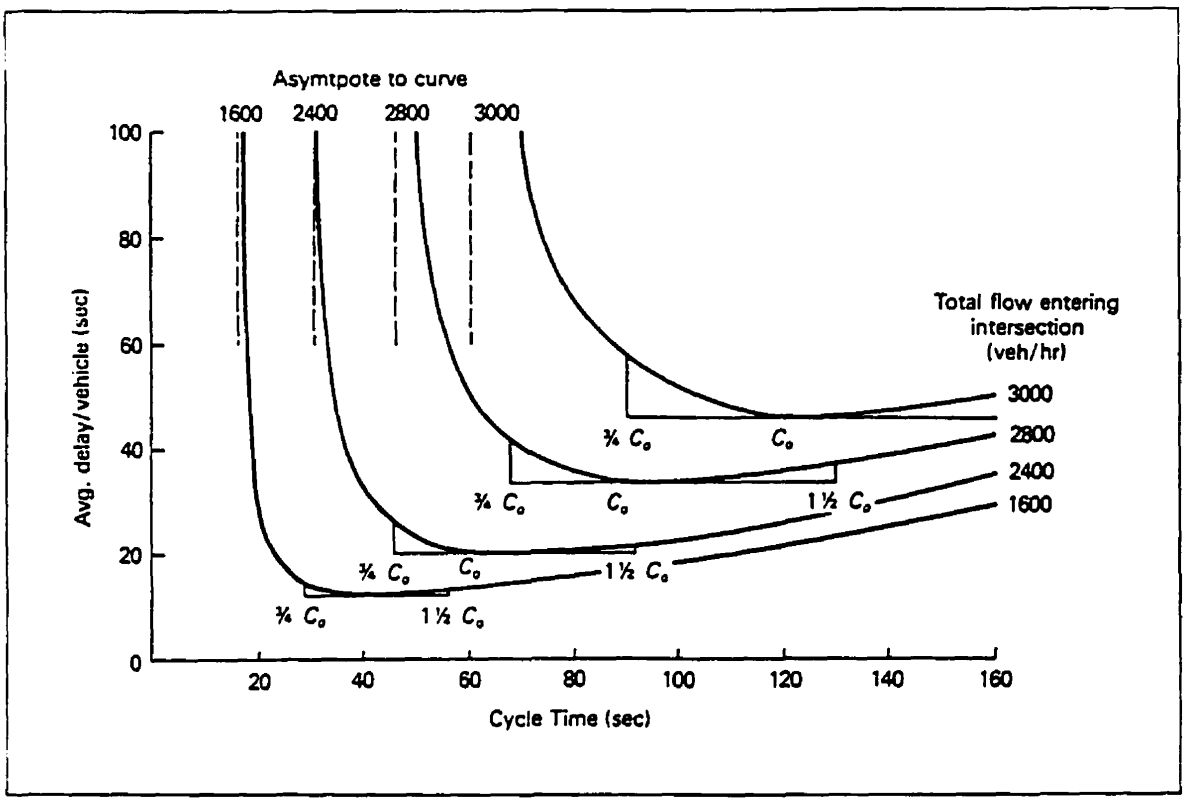


Figure 5.9 - Theoretical Relationship Between Cycle Time and Delay (from McShane, 1990)

favorably arranging progression along groups of adjacent intersections may be completely lost to additional delays attributed to the use of non-optimal cycle times at a large number of intersections.

Based on the above elements, cyclic signal operation is best adapted to networks in which there are only minor or gradual traffic variations, and in which all intersections within coordinated groups share similar levels of traffic demands. A cyclic operation can also cope with a certain degree of variability, provided that this variability does not overshadow the cyclic behavior of traffic demands. For instance, the effect of an emergency vehicle temporarily disrupting flow patterns along an arterial can easily be ignored if such disruptions occur infrequently. In the absence of any other major disruptions, the repetitiveness of traffic behavior, and thus the full benefits of arranging favorable progression patterns on an area basis, will quickly be restored through the implementation of cyclic timings.

However, planning area-wide progression patterns through the implementation of cyclic timings becomes pointless in networks in which there are large and frequent variations in traffic demands and saturation flows. A typical example of a system experiencing substantial traffic disruptions is

Toronto's Queen Street Corridor. As explained in Section 5.2, attempting to establish either one- or two-way progression patterns over the full length of this corridor is almost futile as a result of the severity and frequency of traffic disruptions between signalized intersections caused by on-line transit activities. These disruptions not only make it very difficult to identify average repetitive traffic behavior, but also result in a high degree of discrepancy between projected and actual traffic conditions.

Under such circumstances, it is better to reduce the scope of the signal optimization and provide traffic control on a more local basis. Instead of arbitrarily attempting to create repetitive patterns through the imposition of cyclic timings, individual signal controllers should be allowed to implement their own optimum timings and coordinate as required their operation with adjacent intersections only. While the removal of the common cycle time requirement would remove the guarantee that repetitive patterns will be created by the signal operation, such patterns could nevertheless still be observed. However, these patterns would be the consequence of stable and repetitive demands rather than a precondition of the signal operation.

Given the need to control urban signalized networks with potentially highly variable demands, it is reasoned that the flexibility of signal operation at individual intersections is more important for the SPPORT model than the ability to plan traffic movements on an area basis. For this reason, the use of acyclic signal timings is preferred of the use of cyclic timings. This type of control is fully compatible with distributed control architecture, as it does not require traffic information to be gathered to a single location to perform an area-wide signal optimization. This choice also follows a recent trend towards the use of acyclic signal timings in real-time traffic signal control systems. While early systems such as SCATS and SCOOT rely on the use of cyclic timings, many recent systems, such as PRODYN, and UTOPIA, are based on acyclic signal control principles.

5.5.2. Objective Function

When decisions regarding whether or not to change the signal indications at an intersection are taken at regular intervals, all feasible sequences of decisions may be represented as a decision tree similar to the one shown in Figure 5.10. In this tree, each decision has a cost associated with it that contributes to the total cost of the strategy being followed. This cost may be represented by vehicle stops, delays, a linear combination of stops and delays, or any other relevant criterion. At the end of the decision

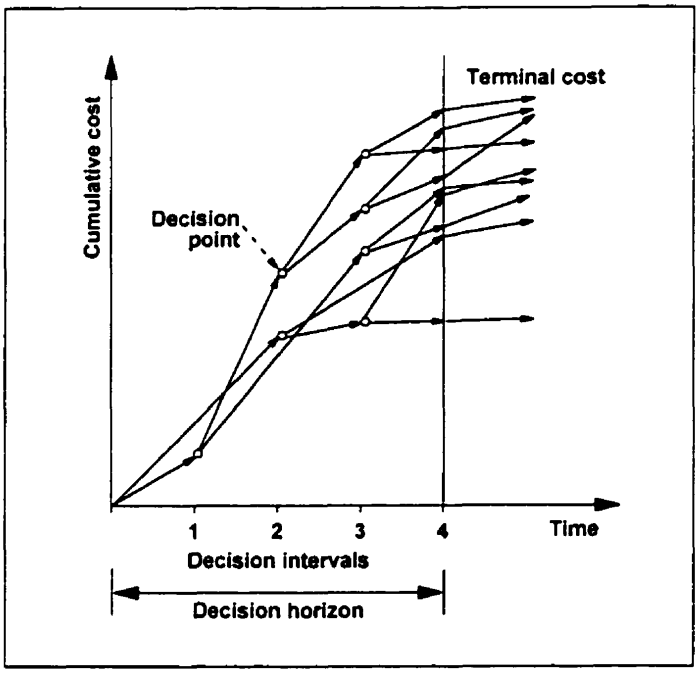


Figure 5.10 - Cumulative Cost along a Decision Tree

process, a certain number of alternative control strategies will usually be available. By comparing the cumulative cost associated with each sequence of timing decisions, the best solution can be identified at the end of the optimization process. In most cases, the best solution is taken to be the sequence of decisions that produces the least cost to vehicles or vehicle passengers over the entire decision horizon. In order to evaluate each alternative strategy, SPPORT requires the user to define the objective function that the signal controllers should use in the signal optimization process. A predefined function has not been imposed in order to permit the signal optimization process to be tailored to the specific needs of traffic signal control around each intersection. The function to use in the optimization process is defined by providing the values of the weighting factors in equation 5.1:

$$PI = \sum_{i=1}^{N_{veh}} o_v \left(k_d \, d_i + k_s \, s_i + k_{TT} \, TT_i \right) + \sum_{k=1}^{N_{int}} \left(k_{TC} \cdot TC_k \right) \qquad [5.1]$$

where: PI = Performance index.

d_i = Total delay incurred by vehicle i in the controlled system (seconds).

k_d = User-specified coefficient defining the relative importance of delay ($k_d \geq 0$).

k_s = User-specified coefficient defining the relative importance of stops ($k_S \geq 0$).

k_{TC} = User-specified coefficient defining the relative importance of the terminal cost ($k_{TC} \geq 0$).

k_{TT} = User-specified coefficient defining the relative importance of travel time ($k_{TT} \geq 0$).

N_{int} = Total number of intersections in controlled network.

N_{sim} = Total number of vehicles entering the controlled network during a given period of time.

o_v = User-specified coefficient representing the average occupancy or relative importance of a vehicle of type v ($o_v \geq 0$).

s_i = Total number of stops incurred by vehicle i in the controlled system.

TC_k = Terminal cost for intersection k at end of the performance evaluation period.

TT_i = Total travel time of vehicle i in the controlled network (seconds).

On the basis of the values provided for the parameters of equation 5-1, signal timing strategies can for instance be evaluated against the minimization of:

- total vehicle delay ($k_v = 1$, $k_s = 0$, $k_{TT} = 0$ and $o_i = 1$),
- total vehicle stops ($k_v = 0$, $k_s = 1$, $k_{TT} = 0$ and $o_i = 1$),
- total vehicle travel time ($k_v = 0$, $k_s = 0$, $k_{TT} = 1$ and $o_i = 1$),
- total person delay ($k_v = 1$, $k_s = 0$, $k_{TT} = 0$ and o_i = Vehicle occupancy),
- total person stops ($k_v = 0$, $k_s = 1$, $k_{TT} = 0$ and o_i = Vehicle occupancy),
- total person travel time ($k_v = 0$, $k_s = 0$, $k_{TT} = 1$ and o_i = Vehicle occupancy),
- a linear combination of total vehicle delay, stops and travel time, or
- a linear combination of total person delay, stops and travel time.

A terminal cost that approximates future costs incurred by vehicles or vehicle passengers beyond the end of the decision horizon as a consequence of the switching decisions taken during the decision horizon can also be added to the performance function. The purpose of this terminal cost is to counteract a bias that could lead the signal optimization process to select signal-switching decisions that yield a low cost during the decision horizon but a high cost thereafter. For example, one solution might minimize the delays incurred by vehicle passengers during the decision horizon and leave large queues at the end, while another one might cause slightly more delays during the decision horizon, but result in much smaller queues after the decision horizon.

The terminal cost function that was developed for use with the SPPORT model explicitly attempts to estimate the delays incurred beyond the end of the decision horizon by all vehicles left in a queue at

that time. In the model, this cost element is estimated on an intersection-by-intersection basis using equation 5.2:

$$TC_k = \sum_{j=1}^{N_{link\,k}} \left(\frac{Q_{end\,j}^{\,2}}{2\,q_{sat\,j}} \cdot \left[U_j + \alpha_R (1 - U_j) \right] + Q_{end\,j}\, R_{min\,j} \left[U_j \right] \right) \qquad [5.2]$$

where: TC_k = Terminal cost for intersection k.

$N_{link\,k}$ = Number of approach links to intersection k.

α_R = User-specified coefficient defining the relative importance of vehicles stopped on an approach with a green signal indication at the end of the decision horizon with respect to vehicles stopped on an approach with a red signal indication ($\alpha_R \geq 0$).

$Q_{end\,j}$ = Queue size on link j at end of decision horizon (passenger car units).

$q_{sat\,j}$ = Saturation flow on approach link j (passenger car units/second).

U_j = Signal display on approach link j (0 if signal is green, 1 if signal is red).

$R_{min\,j}$ = Shortest remaining red duration on approach link j (seconds).

Equation 5.2 estimates the total delay incurred by Q vehicles waiting in a queue that can dissipate at a maximum rate of q_{sat} passenger car units per second. In this equation, passenger car units (ITE, 1995) represent the total number of average passenger cars that can cross the stop line every second. Vehicles are not used in the definition to reflect the fact that transit vehicles occupies more space than passenger car (in the research, it is assumed that one transit vehicle corresponds to two passenger cars). The Figure 5.11 illustrates more precisely how this calculation is done. R_{min} represents the minimum time that vehicles must wait before they can start to cross the intersection. The value of this parameter is a function of the current signal display, the time at which the last phase was implemented, the order in which phases can be displayed, the minimum duration of each phase, and the time lost at the beginning of a green interval due to drivers' reaction time. To avoid overestimating delays by assuming excessively long waits, the maximum remaining red duration, R_{min}, is assumed not to exceed 60 seconds in cases in which the next green interval is beyond the end of the current decision horizon and for which SPPORT cannot predict the soonest green signal indication return time. This value has been chosen as it corresponds to a reasonable approximation for red interval duration in the absence of any other information. Given that a typical cycle time will usually vary between 40 and 120 seconds, a maximum red duration of 60 seconds assume that the signal display would be red, at best, half of the time, and at worse, all the time.

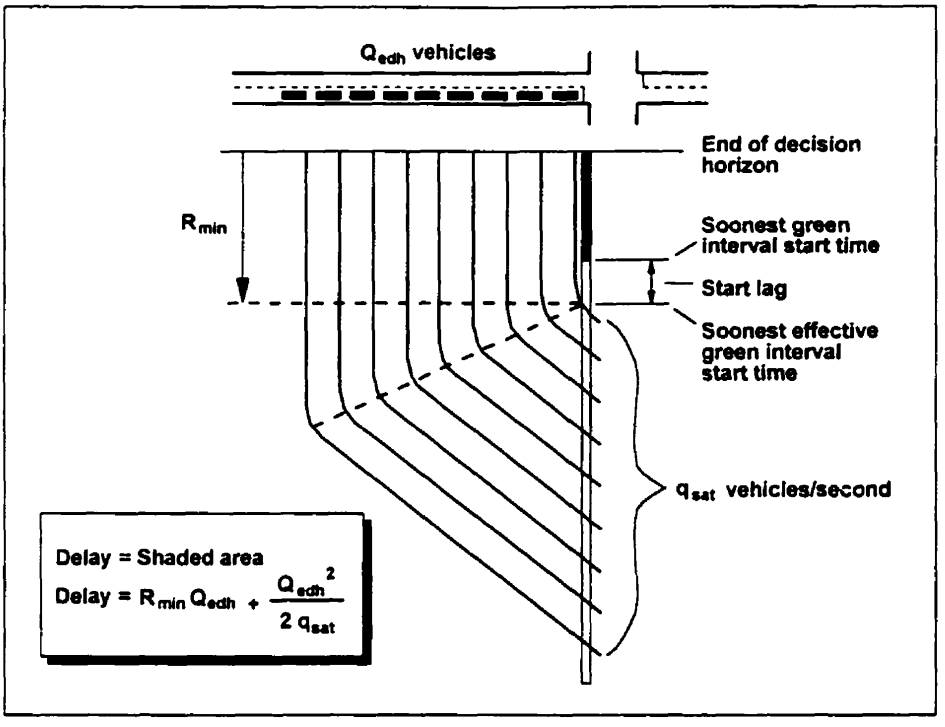


Figure 5.11 - Calculation of Terminal Cost at end of Decision Horizon

To conclude, Figure 5.11 also indicates that the terminal cost calculation conservatively assumes that no vehicle joins the tail of the queue during the dissipation process. This assumption is made to reflect that projected arrivals are only known within the decision horizon. The average observed arrival rates could have been used to represent future traffic demands pass the end of the decision horizon, but it was recognized that that this average rate would not correctly represent the typical cyclic arrival patterns observed at most signalized intersections in coordinated networks. If it is observed that the above assumption is too conservative, the weight of the parameter k_{π} in equation 5.1 could be increased to give more importance to the terminal cost element.

5.5.3. Rule-Based Switching Decision Process

A number of algorithms can be used to find the optimal sequence of signal switching decisions over a given period of time. In OPAC, for instance, a heuristic constrained search procedure based on backward dynamic programming is used to generate signal timings (Farradyne Systems, 1989). In PRODYN, forward dynamic programming is directly used to optimize the signal timings (Henry et al. 1983). In UTOPIA, heuristic and branch-and-bound optimization algorithms are used. In SCOOT, signal timing adjustments are either calculated directly from estimated traffic conditions or determined

through a hill-climbing procedure evaluating the effect of predefined small changes in signal timings and keeping those changes that improve traffic performance.

In the SPPORT model, a rule-based heuristic optimization procedure is used to find the best sequence of switching decisions. This procedure, which is illustrated in Figure 5.12, was initially developed by Yagar, Han and Greenough (1991, 1992) in response to concerns that exhaustive optimization procedures such as dynamic or linear programming may be too computationally demanding for real-time application in urban networks with highly variable demands. The procedure is based on the recognition that signal switching decisions will usually occur after the realization of specific discrete events, such as after a queue of vehicles has reached a certain size, after a queue has just finished dissipating, or after the detection of an incoming transit vehicle. The SPPORT model consequently assumes that considering only a finite number of important events can make efficient signal control decisions. By ignoring all events that have no importance for the signal operation, the model is able to significantly reduce the number of potential switching combinations that need to be considered to find an optimum solution to the traffic control problem. This process, which replicates what many traffic control officers do in their mind when they are asked to manually control traffic at signalized intersections, makes the SPPORT model more amenable to real-time traffic signal control.

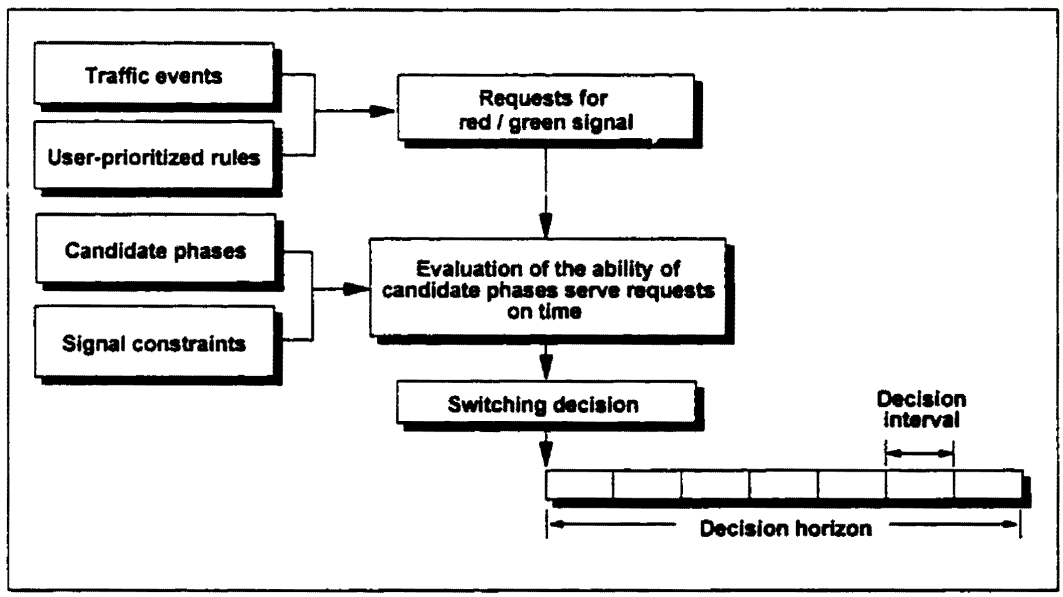


Figure 5.12 - Heuristic Rule-Based Signal-Switching Decision Process

To account for the fact that different traffic events do not carry the same importance, the SPPORT model requires the user to prioritize all the traffic events that are defined to influence the signal optimization process. The heuristic signal optimization algorithm then generates requests calling for either a green or red signal indication on specific approaches at specific times in response to the identification of any prioritized event in a traffic simulation projecting traffic conditions over the decision horizon. Using a sophisticated decision-making process that will be described in detail in Chapter 8, the model finally generates signal switching decisions so as to accommodate as best as possible the list of signal indication requests obtained at the end of the previous optimization step.

5.5.4. Multi-Objective Optimization Process

While the use of prioritized lists of events allows SPPORT to determine the relative importance of various events, it is often very difficult to determine beforehand which event should have the highest priority. For example, it may be established very easily that the need to start serving a queue of 50 vehicles is more important than the need to start serving a queue of only 10 vehicles. However, it may not be determined with such ease if providing priority to transit vehicles should have a higher or lower importance than serving a queue of passenger cars. The problem associated with such situation is that providing priority to incoming transit vehicles may be beneficial in some cases while it may not be in other cases.

To solve the above problem, the user is allowed to provide more than one prioritized list of events for consideration by the SPPORT model. When more than one list is provided, the optimization algorithm is instructed to generate a candidate signal timing plan for each one of them and then to select for implementation the one yielding the best performance measures. While such process adds the need to compare various signal control strategies to the complexity of the signal optimization, it provides the SPPORT model with the ability to simultaneously consider many traffic control strategies and to automatically switch its main strategy when changes in traffic conditions are detected.

Figure 5.13 graphically summarizes the process used by SPPORT to generate signal-switching instructions over a given decision horizon. While many elements on this diagram were briefly introduced in this chapter, they will be described in more detail in the next three chapters. Chapter 6, for instance, will describe the discrete-event simulation model used to project future traffic behavior around controlled intersections and to evaluate candidate phase plans. Chapter 7 will then explain how

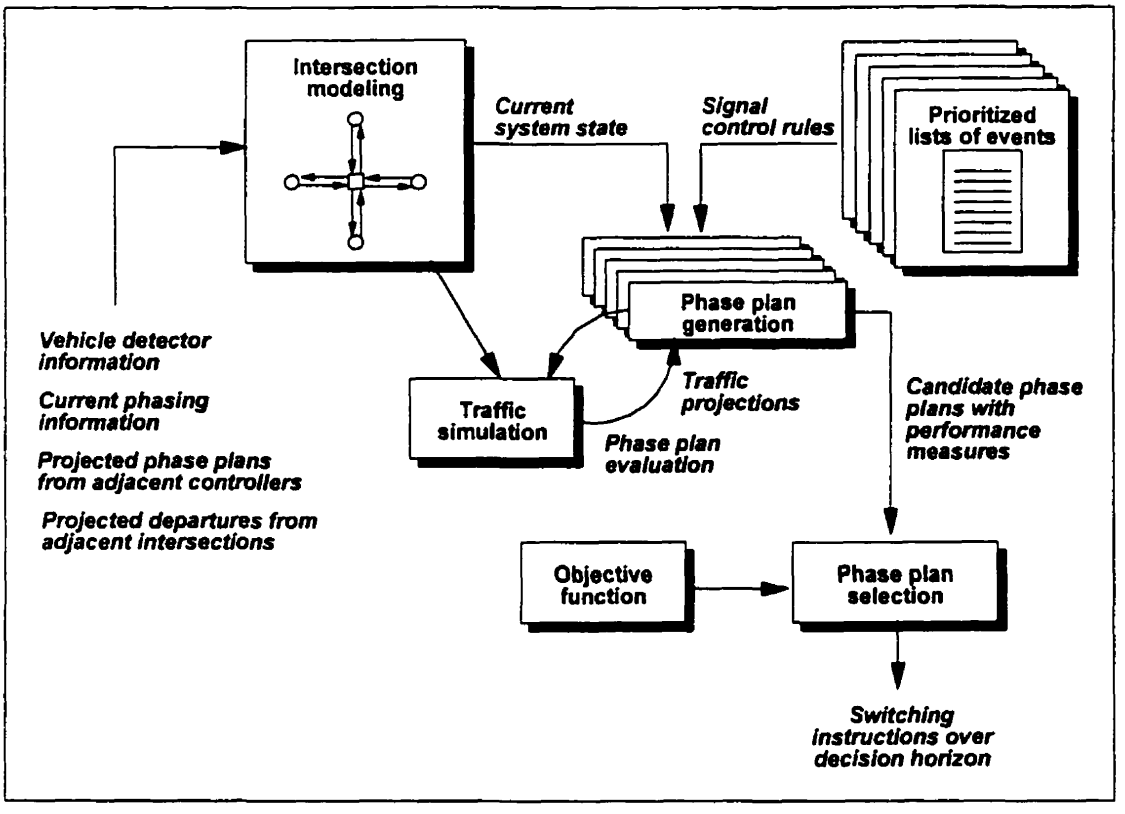


Figure 5.13 - Multi-Objective Decision-Making Process

future traffic demands are estimated for the purpose of generating signal-switching decisions, while Chapter 8 will describe in detail the heuristic rule-based optimization logic used to generate candidate signal timing strategies for each prioritized list of events provided by the user. The following chapter, Chapter 9, will finally presents an evaluation of the signal optimization process, including the results of a study evaluating the benefits of the multi-objective optimization process shown in Figure 5.13.

6. Discrete-Event Traffic Simulation Model

Simulation plays an important role in the signal optimization process of many real-time, traffic responsive signal control systems. In many systems simulation is used to predict future arrival patterns at signalized intersections on the basis of historical data and/or detector information. In some systems, simulation is also used to evaluate the impacts of potential future switching decisions on general traffic performance and to determine which decisions would best meet the needs of current and projected near-future demands.

Similar to other existing real-time traffic signal control systems, SPPORT relies heavily on traffic simulation to generate signal timing strategies that meet the needs of prevailing traffic demands. The SPPORT model makes use of a discrete-event microscopic traffic simulation model that was developed by Conrad (1997) using the C++ programming language and object-oriented software development techniques. This model, which was explicitly developed to be used with the SPPORT model, was designed to be detailed enough to result in the realistic simulation of traffic flow between signalized intersections, but not so detailed as to required excessive amounts of configuration information and processing time. Its main characteristic, when compared to other simulation models, is its ability to model transit interference on other traffic in a way that is computationally efficient, providing sufficiently short simulation times.

During the course of the current research project, revisions were made to the simulation model developed by Conrad to enhance its efficiency and improve its modeling of urban traffic behavior. The intent of this chapter is not to describe every change made to the model. Such a list would not help the reader to understand how the SPPORT model operates. The intend is rather to present an overview of the simulation model in a way that would help the reader understand how SPPORT projects future traffic conditions and evaluate candidate signal timing strategies. One important element of this description is how intersection approaches are modeled, as this modeling has an effect on the ability of the SPPORT model to analyze traffic conditions on intersection approaches. The description of the simulation model starts with a general overview of the principles of discrete event simulation. This is followed by a more detailed description of the main components of the model.

6.1. Discrete-Event Simulation Process

Discrete-event simulation involves the modeling of a system as it evolves over time by a representation in which the variables describing the state of the system change instantaneously at separate points in time (Law and Kelton, 1991). These points in time are the ones at which an event occurs, where an event is defined as an instantaneous occurrence that may change the state of the system.

During a discrete-event simulation, the simulation clock can be advanced using one of two fundamental approaches: next-event time advance and fixed-increment time advance. In the fixed-increment time advance, the simulation clock is advanced in regular steps. In the next-event approach, the clock is initialized to zero and the time of occurrence of future events is determined. The clock is then advanced to the time of occurrence of the most imminent of these future events, at which point the state of the system is updated to account for the fact that an event has occurred. When the update is completed, the simulation clock is again advanced to the time of the most imminent event and the system updated. This process is repeated until a stopping condition is found.

In SPPORT, the next-event time advance approach is used to move the simulation forward. Conrad (1997) has chosen this approach on the fact that it is used by all major simulation languages and by most people coding their model in a general-purpose simulation language. In addition, while the fixed-increment time approach is usually more straightforward and least restrictive, the next-event time advance approach is usually more efficient, as it allows a simulation to jump over periods of inactivity. To further compare the two approaches, the fixed-increment approach can also be seen to be a special case of the next-event approach in which time intervals are defined as events.

6.2. Components of the Simulation Model

Regardless of the time advance mechanism used, all discrete-event simulation models consist of an idealized representation of a real-world system. This idealized system is usually defined using three kinds of elements: resources, traffic units and activities. The resources represent the elements of the idealized system that are subject to utilization, while the traffic units are the elements that seek utilization of these resources. The sequences of activities define the processes that traffic units undertake in order to utilize these resources.

This section describes the basic elements of the simulation model used by SPPORT in its signal optimization process. Specifically, this section describes the vehicles, segments, links, vehicle activities and signal controllers used by the simulation model to realistically simulate traffic behavior in mixed-traffic environments. Unless otherwise noted, the various elements described in this chapter are mainly the contribution of Conrad (1997).

6.2.1. Vehicles

Vehicles are the basic unit of flow in the simulation model. They are the units that successively attempt to engage and utilize the various resources of the system, causing their state to vary. Currently, two kinds of vehicles are defined within the model: private vehicles and transit vehicles. Transit vehicles are required to stop at specific locations to load and unload passengers while private vehicles do not have these requirements. No other types of vehicles have been defined, as current needs do not require the signal control system to make further distinction among vehicle types.

In the simulation model, each vehicle is given upon generation a unique identification number. This number is currently generated by the simulation model, but it could be possible to obtain these numbers from an external source if a vehicle identification system is used. For transit vehicles, the route number must also be specified. This number was not required in the initial model, as it was required to separately model the path followed by transit vehicles on each transit route. In the current version of the simulation model, the same list of vehicle activities is used for both private and transit vehicles, with the transit route number used by the vehicle activity objects to control the movements of transit vehicles at points where a choice of paths is possible.

In the model, each vehicle is also given a specific size in terms of passenger car units (ITE, 1995), and a specific weight for the purpose of recording travel time, delays, stops and any other relevant performance measure that might be found useful. Both these parameters are used-defined. The size of a vehicle determines how much space a particular vehicle occupies while stopped in a queue and how much time is uses when it crosses the stop line of a given approach or the boundary between two segments. Vehicle weights, on the other hand, allow the simulation model to be configured in such way that the recorded performance measures reflect vehicle occupancy, transit vehicle operating costs, or any other considerations. While separate weights for stops, delays and travel times were initially defined by Conrad (1997), it was estimated in the current research project that the use of a single

weight for all three statistics would be more appropriate, as all three statistics could then be compared on the same basis.

6.2.2. Segments

The segments are the resources of the idealized transportation system. In the simulation model, each segment represents a uniform and unidirectional section of roadway typically 25- to 50-metres long along which traffic units attempt to travel. In order to provide full flexibility in the modeling of intersection approaches, segments are not arbitrarily given fixed characteristics. For each segment being used, the user provides the following characteristics:

- Length (metres);
- Number of traffic lanes;
- Maximum queue density (passenger car units/metre/lane);
- Saturation flow (passenger car units/second/lane);
- A statistical distribution of vehicle speeds returning speeds in metres per second;
- Number of downstream segments that an exiting vehicle can engage next (Figure 6.1);
- Identification number of each downstream segment a vehicle can engage next;
- Proportion of vehicles exiting to each defined downstream segment.
- Number of opposing segments carrying flow interfering with the segment's exit flow;
- Identification number of all opposing segments carrying interfering flow.

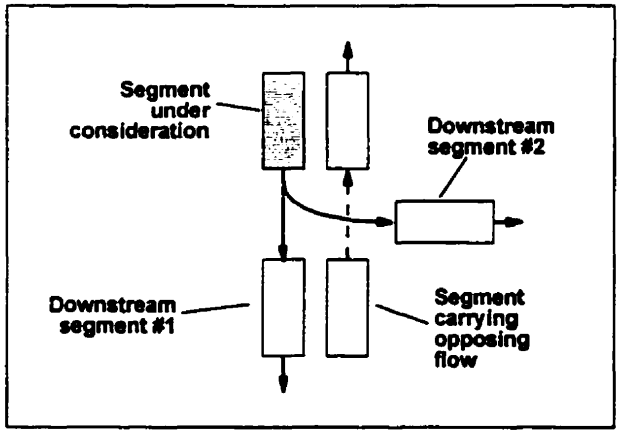


Figure 6.1- Distinction Between Downstream and Opposing Segments

In this list, the elements identifying the downstream segments a vehicle can engage next and the segments carrying interfering flows were added in the current research project. In the model developed by Conrad (1997), segments were not directly linked to their downstream neighbors. Each segment was simulated as an isolated element, with the vehicle activities determining which segment a vehicle should engage next and the probability of choosing a given segment in cases in which there were more than one choice. This modeling was changed as it complicated the analysis of traffic conditions along sequential series of segments. There was also no concept of interfering traffic. For instance, vehicles making a left turn were always assumed not to be affected by vehicles traveling in the opposing direction on the same street.

The characteristics listed above are used to control the process by which traffic units can engage segments and to control traffic behavior within each segment. They are also used to compile a series of performance measures during the course of a simulation, which can later on be used by the signal optimizers to evaluate candidate signal timing strategies (see Section 5.5.2). Two types of performance measures are currently compiled: weighted statistics and non-weighted statistics. The weighted statistics consider the vehicle weights provided by the user for each type of vehicle considered in the signal optimization process (i.e., the parameter o_v in Equation 5.1) while the non-weighted statistics do not (i.e., parameter o_v assumed equal to 1). More specifically, the measures currently recorded by each segment are:

- Total vehicle stops (passenger car units);
- Total transit stops (passenger car units);
- Total weighted vehicle stops (passenger car units);
- Total weighted transit stops (passenger car units);
- Total vehicle delay (seconds);
- Total transit delay (seconds);
- Total weighted vehicle delay (seconds);
- Total weighted transit delay (seconds);
- Total vehicle travel time (seconds);
- Total transit travel time (seconds);
- Total weighted vehicle travel time (seconds);
- Total weighted transit travel time (seconds);

In the list, the primary measures are the total weighted and non-weighted vehicle stops, vehicle delay and vehicle travel time. These measures reflect the stops, delays and transit times compiled for all types of vehicles during the course of a simulation. They are also the performance measures used in Equation 5.1 to calculate the performance index of a given signal control strategy. Currently, the total weighted and non-weighted transit measures are only compiled for analysis purposes. The calculation of these measures allow the user to determine the portion of the total vehicle stops, delay and travel time that was in fact incurred by transit riders.

During the course of a simulation, each segment also maintains a list of all the vehicles it contains at a given moment. This list contains the upstream end arrival time, the projected downstream end arrival time, and the size in passenger car units of each vehicle currently within the segment. In the list, the projected downstream end arrival time is based on the segment's free flow speed. Each time a vehicle enters the segment, the probability distribution modeling the free-flow speed behavior on the segment is sampled and the resulting speed assigned to the entering vehicle. The downstream end arrival time is then determined by calculating the time required by the vehicle to travel across the segment at the sampled speed and by adding this travel time to the upstream end arrival time.

The information contained in the list is used to control traffic behavior and estimate traffic conditions in each segment. For example, the number of vehicles in the list is used to determine whether the segment is full and whether additional vehicles can enter. The information is also used to evaluate the number of vehicles waiting in queue at the downstream end of a segment to exit the segment. Queued vehicles are determined by comparing for each vehicle the current clock time with the time at which the vehicle was expected to exit the segment and by considering as queued any vehicle still present on the segment after its expected exit time.

The first simulation control carried out by a segment is to meter the entry of vehicles at its upstream end. If the segment is full, entry is denied to any new vehicle. If there is still some storage capacity left with respect to the total number of vehicles that the segment can hold, incoming vehicles are allowed to enter the segment. In order to realistically simulate the rate at which vehicles enter a segment, the interval between successive entries is calculated using equation 6.1:

$$t_{next\,m} = Max\left(t\,,\, t_{prev\,m} + \frac{l_{prev\ veh\,m}}{\beta_m\ q_{sat\,m}} \right) \qquad [6.1]$$

where: $t_{next\ m}$ = Next possible entry time at upstream end of segment m (seconds).

t = Current time (seconds).

$t_{prev\ m}$ = Previous entry time at upstream end of segment m (seconds).

$l_{prev\ veh\ m}$ = Size of previous vehicle having entered segment m (passenger car units).

$q_{sat\ m}$ = Saturation flow on segment m (passenger car units/second).

β_m = Gating factor on segment m ($0 \geq \beta_m \geq 1$).

In Equation 6.1, $q_{sat\ m}$ is the maximum flow that can enter the segment m under normal traffic conditions over all traffic lanes. The model simulates road segments as single unidirectional pipes. Individual lanes are not modeled. A segment having a capacity equal to the combined capacity of all lanes models roadways with multiple lanes in a single direction. For example, a segment able to carry 1800 vehicles per pour per lane would allow a vehicle to enter every two seconds if it contains one traffic lane and one vehicle every second if it contains two traffic lanes.

The parameter β allows additional gating effects such as temporary lane closures to be introduced in the simulation. Under normal traffic conditions, this parameter carries a default value of 1. This value indicates that the maximum rate at which vehicles can enter a segment corresponds to the saturation flow of the real-world segment of roadway. Reduced values may then be assigned to this parameter during specific time intervals to represent situations in which the entry rate is temporarily lower than the normal saturation flow. For instance, the blockage of a traffic lane on a two-lane segment could be represented by attributing a value of 0.5 to the parameter β between time A and time B. At the limit, a complete blockage of the segment would be modeled by attributing a value of 0 to the parameter β.

An improvement over the model developed by Conrad (1997) was made in the area of gate factors by allowing more than one gating factor to be imposed at a given time. In the initial model, segments could only store one gating factor, with only one start time and one end time. Each time the gating factor would change, the old information stored in the segment has to be discarded. In order to allow a more flexible modeling of flow interference, each segment can now hold a list of factors, with each one having its own start and end time. When two factors are scheduled to be effect at the same time, the simulation model then imposes the one resulting in the highest saturation flow reduction.

Once a vehicle has entered a segment, its travel speed is determined by sampling the speed probability distribution for that segment provided by the user. Speed sampling for an individual vehicle is done

each time it enters a new segment. As a result, a vehicle may not always travel at the same speed throughout the entire simulated network. The speed given to each vehicle is then used to determine the time at which this vehicle is expected to exit the segment at its downstream end. This time is calculated assuming that the vehicle will not be stopped along the segment.

Typical of many macroscopic simulation models, it is assumed that vehicles are queued vertically at the downstream end of each segment. The major problem with the use of this type of queuing model is that it underestimates the number of vehicles in the queue. Since vehicles have some physical length, as a queue grows, vehicles joining the tail of a queue do so at a location that keeps moving upstream. As a result, vehicles do not necessarily join the queue at, or near, the downstream end of a segment.

To realistically model horizontal queuing behavior, it is assumed that vehicles can only start to enter a segment that has been previously full of queued vehicles when the front of the dissipating queue reaches the upstream end of the segment. The time required for the front of a dissipating queue to move from the downstream end of a segment to the upstream end of the same segment is estimated using equation 6.2:

$$t_{next\,m} = t_{full\,m} + Q_{tot\,m} \cdot \frac{1}{q_{sat\,m}} - \frac{L_m}{V_m} \qquad [6.2]$$

where: $t_{next\,m}$ = Next possible entry time at upstream end of segment *m* (seconds).

$t_{full\,m}$ = Time that a first vehicle exited segment *m* after it became full (seconds).

$q_{sat\,m}$ = Saturation flow on segment *m* (passenger car units/second).

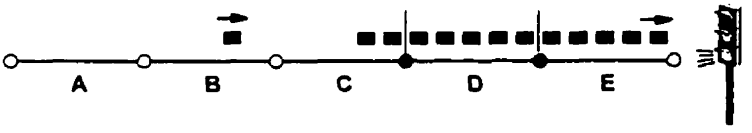
$Q_{tot\,m}$ = Total queue storage capacity on segment *m* (passenger car units).

L_m = Length of segment *m* (metres).

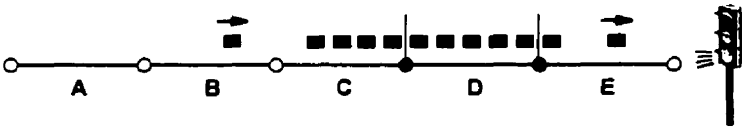
V_m = Average free flow travel speed on the segment (metres/second)

As illustrated in Figure 6.2 and in the simulation output of Figure 6.3, the use of Equation 6.1 combined with the use of short segments contribute to a realistic simulation of horizontal queuing behavior despite the fact that vehicles are assumed to queue vertically within each segment. In Figure 6.3, it can notably be observed that the front of the queue progressively moved upstream after the initiation of the green interval and that vehicles continue to join the tail of the queue until the front of the queue reaches the last stopped vehicle. Such behavior is what one would expect in reality. The queuing process was also further validated by comparing the delays reported by the simulation model

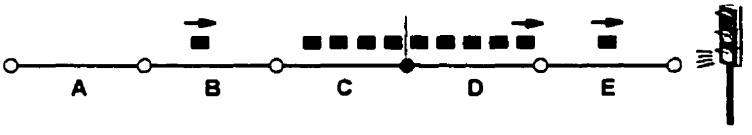
a) Vehicles in queues wait at the stop line just as the signal indication turns green.



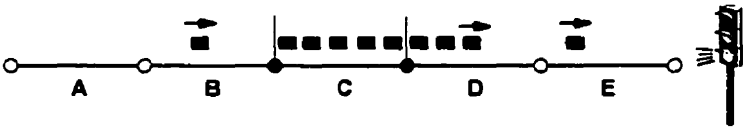
b) The front of the queue has not yet reached the upstream end of segment E. Exit from segments C and D still not allowed.



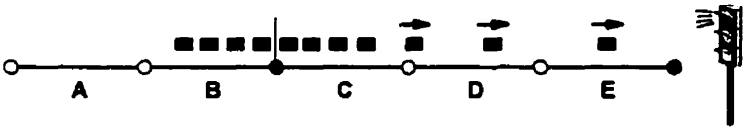
c) The front of the queue has just reached segment D. Exit from segment D allowed. Exit from segment C still blocked.



d) Segment C becomes full. Exit from segment C is still not allowed and exit from segment B is now prohibited.



e) The front of the queue reaches segment C. Exit from C allowed, but not from B. The downstream signal indication turns red.



f) A new queue forms at the stop line.

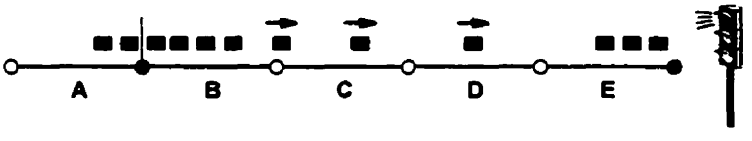


Figure 6.2 - Horizontal Queueing Behavior within Simulation Model

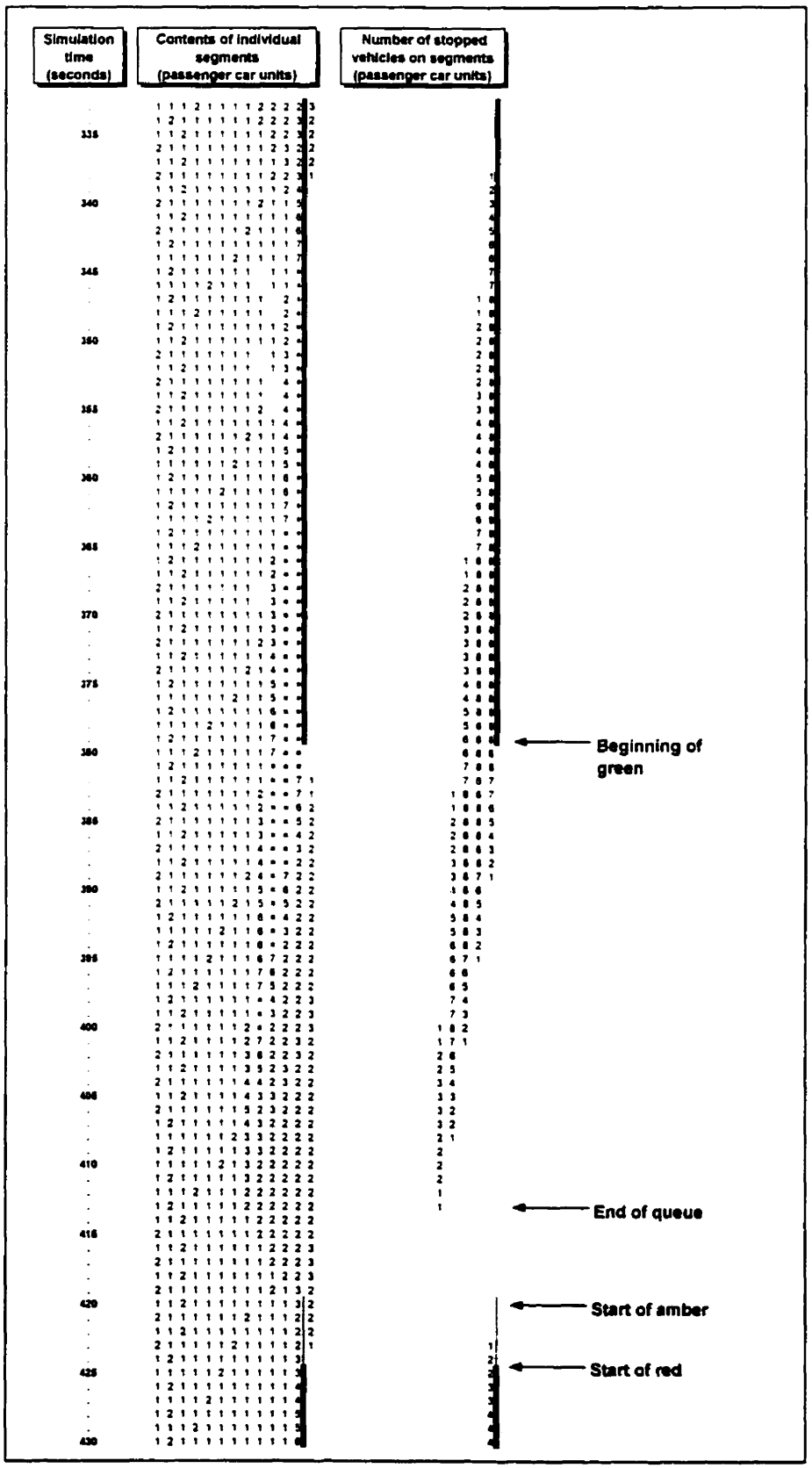


Figure 6.3 - Example of Simulated Horizontal Queue Behavior

in a series of scenarios with the delays obtained by manually projecting the trajectory of individual vehicles in the same scenarios.

In addition to the generic segment object described above, two other specialized segments were also developed for performing particular simulation tasks:

- a signal segment
- a transit stop segment

Signal segments are specialized segments performing additional entry flow control to simulate the presence of a traffic signal at their upstream end. These segments only permit vehicles to engage them while the traffic signal is in an effective green interval. The relationship between displayed and effective green intervals is illustrated in Figure 6.4.

In addition to considering the effects of signal timings, signal segments can also simulate the gap acceptance process followed by drivers attempting to cross an opposing stream. This logic was added during the current research effort, as it was not initially part of the simulation model. While this logic can be used on all segments, some problems exist when it is applied to multilane segments carrying more than one traffic movements, such as segments shared by through and left-turning vehicles (see Figure 6.5). The problem is associated with the gap acceptance logic comes from the fact that traffic progression is not simulated on a lane-by-lane basis. Since all vehicles are assumed to travel in a single pipe, any vehicle stopping at the downstream end of a segment to seek an appropriate gap in an opposing stream cause a complete blockage of the segment, even though it may only blocked one traffic lane in the real world. Solutions to fix this problem were researched; however, no solution completely compatible with the simulation model has been found. Therefore, it is temporarily

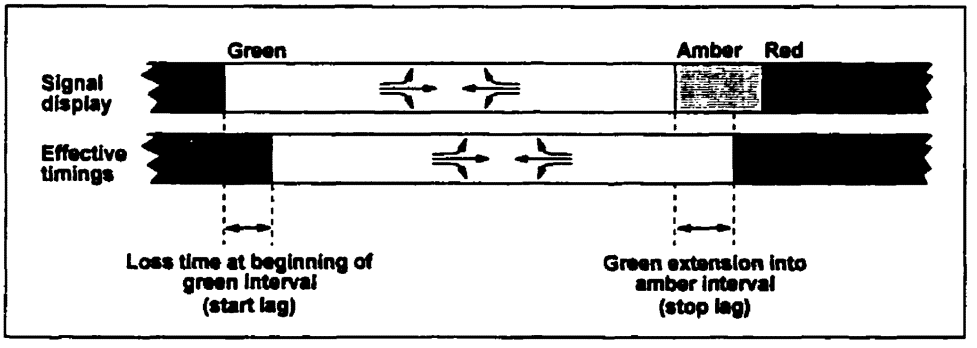


Figure 6.4 - Relationship between Traffic Signal Indications and Effective Timings

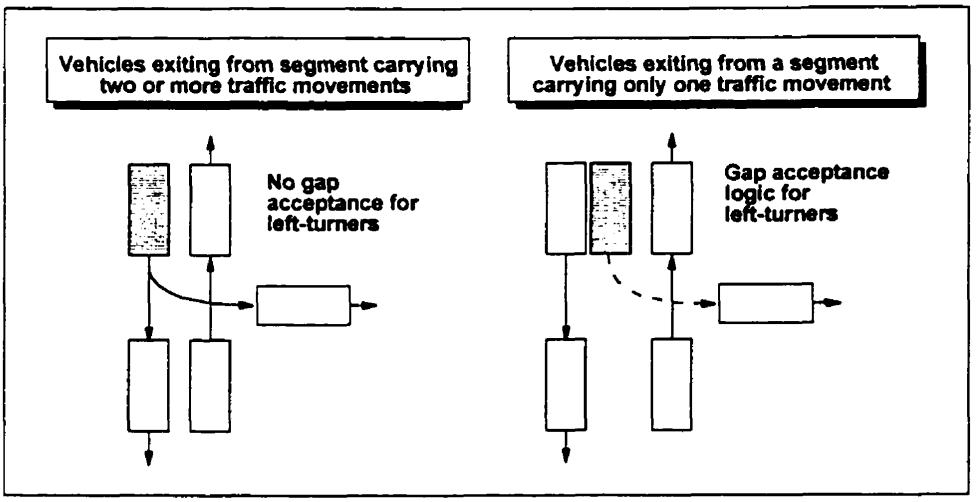


Figure 6.5 - Modeling of Left-Turning Movements within Simulation Model

recommended to use only the gap acceptance logic on single-lane segments and segments carrying only one traffic movement.

A transit stop segment is another specialized segment that extends the gate functionality described above, but this time for modeling transit activities carried on-line. In addition to the characteristics normally found in a segment, each transit stop segment also possesses a user-defined statistical distribution of dwell times and predefined transit gate factors indicating by how much the segment's normal saturation flow rate is reduced when a transit vehicle is loading and unloading passengers. This specialized segment thus possesses the ability to simulate transit loading activities and the interference these activities cause on other traffic. Whenever a transit vehicle enters the segment, a dwell time is sampled from the associated probability distribution and a temporary gate factor is imposed on the segment's maximum allowable entry flow rate for the duration of the load period. The time required by the transit vehicle to reach the downstream end of the segment is also adjusted to account for the time spent loading and unloading passengers.

6.2.3. Links

A link is a collection of segments that carry a given traffic movement. Figure 6.6 illustrates the modeling of a multiple lane intersection approach in terms of segments and links. In this example, all segments carrying though traffic are grouped into one link, and all segments modeling the left turn bay into a second link. Generally, approaches could be divided into components by following a procedure similar to the lane group analysis procedure described in the Highway Capacity Manual (TRB, 1994).

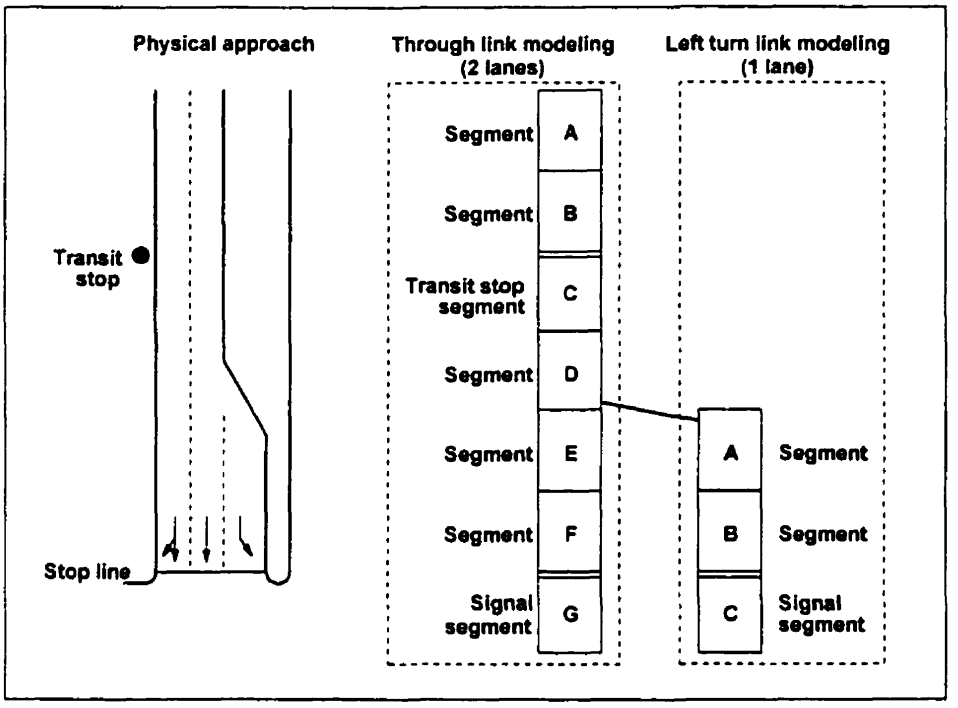


Figure 6.6 - Modeling of a Multiple Lane Approach in Terms of Segments and Links

Each link may contain any number of segments and transit stop segments; however, each link must contain one and only one signal segment, which will usually be the last one in the sequence of segments along a link. Transit stop segments can be placed anywhere upstream of a signal segment. They can be placed just upstream of the signal segment to model a transit stop located near the intersection, or at a more upstream location to model a mid-block transit stop or a stop located just downstream of the upstream intersection. As an example, note the placement of the signal and transit segments in the example of Figure 6.6.

During the course of a simulation, links are also used to compile relevant statistics over groups of segments. For example, the total number of vehicles present on a link can be obtained by summing the number of vehicles on all the segments constituting the link. Similarly, the cumulative travel time, number of stops and amount of delay spent by all vehicles that have traveled on the link can be obtained by summing the corresponding statistic over all segments in the group. The stop line queue size can be determined by counting the number of stopped vehicles in the queue starting on the segment located immediately upstream of the link's signal segment. The reach of the stop line queue is determined by searching for the first upstream segment that is not completely full of queued vehicles. Any additional queuing occurring upstream of the identified segment is then assumed to belong to

different queues and not to be the result of the current red signal indication. Consequently, the reach of the most upstream queue is associated with the most upstream segment on which there are queued vehicles.

6.2.4. Vehicle Activities

Vehicle activities represent the tasks that individual vehicles must undertake in order to move along a given path through the idealized system. When undertaking these activities, the traffic units engage and disengage the associated resources, causing their state to vary. Queuing then occurs each time that traffic units are denied access to a resource for any reason.

To realistically model traffic behavior, Conrad (1997) defined the six basic tasks based on the principles followed by the GPSS simulation program (Shriber, 1974). While many technical modifications were made in the coding of these objects in the course of the current projects, their basic operating principles remain the same. These six tasks are:

- *Begin Path* - Enter the simulated system on the basis of user-specified flow parameters.
- *Enter (Segment X)* - Enter the specified segment, if permitted.
- *Travel (Segment X)* - Travel the length of the specified segment.
- *Exit (Segment X)* - Exit the specified segment.
- *Choose Path* - Select at random using a user-defined probability distribution the next activity among a set of alternatives.
- *End Path* - Exit the simulated system.

Figure 6.7 illustrates how this set of tasks would be used to model vehicle activities for the approach illustrated in Figure 6.6. In this example, *Begin Path* objects mark the beginning of any activity process and introduce vehicles into the idealized system. These objects randomly generate vehicles on the basis of a user-specified statistical distribution of inter-arrival times. The user may also specify the duration of platooned arrivals ad the interval between successive platoons to generate cyclic platooned arrivals. During a single simulation, a specific *Begin Path* activity can only generate one type of vehicle and all the vehicles of this type that are generated are assigned the same size and same weight. Therefore, to simultaneously introduce private and transit vehicles, more than one *Begin Path* activity must be used. This is illustrated in Figure 6.7, in which two *Begin Path* objects are used in parallel to introduce both private and transit vehicles at the upstream end of segment A.

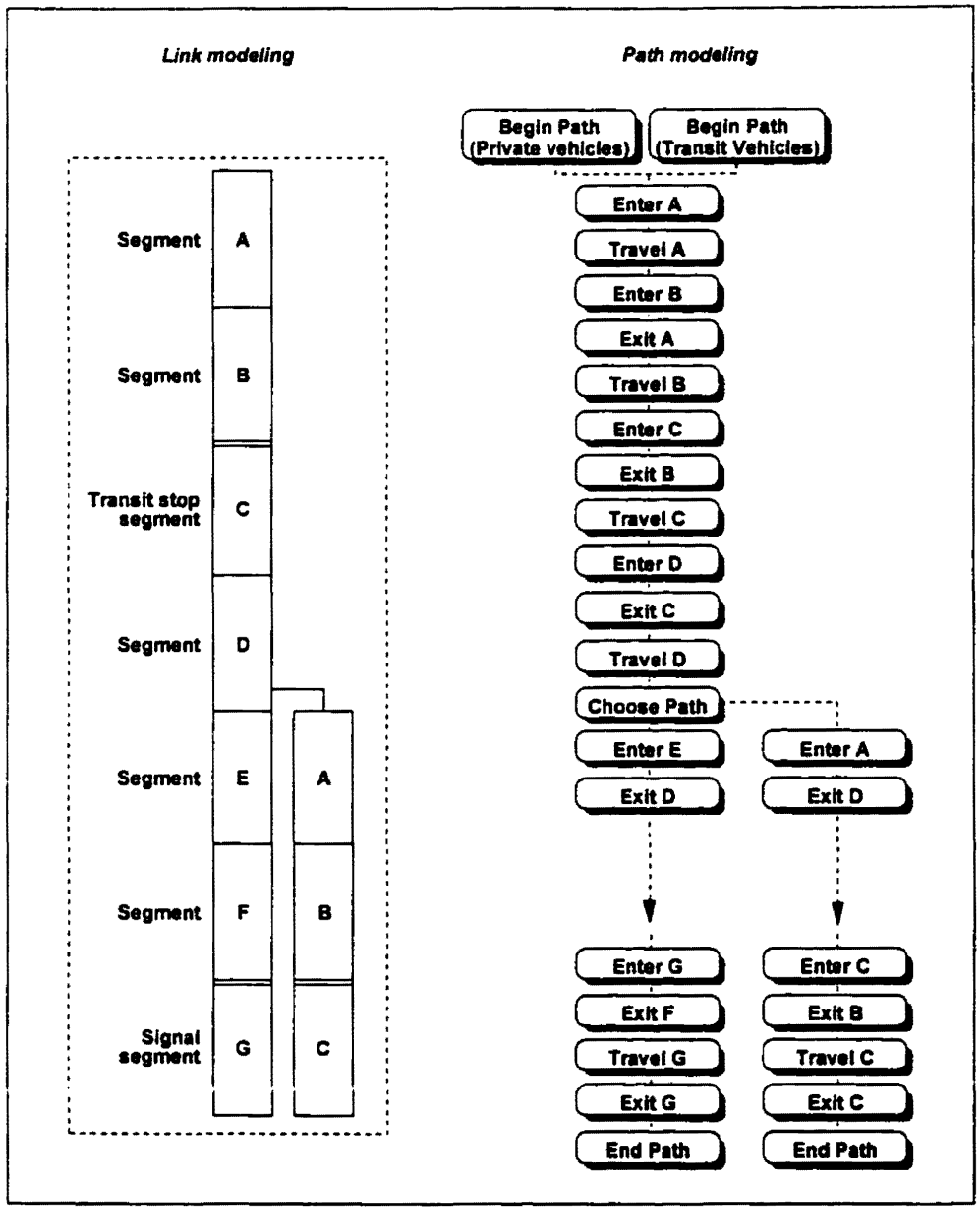


Figure 6.7 - Modeling of Vehicle Paths along an Intersection Approach

Enter, *Travel*, and *Exit* activities control how vehicles use roadway segments. Each time a vehicle engages an *Enter* activity, it attempts to engage the segment associated with that activity. If the segment is blocked, the vehicle is then forced to wait in queue at the downstream end of the upstream segment until the entering blockage is lifted. Each time a vehicle undertakes an *Exit* activity, it then disengages the segment it is currently using. Finally, *Travel* activities control the minimum amount of time a vehicle will remain within a given segment. This time is calculated in two stops. First, a speed value is sampled in the speed probability distribution associated with the segment. This probability

distribution is defined by the user when setting the simulation model and will usually reflect observed traffic behavior and conditions on the segment on which the vehicle is traveling. Once a speed value has been determined, the *Travel* activity determines the time required by the vehicle to travel non-stop along the full length of the segment at the sampled speed. As illustrated in Figure 6.7, the *Enter*, *Travel* and *Exit* activities are always arranged to ensure that any vehicle currently waiting to enter a given segment will continue to occupy the upstream segment until it can exit it. As it can be deduced, this process is another important element of the horizontal queuing process described in Figure 6.2.

Choose Path activities allow vehicles to continue along one of a number of alternative paths at specific points. The path chosen is again randomly selected on the basis of the user-defined probability of selecting each alternative. The only exception to this scheme is for transit vehicles, which are directed along specific paths on the basis of their assigned transit route number.

End Path activities mark the end of an activity process. These objects remove the vehicles from the idealized system once they have completed all the activities along their path. Consequently, these activities are used to mark the downstream boundary of a simulated system or to model any sink node within a given area.

Two other specialized activity objects have also been created by Conrad (1997) to specifically handle traffic detector inputs. These additional activities are:

- *Insert Detection (at Segment X)* - Insert new vehicles in the simulation model on the basis of information provided by a real-world traffic detector and allow the new vehicles to enter the specified segment.
- *Remove Detection (at Segment X)* - Remove from the simulation model vehicles projected from upstream detection stations.

To illustrate the use of these two objects, the example of Figure 6.8 depicts a vehicle process crossing two detector stations. The first detection station is located at the upstream end of segment A, while the second detection station is located at the upstream end of the segment D (currently, the simulation model only allows detectors to be placed at the entrance of a segment). At the beginning of the path, new vehicles are inserted in the simulation model at the upstream end of segment A on the basis of the traffic information provided by detector 1 using an *Insert Detection* activity object. Vehicles are generated on the basis on their vehicle type and introduced in the simulation model at the simulated

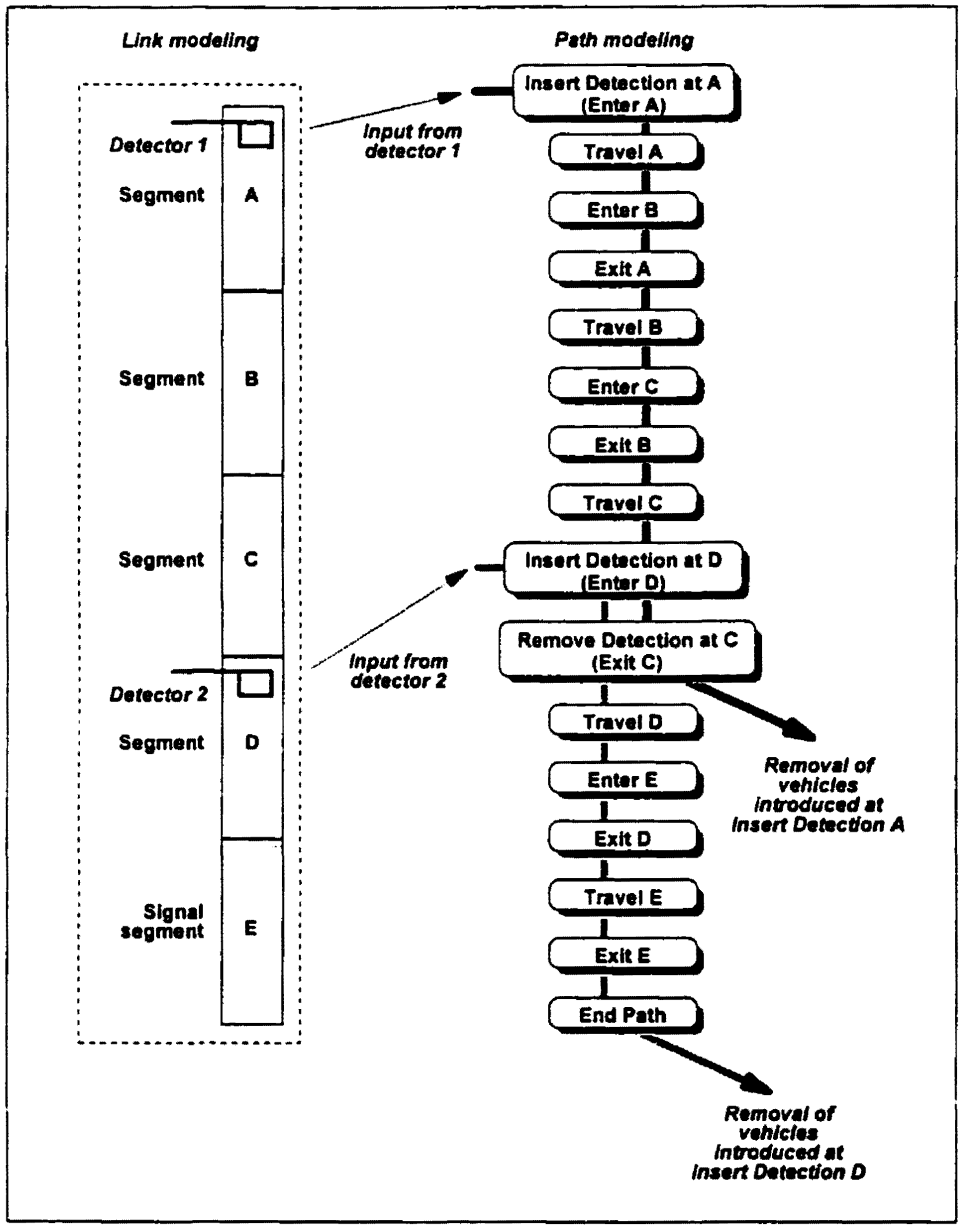


Figure 6.8 - Modeling of a Vehicle Path with Two Detector Inputs

time corresponding to their real-world detection time. Further downstream along the process, the second *Insert Detection* object introduces at the upstream end of the segment D the vehicles that are detected by the detector 2. At this point, since there is no purpose in continuing simulating the progression of vehicles introduced in the model at the first *Insert Detection* activity, these vehicles must now be removed from the simulation. The *Remove Detection at C* object accomplishes this task by removing from the simulation model all vehicles that have completed their travel along segment C.

As a result, all vehicles projected downstream of the detection station 2 will only be vehicles that have been detected at this station.

Insert Detection and *Remove Detection* objects are specialized *Enter* and *Exit* objects. When the simulation model is used to project future traffic conditions, for which case there is evidently no input from traffic detectors, these two objects perform exactly the same tasks as the *Enter* and *Exit* objects, respectively. In this case, the model is said to be in *projection mode*. When the simulation model handles traffic detection, such as when updating the estimated state of a real-world system, the tasks described in the previous paragraph are performed on top of the tasks normally performed by the *Enter* and *Exit* activities. In this case, the model is said to be in *tracking mode*.

In order to introduce in the simulation model detected vehicles that are compatible with the ones generated by *Begin Path* objects, an *Insert Detection* object must know the vehicle size and vehicle weight assigned by the user to the detected vehicles in addition to the detected time and vehicle type. In the test runs that will be described in Chapter 9, it will be assumed that this information could be provided by the traffic detection system. In a real-world system, however, it is very unlikely that the traffic detection system be able to provide such information. In such systems, the required information could be obtained by matching the detected vehicle type with pre-defined characteristics stored in look-up tables.

6.2.5. Signal Controllers

Signal controllers model the operation of traffic signals at individual intersections. These objects do not optimize the signal timings. They simply control the operation of signal timings. Their primary role is to impose signal changes in intersection approaches according to the phase plan in operation at each intersection.

In order to realistically model the operation of real-world traffic signal controllers, it is currently required that the user provides the following information for each signal controller defined within the idealized system:

- The number of phases available for implementation at the intersection under control;
- A phasing table indicating the links on which signal heads display a green signal indication during each possible phase;
- An intergreen period matrix indicating for each phase the duration of time separating the

end of the displayed green interval from the beginning of the next conflicting displayed green interval (the intergreen period includes the amber interval and any all-red period between two conflicting green intervals);

- A table indicating which phase can follow each defined phase;
- A table defining the minimum duration of each phase;
- A table defining the maximum duration of each phase;
- A start lag table indicating for each phase the amount of time lost at the beginning of each green interval (see Figure 6.4);
- A stop lag table, indicating for each phase the time elapsing between the end of the green signal indication and the end of the effective green interval (see Figure 6.4).

During the course of a simulation, signal controllers implement signal phases on the basis of the phase plan stored in their memory. The phases to implement are determined using two different methods depending on the type of signal control performed. If the signal controller is operated in fixed-time, the phases are determined on the basis of the cycle time, green allocation and signal offset information provided by the user. If the controller is operated in real-time, the phases are determined by a chronological sequence of switching decisions stored within the signal controller. This sequence of switching decision is generated by the signal optimizer attached to the signal controller and indicates which phase to implement at which time.

In order to keep track of past decisions, the simulation model has been modified in the course of the current research project to allow individual signal controllers to keep a record of the ten previous signal switches that they have implemented. No such records were kept in the previous version of the simulation model, resulting in an inability to interrogate signal controllers about their previous decisions. As it will be seen in Chapter 8, this ability to obtain information about previous signal-switching decisions at individual intersections will provide valuable information when attempting to establish efficient signal coordination patterns between adjacent intersections. It will also allow signal controllers to evaluate the current average duration of the various phases that they have implemented in a recent past. Such information will notably be valuable in predicting future signal displays when projecting future traffic demands or attempting signal coordination.

To summarize, this chapter has presented an overview of the simulation model used by SPPORT. As indicated in the introduction of the chapter, this model plays a vital role in the signal optimization

process as it allows future traffic conditions to be projected and candidate signal timing strategies to be evaluated. The network modeling associated with the simulation model also determine the type of information describing traffic conditions that can be obtained by SPPORT and imposes in that way some constraint on the development of the signal optimization process. The next chapter describes a first application of the simulation model by presenting how the model is used to estimate and predict the demand placed on each approach to a signalized intersection. The following chapter, Chapter 8, will then describe how this information obtained through simulation is used to generate signal-timing strategies.

7. Traffic Demand Estimation

For real-time, traffic responsive signal control logic to be effective, it must have an accurate view of the state of traffic conditions in the network under control and be able to predict, at least over short periods of time, how these conditions will evolve. While the previous chapter described the components of the simulation model used by SPPORT to predict future traffic conditions, this chapter presents how the model is used to make such predictions. The chapter describes in detail the method developed for estimating the expected traffic demand on each intersection approach. It begins with a description of the process allowing individual controllers to keep track of the existing traffic conditions around the intersection they each control. This is followed by a description of the data exchange process allowing adjacent controllers to exchange valuable information about future vehicle arrivals, and by a presentation of how stop line arrival patterns are estimated. Finally, the last section describes the algorithm used to identify platoons within the projected arrival patterns

7.1. Tracking of Current Traffic Conditions

Before being able to project any future traffic conditions, individual signal controllers must obtain knowledge of the traffic conditions existing around the intersection they each control. As illustrated in Figure 7.1, estimates of current conditions are obtained through simulation by continuously updating the status of an idealized system representing the intersection under control using detector data and signal change information from the real-world.

In the above process, the use of simulation allows the demand estimation process to simultaneously consider a variety of factors that would be otherwise difficult to combine in analytical evaluations. For example, simulation allows the combined effects of transit interference, traffic signal operation and queuing process to be simultaneously considered in an efficient way. However, the use of simulation to estimate traffic demand estimates that are later used to generate signal timing decisions for a real-world system also raises some concerns about the validity and accuracy of the estimations with respect to real-world traffic demands.

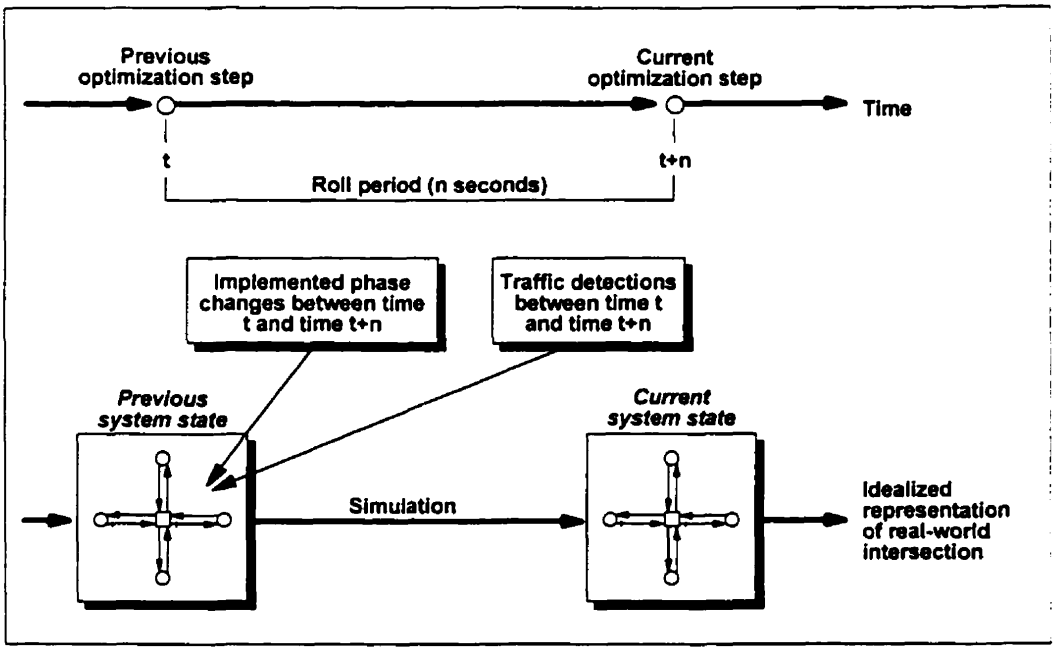


Figure 7.1 - Current System State Evaluation Process at Individual Intersections

Yagar and Han (1994) raised one major concern regarding the effects of inadequately simulating the speed at which vehicles travel along intersection approaches. As explained in Section 6.2.4, the simulation model only projects vehicles from one detection station to the next when operating in *tracking mode* (see Figure 6.7 in previous chapter). In this mode of operation, any vehicle that is projected to have reached a downstream detection station is removed from the simulation model on the assumption that this vehicle will be replaced by a new one as soon as the corresponding real-world vehicle will pass over the traffic detector. The problem with this simulation scheme is that it may cause vehicles to be temporarily missing or double counted. For example, a vehicle may be double counted if its assumed speed is less than its real-world speed. In this case, a new vehicle will be generated by the *Insert Detection* activity object associated with the downstream detector and introduced in the simulation model before the vehicle generated at the upstream detection station will be removed from the simulation. If the assumed speed is greater than the real-world speed, a vehicle will then be removed from the simulation before its replacement will be generated, therefore causing that vehicle to be temporarily missing from the simulation.

The above problem only occurs when the simulation model is in *tracking mode*. When operating in *projection mode*, the simulation model ignores all detection stations. As a result, *Insert Detection* and *Remove* Detection activity objects are treated as regular *Enter* and *Exit* objects and no vehicles are introduced or removed from the simulation at these points. Despite the problem mentioned above, the

problem of missed and double counted vehicles when simulating in *tracking mode* is considered to be acceptable for passenger cars, as these vehicles are always considered collectively in terms of queues and platoons in the signal optimization process. For passenger cars, the impact of considering one less or one more vehicle should not significantly affect the optimization.

However, such errors are not acceptable for transit vehicles, as these vehicles are always treated individually. For this reason, transit vehicles are treated somewhat differently than passenger cars in the simulation model. Contrary to all other vehicles, transit vehicles are only removed from the simulation when a downstream real-world detector has actually detected them. As a result, double counting may still occur, but transit vehicles will never be missing from the simulation. The match between simulated and detected vehicle can be done by comparing the route number of transit vehicles or the identification code assigned to the transit vehicle at the time of its generation in the model (see Section 6.2.1). However, such matching process can only be achieved if the required information can be obtained from the selective detection system. For example, unique identification code can be obtained from the use of transponders mounted on board on each vehicle. If this information cannot be obtained, there will then be no way to verify the identity of specific vehicles. The same matching process could theoretically be used for passenger cars, but major problems prevent its practical application. First, to uniquely identify each vehicle a very sophisticated detection system must be used. In addition, powerful computers would have to be used to perform the complex task of matching in real-time the identity of hundreds of vehicles during the course of a simulation. At this time, the cost and risks of false detection associated with such systems are too great to treat passenger cars similarly to transit vehicles.

The ability of the SPPORT model to accurately follow real-world traffic conditions depends heavily on how well the simulation model it uses is calibrated to the system under control. As with any other simulation model, the user is required to provide the simulation model with a number of parameters describing the geometric layout of controlled intersections and the behavior of drivers at these intersections. As a result, any value misrepresenting reality may have significant effects on the simulation, especially when simulations are run for long periods of time without calibration.

Even if all simulation parameters are reasonably well determined, there is always a risk that discrepancies may slowly be introduced between the simulation model and the real world. Being idealizations of real-world systems, no simulation model can perfectly replicate real-world traffic behavior. The most damaging effect of calibration errors is probably the simulation of queues being

significantly longer or smaller than the existing ones. In order to minimize the impact of these errors, algorithms can be developed to allow simulation models to correct the size of queues they estimate on the basis of information collected by traffic detectors. For example, similar to what is currently done in SCOOT and PRODYN, queue sizes could be corrected in the simulation model each time the tail of a real-world queue would reach a detection station. Queue sizes could also be corrected if a simulated queue is projected to extend over a detector that does not currently report a real-world queue.

A first step in this direction was taken by Conrad (1997), who introduced algorithms in the simulation model that impose gate factors when queues are detected to spill across a detector and remove these factors when the real-world queues disappear. However, these algorithms are not currently designed to handle cases in which real-world queues are smaller than simulated ones, as they only impose traffic gates and do not cause any change in the number of vehicles being simulated. The major problem is to determine if there is too many vehicles in the system and how to remove those extra vehicles. Changing queue sizes is relatively easy in simulation processes that only keep track of the number of queued vehicles. However, queue sizes are obtained within SPPORT by analyzing the status of individual vehicles. As a result, queue sizes can only be changed by modifying the status of individual vehicles, which cannot easily be done given the complex interaction between all simulated vehicles.

7.2. Estimating of Departures from Upstream Intersections

In Section 5.3, it was indicated that SPPORT promotes signal coordination between adjacent signalized intersections by instructing each signal controllers to exchange pertinent signal and traffic information with its immediate neighbors. As it was illustrated in Figure 5.3, each controller receives a list of projected signal-switching decisions that its upstream and downstream neighbors intend to implement in the near future, as well as a list of projected vehicle departures from the intersection located at the upstream end of each approach. This information allows each signal controller to extend the horizon over which arrivals can be projected, permitting the controllers to better plan their signal timing strategy as a function of what adjacent signal controllers are expected to do.

Figure 7.2 illustrates the process by which estimated vehicle departures from one intersection are used in the demand estimation process of a downstream intersection. In the first decision horizon, traffic detector data are collected at station A and used by signal controller A to generate new projected signal timings for the current decision horizon. After the signal optimization, the controller is instructed to simulate traffic departures from the intersection it controlled on the basis of its new signal strategy.

During the course of this simulation, traffic arrivals are recorded at detection station B. This information is then passed down to signal controller B, which adds the projected departures to the data previously received from detector B before using these data to produce an estimate of the near-future arrival patterns at the stop line of the second intersection.

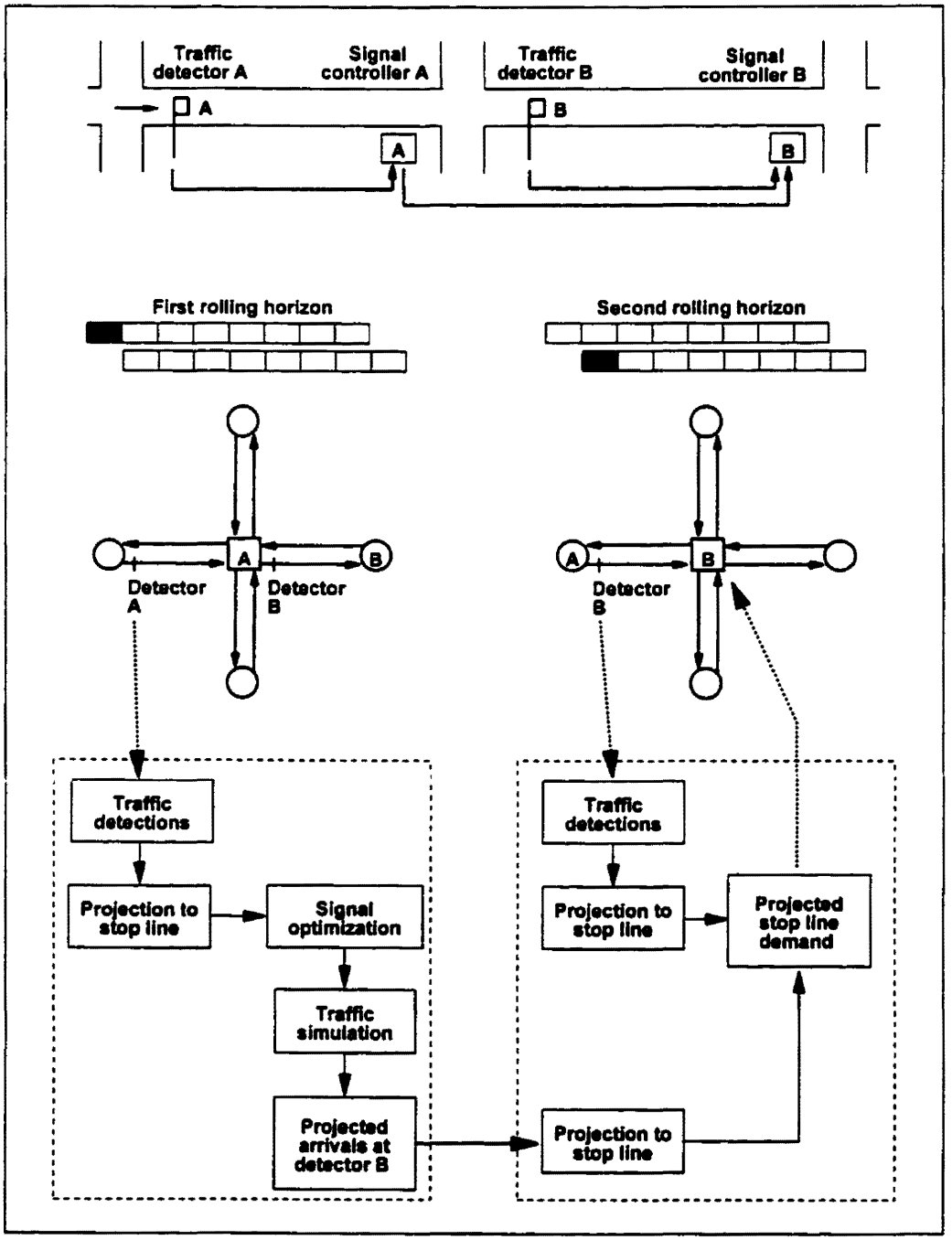


Figure 7.2 - Generation of Projected Stop Line Arrivals using Traffic Detector Information and Projected Departures from Upstream Signalized Intersection

In the above process, vehicle arrival times at detection station B instead of stop line departure times from intersection A are passed to signal controller B in order to keep the demand estimation process as simple as possible. If stop line departure times from intersection A were passed to signal controller B, this controller would have to simulate the effects of turning movements within intersection A in order to determine which vehicles would reach the intersection it controls. This would result in a duplication of simulation tasks, since both signal controllers A and B would simulate traffic movements across intersection A. Signal controller B would simulate traffic movements to estimate future arrivals at intersection B, while signal controller A would simulate the same movements to evaluate the signal control strategy it had just developed.

As indicated in Section 6.2.4, the *Insert Detection* activity object has been developed to allow vehicles to be introduced in the simulation model used by each signal optimizer on the basis on traffic detector inputs. Instructing each simulation model to record departures from a given intersection at the location of a real-world detection station along the exits from that intersection thus offer the added benefit of being able to use the same *Insert Detection* object to introduce in the simulation model the projected vehicle departures provided by adjacent signal controllers. In more practical terms, this treatment allows the simulation model to treat the projected vehicle arrivals as if they were real-world detector inputs.

In order to provide the means of recording projected arrivals at a detection station, an option has been built in the *Insert Detection* object to allow it to create on demand a record of all vehicles engaging it during the course of a simulation. The effect of this option is illustrated in the example of Figure 7.3. In this example, the *Insert Detection* object associated with detector 10501 is instructed to maintain a record of all the vehicles engaging it from the *Travel A* activity object, while the *Insert Detection* object associated with detector 10502 is instructed not to do so. In order to keep full compatibility with the data required by the *Insert Detection* objects of the simulation models receiving the recorded data, the following information is recorded each time a vehicle engages the object: the simulation clock time, the type of vehicle, the size of vehicle in passenger car units, and the weight assigned to the vehicle by the user. As indicated in Section 6.2.4, it can be assumed that traffic detection systems can provide the detection time and vehicle type, while the other information can be obtained by matching the characteristics of the detected vehicles to information stored in look-up tables.

To ensure the correct correlation between detection stations when transferring data from one simulation model to another, a unique identification number is assigned to each real-world detector. As a result,

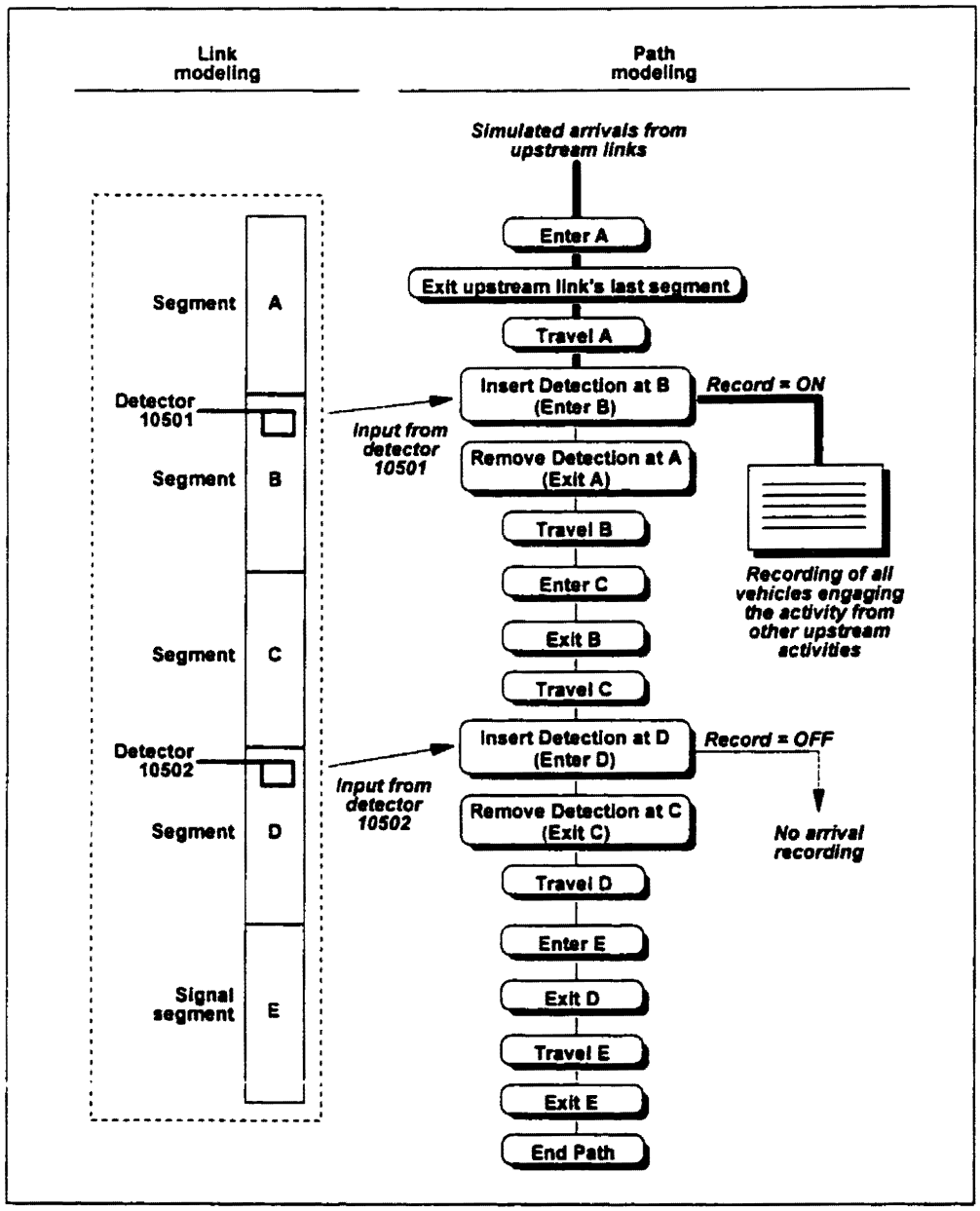


Figure 7.3 - Modeling of a Vehicle Path with a Traffic Arrival Record Point

the same real-world detection station always has the same identification number in all the simulation models in which it is modeled. This allows individual signal controllers to search for specific arrival lists by successively interrogating adjacent controllers and asking each one of them whether they have the list recorded at the requested detection station.

At this point, it must be noted that the arrival lists recorded by *Insert Detection* objects is kept until the beginning of the next optimization step, at which time it is cleared to allow the recording of a new

projected list. This means that data exchanges between adjacent controllers must be performed at the end of an optimization step, when all signal controllers have finished generating new signal timings and are waiting to start their next optimization.

7.3. Estimating Stop line Traffic Demand

Signal optimizers need to know the temporal distribution of traffic demand on each approach to the intersection they each control in order to generate optimum signal switching decisions. This temporal distribution is estimated in SPPORT by:

- Maintaining a record of the times at which vehicles departed the intersection from each approach link over the previous 10 seconds.
- Projecting the arrival time of individual vehicles at the stop line of each approach link over the entire duration of the decision horizon (see Figure 7.4), assuming that the signal head displays on each link will be returned to green as soon as possible and then kept in that state indefinitely. As it will be explained below, this information, together with the record of vehicle departures in the previous 10 seconds, is used to evaluate when platoons are scheduled to reach the controlled intersection on each approach and to determine whether queues are dissipating at an acceptable rate.
- Recording the size of the stop line queue at the beginning of the look-ahead period, to allow signal optimizers to determine when it is time to start serving a growing queue.
- Recording the projected location of the most upstream tail of queue at the beginning of each decision point within the look-ahead period, to allow signal optimizers to monitor queue spillback conditions over the look-ahead period on both intersection approach and exit links.

The look-ahead period is the period over which future traffic conditions are simulated in order to generate a temporal picture of the demand placed on the intersection under control. As shown in Figure 7.4, this period does not necessarily correspond to the decision horizon, which is defined as the period over which signal-switching decisions are made. Instead, it can be seen that there is a distinct look-ahead period for each decision point within the decision horizon. For each decision point, the look-ahead period starts at the time at which a signal-switching decision is required and extends for a user-defined fixed number of seconds into the future. As an example, many of the test runs that were performed during the current research project used a 60-second decision horizon, decision points five seconds apart, and a 45-second look-ahead period.

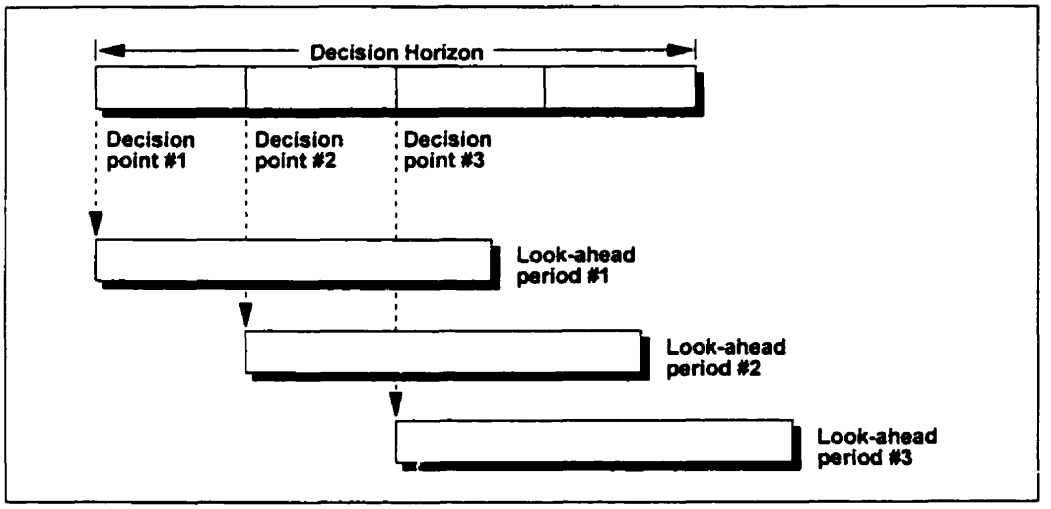


Figure 7.4 - Relation between Decision Horizon and Look-Ahead Periods

Over the look-ahead period, it is assumed that the projected signal-switching decisions at adjacent intersections are kept unchanged, as different signal controllers determine these decisions. Projected stop line vehicle arrivals at the intersection under control are then determined assuming that all signal head displays can be returned to green as soon as possible on every approach link. As a result, on approaches on which a green signal indication is displayed at the beginning of the look-ahead period, the indication is kept the same for the whole duration of the look-ahead period. On approaches with a red signal indication, it is assumed on the other hand that the signal displays are returned to green as soon as minimum green interval and phase sequence restrictions allow it.

The above simulation scheme results in all the signal displays at the controlled intersection being green at the same time. The basis for this simulation is to obtain an estimate of the temporal distribution of the demand placed on the intersection that do not reflect the projected signal timing decisions made during previous optimization. If projected signal-switching decisions from previous optimizations were included in the simulation, the resulting stop line departure patterns would then heavily reflect past projected decisions. By simulating an "all-green" situation, the effects of the signal timings on traffic demand patterns can therefore be removed from the demand estimation process.

In this process, the minimum remaining red intervals must still be simulated for the approaches on which a red signal is currently displayed to reflect the fact that signal changes cannot be made instantaneously. While it is generally desirable to remove the effects of possible future signal timings on the demand estimation process, it is also desirable to maintain those signal effects from commitments made before the start of the current optimization and that can therefore not be changed.

For example, there is no point in simulating immediate departures from an approach on which the signal indication was changed to red five seconds ago if phasing constraints in the real-world system prevent the signal indication to be returned to green on that approach within the next 20 seconds.

To simulate unrestricted access to the intersection being optimized, it is also assumed that all left-turning vehicles are not required to seek the availability of appropriate gaps in opposing traffic streams before being allowed to enter the intersection. In the initial versions of the SPPORT model (Yagar, Han and Greenough, 1991, 1992; Han and Yagar, 1991, 1992; Yagar, Han and Wang, 1992; Yagar et Han, 1994; Conrad, 1997; Conrad et al., 1998), there were no facilities for simulating the gap acceptance logic process followed by left-turning vehicles. As a result, all traffic demands were estimated assuming the utilization of protected phasing.

As indicated in Section 6.2.2, the logic according to which vehicles can seek the availability of appropriate gaps in opposing streams was added in the simulation model used by SPPORT as part of the current research effort. However, it has been found that its integration with the current demand estimation process presents some major problems:

- First, as a result of the design assumption that all signal displays should be returned to green as soon as phasing restrictions allow it on all approaches, protected phasing demands cannot be properly evaluated if the gap acceptance logic is kept active. In this case, vehicles would be simulated to wait for the availability of appropriate gaps while they would not do so in the real world. Consequently, simulated vehicle departures could be more spaced in time than real-world departures, which could then wrongly lead the signal optimizer to assume that something interferes with the discharge of vehicles during the protected phase and that it may be better to switch to another phase.
- Second, since SPPORT is being designed with the capability of choosing which phase should follow the current one, the use of the gap acceptance logic creates a dilemma about how to evaluate traffic demands on intersection approaches on which both permissive and protected phasing can be used. A typical example of such approach is an approach on which a flashing green signal indication can be used in conjunction with a regular circular green indication. One way of correctly estimating the demand for both protected and permitted phasing would be to perform two simulations: one with the gap acceptance logic on, and one with the logic off. While such approach is currently feasible, it would greatly increases the time needed to perform an optimization and would consequently reduce the real-time applicability of the SPPORT model. It would also not solve the problem of

determining when to end the protected flashing green display and replace it with a regular permissive green signal indication.

- Finally, the introduction of the gap acceptance logic in the demand estimation process creates the problem of correctly estimating the time required to dissipate a queue, as such information plays an important role in the process by which the signal timings of adjacent intersections are coordinated. For example, during a protected left turn movement, the rate at which vehicles cross an intersection is essentially a function of the flow capacity of the intersection approach and of the receiving downstream link. For permitted movements, the discharge rate is also a function of the traffic patterns in the opposing direction. While it is possible to use procedures described in the Canadian Capacity Guide for Signalized Intersections (ITE, 1995) and the Highway Capacity Manual (TRB, 1994) to estimate the maximum possible discharge rate for the left-turning vehicles, the results will still be a function of the accuracy of way the simulation model represents real-world traffic.

It is also important to note that the duration of the look-ahead period should be sufficiently long to allow the effects of the signal-switching decisions taken at the beginning of the period to be properly evaluated. For example, a decision to change a signal display from red to green may be taken at a given time solely on the basis of the estimated queue sizes on the various intersection approaches at that time. While such a decision may satisfy current needs, it does not consider the potential future impacts of the decision on other traffic. For instance, switching the signal displays now may satisfy the needs of existing queues but may also cause large platoons of vehicles to be stopped on another approach in 10 or 20 seconds. The switch may also prevent future incoming transit vehicles from getting priority of passage as a result of signal phasing constraints. Consequently, the look-ahead period should be long enough to allow traffic conditions on all approaches to be evaluated up to at least the next possible green interval. In addition, the period should also be long enough to allow future incoming platoons to be identified on approaches for which uninterrupted platoon progression is desirable. Theoretically speaking, the longer the look-ahead period is, the better the future impacts of current switching decisions can be evaluated.

From a practical point of view, however, the duration of the look-ahead period is constrained by the fact that the overall accuracy of traffic projections decreases with every attempt to move further into the future. The duration of the look-ahead period is also constrained at each intersection by the amount of advance information about future arrivals available on each approach. For example, consider an intersection on which advance arrival information on each approach is available for the next 30, 25, 36

and 20 seconds, respectively. In this case, the duration of the look-ahead period associated with the first decision point of the decision horizon should theoretically not exceed 20 seconds, as traffic information is not available for all approaches past this point. If periods longer than 20 seconds are used, a bias could be introduced in the demand estimation process in the fact that no demand may be assumed to exist in the later part of the look-ahead period for the approaches for which advance traffic information is not available. Similarly, for the second decision point of the decision horizon, which is assumed to occur five seconds later, the duration of the look-ahead period should not exceed 15 seconds. Using the same logic, the look-ahead period for the third and fourth decision points should also be constrained to ten and five seconds, respectively.

The above constraint lead to a situation in which the duration of the decision horizon could be constrained at an intersection by the amount of advance traffic information available, which is not desirable. To solve this problem, artificial traffic generators can be used in the simulation model to fill in data for the periods over which no detector information is available. This approach is illustrated in Figure 7.5. For each approach for which insufficient information is available, a *Begin Path* activity object can be inserted in the simulation model to generate vehicle arrivals at the upstream end of the approach on the basis of a user-defined probability distribution of inter-arrival times (see Section 6.2.4). Within the period for which future arrivals are known, any artificially generated traffic would then

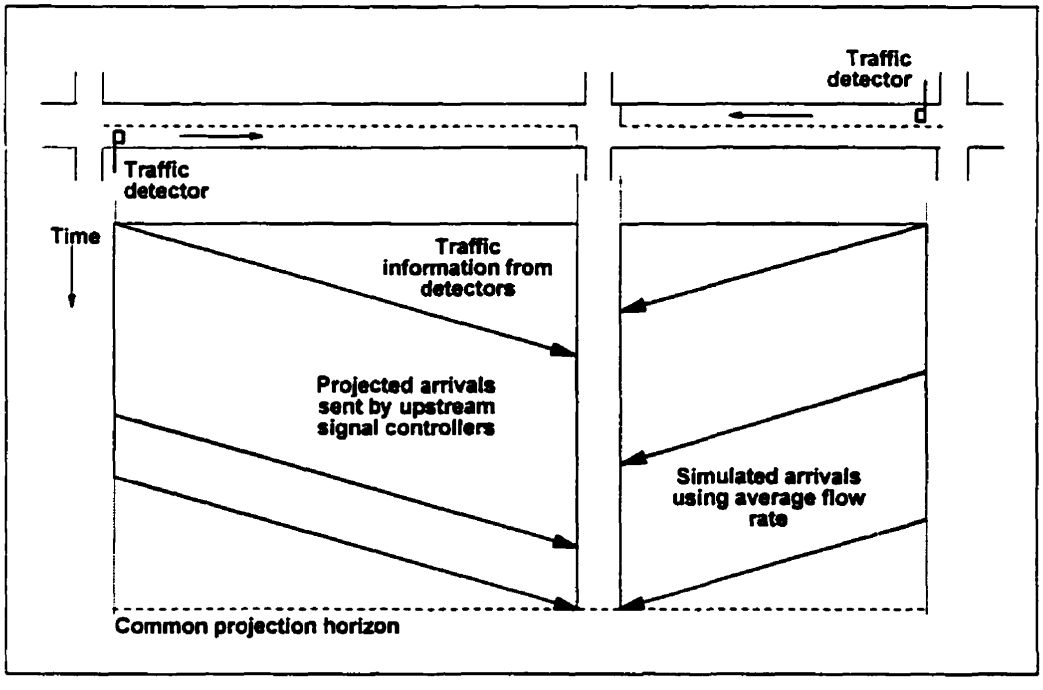


Figure 7.5 - Use of Artificially Generated Arrivals within the Demand Estimation Process

automatically be replaced by detected traffic when the generated vehicles would reach an *Insert Detection / Remove Detection* activity object combination (see Figure 6.8). As a result, near future stop line arrivals on each approach would reflect the information provided by the traffic detectors located on the intersection approach and by adjacent upstream signal controllers, while arrivals in the later part of the look-ahead period would reflect a continuous arrival process based on some observed probability distribution of inter-arrival times.

In this case, while the artificially generated arrivals might not truly reflect future arrival patterns, they at least allow the signal optimizers to consider that vehicle arrivals may still occur on each approach past the point for which real-world traffic information becomes unavailable. These artificially generated arrivals may also allow the signal optimizers to plan signal switches over longer decision horizons. In such scenario, however, any signal switches scheduled to occur in the later part of the decision horizon should only be considered as a first approximation of the intended future signal control strategy, as traffic projections associated with the last few decision points of the decision horizon might solely rely on artificially generated traffic information (see Figure 7.6). In this case, the rolling horizon optimization process will compensate for any inaccuracy by allowing the signal optimizers to revise their traffic signal control strategy at regular intervals as updated traffic information is obtained from traffic detectors and adjacent traffic signal controllers.

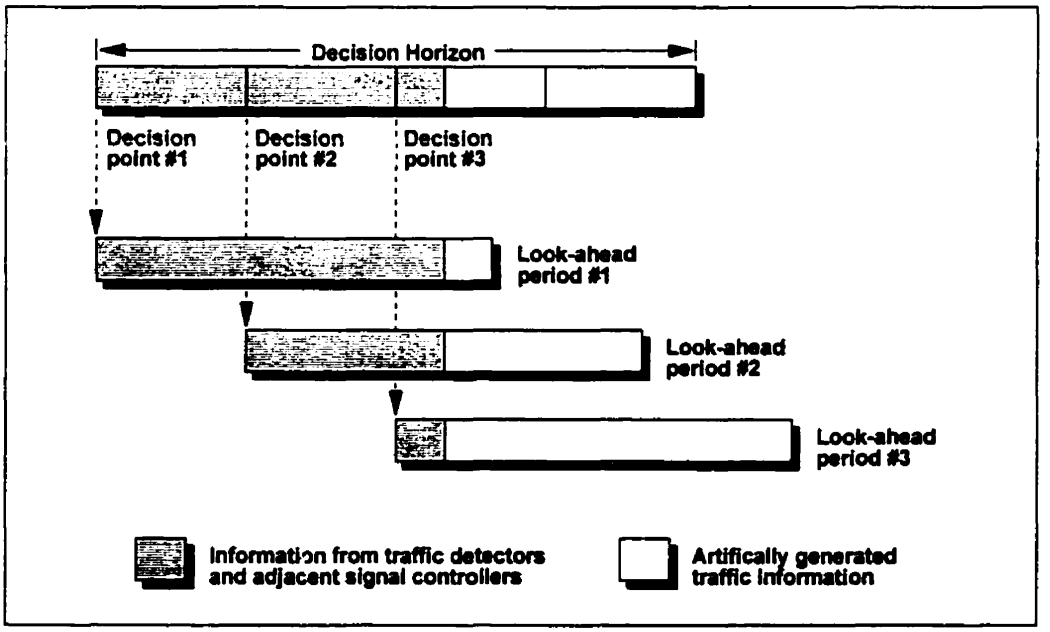


Figure 7.6 - Source of Traffic Information within Decision Horizon and Look-Ahead Periods

At any given time, it is important that the demand be correctly estimated in the first part of the look-ahead period, as this demand may cause a signal switch to be implemented in the real world in the next few seconds. Since many events may still affect future arrival patterns, demands in the later part of the look-ahead period are not subject to the same accuracy, even if the vehicles projected to arrive further in time are given the same weight in the optimization process as the vehicles projected to arrive imminently. In this case, if projected traffic demands are reevaluated every five to ten seconds, as it is anticipated to be done within the SPPORT model, there will be ample time to correct any mistakes initially made by the signal optimizers. For instance, consider a signal-switching decision made at the last decision point of a 60-second decision horizon. Also, consider that 5-second decision intervals are used within the decision horizon. In this case, the switching decision will be reevaluated 12 times before being implemented in the real world, leaving enough opportunities for the signal optimizer to correct any mistake it may have initially made. This important aspect of the signal control system will be examined in more detail in Chapter 9, where a sensitivity analysis will evaluate the impacts on the performance of the SPPORT model of using decision horizons of different duration while maintaining constant decision intervals within the horizon.

7.4. Platoon Identification Algorithm

On flow facilities controlled by traffic signals, traffic engineers must deal with the constant stopping and restarting of traffic streams caused by the operation of traffic signals. As a result of these periodic interruptions of flow, traffic tends to occur in platoons, where a platoon is defined as a group of relatively fast moving vehicles traveling closely together along a facility. When such traffic conditions exist, there are typically significant potential benefits in attempting to align the green signal display intervals at each intersection along a given route with the scheduled platoon arrival times at these intersections, that is, in attempting to coordinate the operation of signalized intersections.

In the SPPORT model, platoons of vehicles are identified by examining the projected stop line arrival patterns produced by the "all-green" simulation process described in the previous section. More specifically, platoons are identified by analyzing the headway, or time interval, between successive arrivals on a given approach. The headway of each vehicle in the traffic stream is examined in turn, beginning with the first projected arrival, and then by considering successive arrivals. The downstream end, or front, of a platoon is associated with the first headway that is less than some critical headway. Each vehicle with a headway less than the critical headway is then associated with the platoon, until a headway greater than the critical headway is found, at which point the upstream end of the platoon is

defined. At the end of the process, a platoon is only declared if the number of vehicles found to travel closely equals of exceeds the user-defined minimum number of vehicles assumes to constitute a platoon. This number is based on engineering judgement and may reflect the various preferences of the engineers in charge of the traffic signal control system. For example, in the various tests that will be conducted in Chapter 9, a minimum of 10 vehicles will be assumed to constitute a platoon. This number was obtained through an evaluation process in which the effects of using various platoon thresholds in the signal optimization process were analyzed.

In the current platoon identification process, two vehicles are said to be following each other closely and to belong in the same platoon if the following criteria are simultaneously met (see Figure 7.7):

$$h_i \leq h_{crit} + h_{res\,m} \qquad [7.1]$$

$$h_i \leq 2 * h_{avg} \qquad [7.2]$$

where: h_i = Time headway between the i^{th} and i^{th-1} vehicles (seconds).

h_{avg} = Average headway of vehicles in platoon (seconds).

h_{crit} = Critical headway (seconds).

$h_{res\,m}$ = Current reserve headway on segment m (seconds).

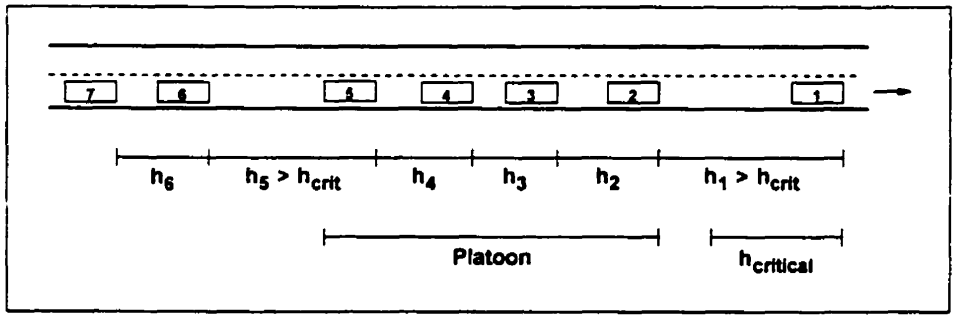


Figure 7.7 - Notions of Headway and Critical Headway in Platoon Identification Algorithm

In Equation 7.1, the critical headway parameter represents the maximum allowed headway between successive vehicles in a platoon. This time parameter must be defined by the user for each link in the simulation model and is used as a threshold to determine where a platoon starts and where a platoon ends. For example, a critical headway of three seconds indicates that two vehicles following each other with a headway of two seconds can be assumed to belong to the same platoon, but not two vehicles following each other with a headway of four seconds.

The reserve headway parameter, also found in Equation 7.1, allows the platoon identification algorithm to consider the natural randomness of vehicle arrivals when determining if two vehicles following each other should be grouped or not. This parameter, which was introduced in the SPPORT model by Conrad (1997), permits the search algorithm to include in a platoon a vehicle that follows the last registered vehicle of the platoon with a headway larger than the critical one. During the course of a given search, the value of this parameter is adjusted as follows:

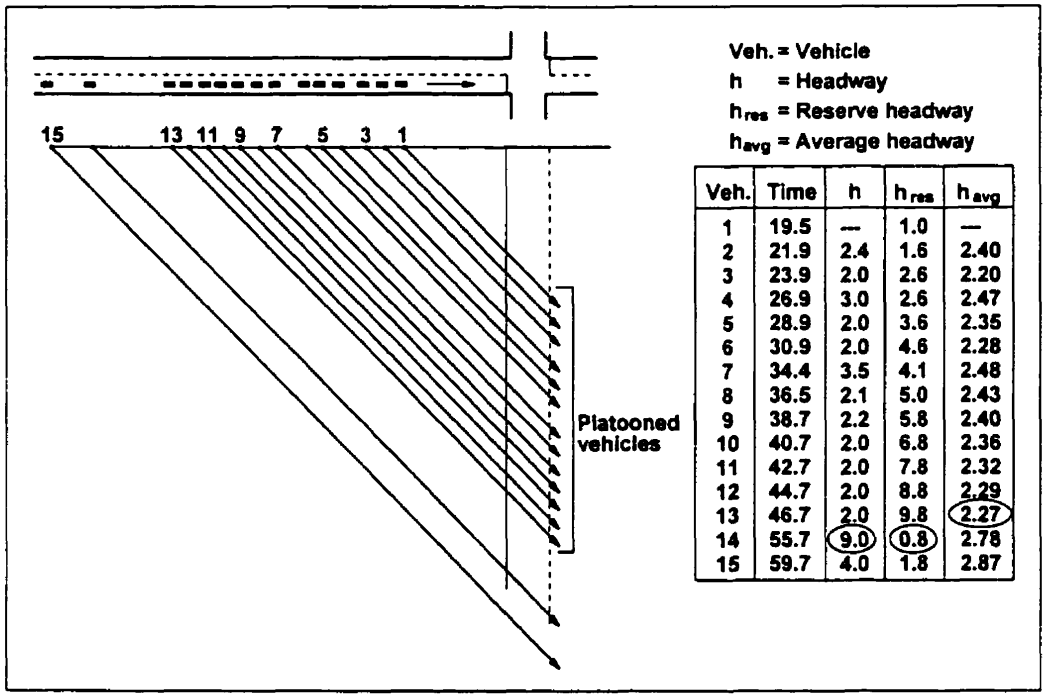
- When the search begins, a reserve headway of one second is arbitrary allowed. This means that any two vehicles following each other with a headway exceeding the critical headway by no more than one second will be assumed to be close enough to be grouped.
- Each time a new vehicle is found to be close enough from its predecessor for being included in the same platoon, the value of the parameter is then increased, or decreased, depending on how close the vehicle follows its predecessor. Equation 7.3 is used to adjust the value of the reserve headway on each segment:

$$(h_{res\,m})_{n+1} = max\left[0\,,(h_{res\,m})_n + t_{m\,i+1} - t_{m\,i}\right] \qquad [7.3]$$

where: $(h_{res\,m})_i$ = Reserve headway on segment m after arrival of i^{th} vehicle (seconds).
$t_{m\,i}$ = Time vehicle i entered segment m at its upstream (seconds).

- Once the reserve headway reaches a value of 0, vehicles can only be added to a group of vehicles if their headway with the previously observed vehicle does not exceed the critical headway. In this case, each a vehicle is added to the platoon, the value of the reserve headway is still adjusted using Equation 7.3.
- When a vehicle is found to travel too far behind its predecessor to be included in the same platoon, the value of the reserve headway parameter is finally arbitrarily reset to one before continuing the search with the remaining vehicles.

While the reserve headway criterion allows the platoon search algorithm to account for some of the natural randomness of vehicle arrivals, it may also allow vehicles that are relatively far apart to be wrongly grouped. This is particularly true when dealing with large platoons. For example, consider Figure 7.8 in which 10 vehicles are following each other very closely with headways varying between 2.0 and 3.5 seconds and a critical headway of three seconds. As expected, the reserve headway parameter allowed the seventh vehicle to be included in the platoon on the basis that enough reserve headway has been cumulated to allow the vehicle to be considered in the platoon despite the fact that its headway is greater than the critical headway. The same phenomenon is again observed when the



Veh.	Time	h	h_{res}	h_{avg}
1	19.5	—	1.0	—
2	21.9	2.4	1.6	2.40
3	23.9	2.0	2.6	2.20
4	26.9	3.0	2.6	2.47
5	28.9	2.0	3.6	2.35
6	30.9	2.0	4.6	2.28
7	34.4	3.5	4.1	2.48
8	36.5	2.1	5.0	2.43
9	38.7	2.2	5.8	2.40
10	40.7	2.0	6.8	2.36
11	42.7	2.0	7.8	2.32
12	44.7	2.0	8.8	2.29
13	46.7	2.0	9.8	2.27
14	55.7	9.0	0.8	2.78
15	59.7	4.0	1.8	2.87

Figure 7.8 - Platoon Identification Example

fifteenth vehicle arrives. However, this time the observed headway between this vehicle and the fourteenth one is 9.0 seconds, which is too big in practice to consider these vehicles grouped. At the end, it can be observed that the algorithm has successfully identified all the vehicles in the platoon.

To solve the above problem and improve the platoon identification algorithm, Equation 7.2 has been introduced in the platoon search algorithm as part of this research project. This equation states that a vehicles following another one can only belong to the same platoon as the leading vehicle if the interval separating the two vehicles does not exceed twice the value of the average headway between the vehicles in the platoon to which the leading vehicle belongs to. As shown in the example of Figure 7.8, the use of this criterion results in the realistic conclusion that the platoon being investigated ends immediately after the fourteenth vehicle.

The above procedure for the identification of incoming platoons, together with the other demand estimation procedures described in this chapter, will play an important role in the signal optimization process follow by SPPORT. While Chapter 5 briefly presented the main conceptual elements of the SPPORT model, the next chapter, Chapter 8, will describe in details how the demand estimation procedures detailed in this chapter and using the simulation model described in Chapter 6 will be used to generate near-optimal signal-switching decisions.

8. Signal Optimization

As described in Chapter 5, the SPPORT model relies on the use of a heuristic rule-based decision making process to make near-optimal signal-switching decisions and generate signal timing plans. This signal optimization procedure is based on the recognition that signal-switching decisions usually occur after the realization of specific traffic events in traffic responsive signal control environments. This chapter presents this process in detail by successively describing the heuristics on which the signal optimization process is based, the signal-switching decision process, and the logic used to promote coordination with adjacent signalized intersections.

8.1. Signal Switching Rules

In order to allow SPPORT to react to major traffic events affecting signal timing decisions, a series of signal switching rules were developed. These rules are based on fundamental traffic control principles and represent the various and often competing objectives of the decision making process. Each one is composed of two distinct elements. The first element is a condition that merits some sort of a response from the signal controllers. Examples of conditions include queues of vehicles containing more than a certain number of vehicles, the arrival of a platoon at the intersection stop line, the arrival of a transit vehicle at a transit stop, or the arrival of such vehicle at the intersection. The second element is the response to the event. Depending on the identified traffic conditions, the response could be a call for a green or red signal indication to be displayed at the current time or some projected future time.

This section describes in detail the rules that were developed to allow SPPORT to efficiently control traffic at urban intersections. The description is organized by the function of the main control objective that each set of rules attempts to satisfy at each intersection:

- Queue management,
- Congestion management,
- Uninterrupted traffic progression, and
- Transit priority.

The distinction between the various objectives is important, as each one looks at specific elements of the traffic demand and proposes different actions to deal with the identified problems. For example, the rules for queue management deal with the queues of vehicles that form at the stop line of each intersection approach when traffic is facing a red signal indication. Their main objective is to minimize the delays experienced by vehicles while waiting for the return of a green signal indication. Such an objective is attained by promoting a return to a green signal indication as soon as possible after a queue of a certain size has been detected and by ensuring that enough time will be offered during the scheduled green interval to completely dissipate the queues that have formed during the red interval.

The rules for congestion management also deal with the queues of vehicles that form at the stop line of every approach as a result of the signal operation. However, these rules are also concerned with the queues that are caused by other traffic events, such as the queues caused by transit vehicles stopping in the right-of-way in front of a transit stop to board and discharge passengers. The main objective of these rules is to prevent these queues from extending across upstream intersections and blocking traffic movements at these intersections. When such blockage occurs and is not fixed quickly, widespread congestion could then occur as queues from the upstream intersections start in turn to spill across other intersections further upstream. Consequently, the role of these rules is twofold. First, it is to promote queue dissipation on approaches on which stop line queues threaten to spill across upstream intersections. This is achieved by promoting the return of a green signal indication to these approaches as soon as possible. Second, it is to prevent the blockage of the intersection under control whenever queues on exit links threaten to spill across the intersection. This objective is achieved by implementing a red signal on the approach sending traffic onto the congested exit link so that the growth of the problematic queue can be stopped.

The rules for uninterrupted traffic progression attempt to prevent unnecessary stops and delays on intersection approaches by promoting coordinated traffic signal operation with upstream signalized intersections. Contrary to queue management rules, which deal with existing queues, these attempt to prevent queuing from occurring in the first place by attempting to align the periods during which a green signal indication is provided on each intersection approach with the periods of high arrival flow rates on these approaches.

Finally, the transit priority rules attempt to minimize the stops and delay incurred by transit vehicles around controlled intersections as a result of the signal operation. In this case, the control objective is

also twofold. The first objective is to try to provide clear passage to transit vehicles up to their loading point along a given approach. The second objective is to try to provide these vehicles with uninterrupted progression across controlled signalized intersections.

8.1.1. Rules for Queue Management

In SPPORT, queue management objectives are implemented through the three following rules

- The *Begin Serving Queue* rule,
- The *Continue Serving Queue* rule, and
- The *Begin Serving Maximum Wait* rule.

In the SPPORT model for isolated intersections developed by Conrad (1997), only the *Begin Serving Queue* and *Continue Serving Queue* rules were present. The *Begin Serving Maximum Wait* rule is an addition of the current research project. In addition, both the *Begin Serving Queue* and *Continue Serving Queue* rules have been subject of revisions. While the principles applied by these two rules remain the same, modifications were made in the way that these two rules evaluate traffic conditions on intersection approaches and calculate the time at which a green signal indication is requested.

The *Begin Serving Queue* rule only generates signal requests on intersection approaches on which a red signal is currently displayed. It calls for the start of a green interval on these approaches when the stop line queues exceeds a user-defined threshold. No other queues than the stop line queue are considered to avoid generating requests for queues that are not caused by the signal operation, such as the queues created a mid-block by transit vehicles blocking one traffic lane while dwelling in the right-of-way or by any other source of traffic disruption.

To avoid providing or extending a green signal indication to intersection approaches on which stop line queues cannot be served efficiently, requests to serve existing queues are only generated if the rate at which queued vehicles are projected to be able to move across the stop line exceeds a user-defined *critical flow rate*. This critical flow rate is uniquely defined on each link and is expressed as a fraction of the link's saturation flow rate. It represents the minimum rate at which traffic is assumed to be efficiently served on a given link. This criterion plays an important role in situations in which traffic conditions on one or more the exit links from an intersection hinder the maximum rate at which

vehicles can flow across the intersection, such as when a transit vehicle stopped in the right-of-way just downstream of an intersection completely blocks one traffic lane during the green interval.

For most of the tests carried out in the course of this research project, a critical flow rate corresponding to 2/3 of each link's saturation flow rate was specified. This resulted in no *Begin Serving Queue* requests being generated on links on which the projected average exit flow rate over the time required to dissipate the stop line queue was found to be less than 2/3 of the link's stop line saturation flow rate. This value was chosen on the basis of simulation runs evaluating the effectiveness of the signal timings produced by SPPORT with different critical flow rates. The results of these simulations will be presented in Chapter 9.

The *Continue Serving Queue* rule complements the *Begin Serving Queue* rule. It is only applied on links with a green signal indication and calls for a continuation of this indication until the number of vehicles that have been simulated to cross the stop line since the beginning of the green interval exceeds the number of vehicles that were part of the stop line queue at the end of the red interval. In this case, however, requests are only generated on links on which the exit flow rate is projected to remain above the link's critical flow rate over the next few seconds. This criterion is currently evaluated by estimating the link's projected exit flow rate over the next five seconds, and then over the next 10 seconds. Two distinct evaluations are made to add robustness to the request generation process. Following this evaluation, a *Continue Serving Queue* request is only generated if any one of the two estimated rates exceeds the link's critical flow rate. On the opposite side, no request is generated when none of the two estimated rates meets the critical flow rate criterion. This criterion has been designed by Conrad (1997) to avoid holding up the green signal indication on an approaches on which the queues that cannot be efficiently served. For example, *Continue Serving Queue* requests may stop being generated on an approach on which a queue is currently being served if a traffic event, such as a transit vehicle stopping in the right-of-way to board and discharge passengers, causes a blockage of the approach and a sudden drop in the stop line arrival flow.

The last rule, *Begin Serving Maximum Wait*, prevents excessive queueing time to occur as a result of long red intervals by issuing a call for an immediate green signal indication on links on which vehicles have been waiting for a period exceeding a user-defined threshold. This rule is applied each time the total delay incurred by a single vehicle while waiting at a red signal exceeds the defined threshold. It is

also applied when the waiting is projected to exceed to user-defined threshold by the soonest time a green interval can be implemented to serve queued vehicles on the link.

Similar to the *Begin Serving Queue* rule, the *Begin Serving Maximum Wait* rule only generates requests on links on which it is projected that queued vehicles can be served at a rate exceeding the user-defined critical rate. This rule is different from a maximum red interval constraint in that it is demand-sensitive. First, the rule is only applied on links on which there is at least one queued vehicle at the stop line. Second, the waiting time triggering the generation of a request calling for a signal display change is calculated from the moment the first vehicle has stopped, and not simply from the beginning of the red signal. Consequently, no request for a green signal indication is generated on links on which there is no stopped vehicle. No requests are also generated on links on which the red signal has been displayed for a long time but on which vehicles have only been stopped for a relatively short time.

8.1.2. Rules for Congestion Management

To avoid queue spillbacks, it is desirable to control traffic signals so that they do not generate excessively long queues and do not send large numbers of vehicles onto links on which queues threaten to spill across the intersection under control. As shown in Figure 8.1, different signal responses are required depending on where the spillback problem is occurring. If the spillback occurs across the controlled intersection as a result of a bottleneck on one of the exit links, the ideal response is to stop sending traffic across the intersection and onto the congested link until traffic starts to move again. However, for spillbacks occurring at the upstream end of an approach, the response should be to start serving traffic on the congested approach as soon as possible.

To provide the above responses, the three followings are currently being used by SPPORT:

- The *Begin Serving Upstream Spillback* rule,
- The *Continue Serving Upstream Spillback* rule, and
- The *Stop Feeding Downstream Spillback* rule.

Similar to the rules for queue management, the first two rules were present in the SPPORT model for isolated intersections developed by Conrad (1997), while the third rule is an addition of the current research project. However, similar to the *Begin Serving Queue* and *Continue Serving Queue* rules, the *Begin Serving Upstream Spillback* and *Continue Serving Upstream Spillback* rules were also the

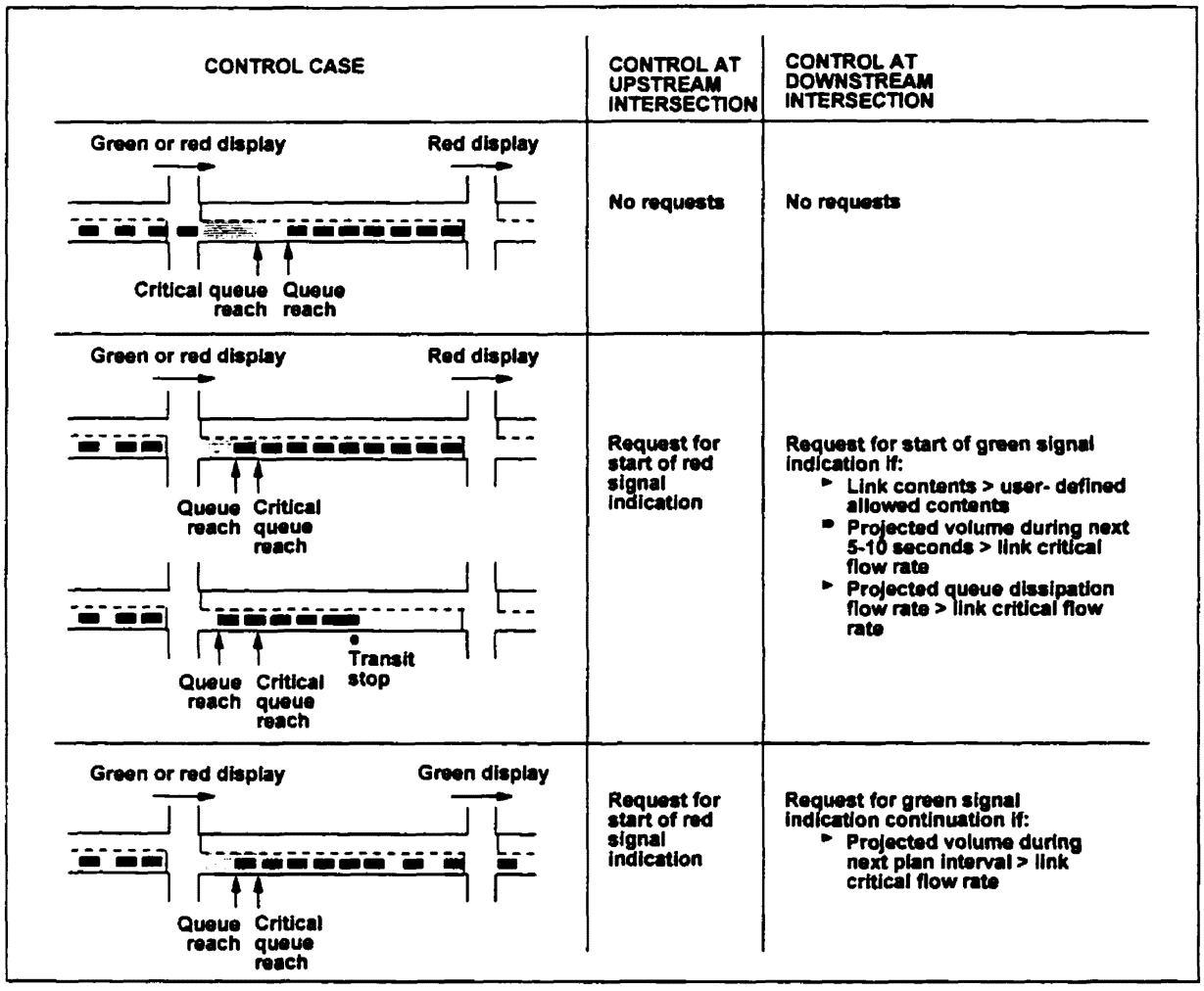


Figure 8.1 - Response to Queue Spillbacks across Signalized Intersections

subject of numerous minor modifications in the way they evaluate traffic conditions and calculate signal indication request times.

The *Begin Serving Upstream Spillback* rule is only called on links on which a red signal is displayed. This rule provides the following responses to spillbacks occurring on approach links:

- A call for an immediate green signal indication on links on which the tail of the stop line queue currently exceeds the *critical queue reach*, which is a used-defined parameter identifying the most upstream location where queuing is tolerated on the link.
- A call for an immediate green signal indication when it is projected that arrivals at the intersection will cause a stop line queue to exceed the critical queue reach on one of the intersection approaches within the look-ahead period.

In these responses, the tail of the stop line queue is considered instead of the tail of the most upstream queue to avoid switching the signal display to try to fix problems that are not caused by the signal operation. A typical example of such a situation is shown in the bottom diagram of Figure 8.1, in which a dwelling transit vehicle creates a queue spillback at the upstream intersection. In this case, it is obvious that switching the signal heads to a green display would not solve the queuing problem. In order to avoid wasting intersection traffic capacity, calls for green signal indications are also only generated by this rule when the projected average flow rate over the whole queue dissipation period is projected to remain above the link's critical rate.

The man difference between the two responses listed above is based on whether a spillback currently exists or is projected to occur. While the first response deals with current spillbacks, the second attempts to reduce the likelihood of future spillbacks. As indicated in Section 7.3, the demand placed on a given intersection is estimated by assuming that a green signal indication can be returned to each approach for which signal-switching decisions are being made as soon as phasing requirements allow it. This results in a projection of stop line departures under an ideal scenario in which the maximum amount of time is made available to serve the incoming traffic. At the same time as vehicles as vehicles are simulated to exit the intersection approach at its downstream ends, the arrival of vehicles from upstream intersection is simulated at the upstream end. While these projected arrivals are based on the signal timing strategy currently considered at the upstream intersection, situations could then still arise in which the projected arriving traffic may be forced to join the tail of an existing queue. Consequently, if an upstream spillback is projected to occur within the look-ahead period, it is an indication that the problem has not been successfully considered by the upstream controller and that an immediate call for a green signal indication should be made to start the queue dissipation process as soon as possible and avoid worsening the situation.

The *Continue Serving Spillback* rule calls for an extension of the current green signal indication until the tail of the most upstream queue ceases to exceed the critical queue reach. In this case, upstream queues are considered since the simulation model cannot distinguish between a stop line queue currently dissipating and a mid-block queue. Similar to the *Continue Serving Queue* rule, *Continue Serving Spillback* requests are only generated if the rate at which vehicles are projected to flow across the stop line in either the next five or ten seconds remains above the link's user-specified critical flow rate. As a result, queues that are located more than ten seconds away from the intersection and that are currently dissipating will not result in the generation of a *Continue Serving Queue* request. Instead, if a

sufficient number of vehicles dissipating from this queue are projected to reach the intersection stop line before the end of the look-ahead period, these vehicles will be considered as part of an incoming platoon and could result in the generation of a *Begin Serving Platoon* request.

To avoid unduly holding up the green signal indication for long periods of time, the *Continue Serving Queue* requests are also only generated when the number of vehicles on a link exceeds half its total queuing capacity. The rationale for this criterion is to allow service to be given to other links once sufficient storage capacity has been gained on a congested link. Experience has shown that waiting for the complete dissipation of large queues often results in the generation of similarly long queues on other approaches, which could in turn cause spillbacks at other intersections.

The *Do not Feed Downstream Spillback* rule addresses spillbacks occurring on links leading traffic away from controlled intersections. Having the goal of stopping the growth of queues that may cause intersection blockages, this rule provides the three following responses to potential queue spillbacks:

- A call for an immediate red signal on links discharging traffic onto links on which the tail of the most upstream queue currently exceeds the critical queue reach.
- A call for an immediate red signal on links on which it is projected that vehicles released from these links will cause a spillback to occur across the controlled intersection before the current green interval can be ended.
- A call for a red signal at the time future queue spillbacks are projected to occur within the look-ahead period when no other calls for red have been made.

Here, the tail of the most upstream queue is used as a control criterion instead of the tail of the stop line queue as this queue would be the first to start to spill across the controlled intersection if something is constraining queue dissipation on the exit link. In addition, the rule does not necessarily evaluate queuing conditions on all exit links before determining if a *Do not Feed Downstream Spillback* request should be generated. The decision as to which link to consider is left to the user. For each approach link, the user must indicate with which exit links coordination should be attempted. This design allows the user to set up the signal optimization in a way that could prevent traffic from being held at an intersection only in response to the threat of a potential queue spillback from a minor street. For example, it makes no sense to hold up 30 vehicles attempting to go straight across an intersection onto an exit link with no queuing problem just because two or three vehicles might want to turn right or left on a congested exit link. In many situations, drivers would even alter their path to avoid the problem.

8.1.3. Rules for Uninterrupted Traffic Progression

The SPPORT model defines three rules for providing uninterrupted progression along sets of intersections:

- The *Begin Serving Platoon* rule,
- The *Continue Serving Platoon* rule, and
- The *Continue Serving High Demand* rule.

Once more, the first two rules were present in the initial model developed by Conrad (1997), while the last rule, *Continue Serving High Demand*, was added as part of the current research project. In this case, the only modifications made to the existing rules are associated with how they identify incoming platoons on intersections. The principal modifications to the platoon identification algorithm were discussed in Section 7.4 of the previous chapter. Another important modification deals with the period over which traffic arrivals are considered in the platoon identification process. In the initial model, only the stop line vehicle arrivals at and after the current decision point were considered. In the current model, the analysis also includes the arrivals that occurred a few seconds before the decision point. This allows the rules, particularly the *Continue Serving Platoon* rule, to better identify whether a platoon is currently being served or not.

The *Begin Serving Platoon* rule calls for a green signal indication, regardless of the current signal display, at the times platoons are projected to reach the stop line. As shown in the time-space diagram of Figure 8.2, this rule does not necessarily call for an immediate switch to a green signal indication. The time at which a call is placed depends on the arrival time of the first vehicle in the platoon, the time needed to clear vehicles ahead of the platoon, and on the reaction time of drivers at the onset of the green interval.

The *Continue Serving Platoon* rule operates similarly to the *Continue Serving Queue* rule. This rule calls for a continuation of the current green signal indication on links on which a platoon is currently being served. In this case, a platoon is assumed to be currently served when the two following conditions are simultaneously met:

- A platoon of vehicle has started to cross the stop line within the previous 10 seconds,
- The projected stop line departure flow rate is projected to remain above the link's critical flow rate over either the next five or ten seconds.

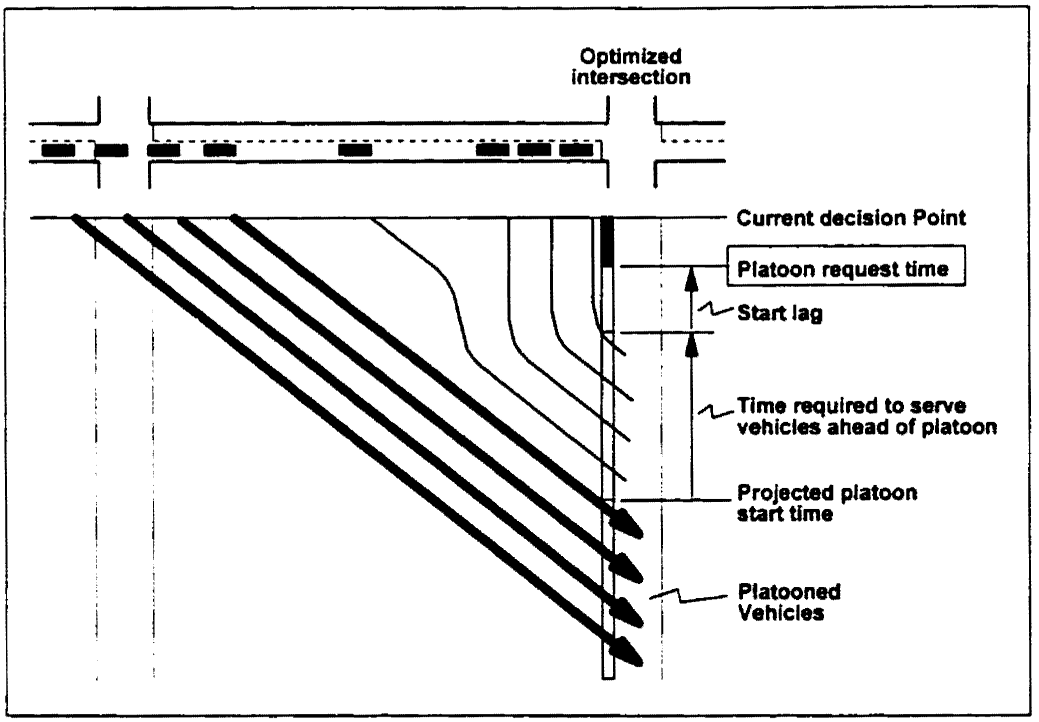


Figure 8.2 - Example of Platoon Request on Intersection Approach

The *Continue Serving High Demand* rule calls for the continuation of the current green signal indication on links on which the exit flow rate is projected to remain above the user-specified critical flow rate over either the next five or ten seconds and on which the application of the *Continue Serving Queue* and *Continue Serving Platoon* rules has not resulted in the generation of signal requests. In this case, however, the main criterion determining if a request should be generated is whether there are benefits, in terms of stops and delays, associated with extending the current green signal indication, as opposed to initiating an immediate signal change. The benefit is evaluated for each link using Equations 8.1, 8.2 and 8.3:

$$B_k = \sum_{j=1}^{N_{link}} \left\{ (1 - U_j) \left[k_d \, d_{saved\,j} + k_s \, s_{saved\,j} \right] - (U_j) \left[k_d \, d_{induced\,j} \right] \right\} \qquad [8.1]$$

$$d_{saved\,j} = q_{next\,\Delta t\,j} \, R_{min\,j} \qquad [8.2]$$

$$d_{induced\,j} = Q_{stop\,j} \cdot \Delta t \qquad [8.3]$$

$$s_{saved\,j} = q_{next\,\Delta t\,j} \qquad [8.4]$$

where: B_k = Benefit at intersection k of extending the green signal indication on link j for one plan interval Δt.

k_d = User-specified coefficient defining the relative importance of delay ($k_d \geq 0$).

k_s = User-specified coefficient defining the relative importance of stops ($k_s \geq 0$).

$d_{saved\,j}$ = Delay saved on link j by extending the green signal indication by Δt seconds (seconds).

$d_{induced\,j}$ = Delay induced on link j by extending the red signal indication by Δt seconds (seconds).

$s_{saved\,j}$ = Stops saved on link j by extending the green signal indication by Δt seconds (passenger car units).

$Q_{stop\,j}$ = Number of vehicles queued at the stop line of link j (passenger car units).

Δt = User-specified interval between decision points (typically five seconds).

$q_{next\,\Delta t\,j}$ = Projected arrival on link j in the next plan interval (passenger car units).

$R_{min\,j}$ = Minimum red interval on link j based on phasing constraints if a switch from a red to green signal display is initiated now (seconds).

Equation 8.1 compiles the stops and delays saved on links currently having a green signal indication by extending the display by Δt seconds, as well as the delays induced on other links by pushing further in time the soonest moment at which a green interval can be initiated on them. The formula also indicates that stops and delays are strictly compiled on a vehicular basis. Contrary to the performance evaluation of candidate phase plans, there is no weighting for the stops and delays specifically incurred by passenger cars and transit vehicles. The rationale for performing such a compilation comes from the fact that specific requests have been developed to handle the priority requirements of transit vehicles. As a result, a need only remains for providing uninterrupted progression for passenger cars.

8.1.4. Rules for Transit Priority

To provide priority to transit vehicles on an active basis, the two following rules are provided:

- The *Bring Transit Vehicle to Transit Stop* rule, and
- The *Bring Transit Vehicle to Stop Line* rule.

Both these two rules were present in the model developed by Conrad (1997) and were only the subject of minor modifications.

The *Bring Transit Vehicle to Transit Stop* rule attempts to provide transit vehicles with a clear passage up to the loading point. This rule is only applied on links on which the number of vehicles currently ahead of an incoming transit vehicle is greater than the total queue capacity downstream of the loading point. It is not applied in other cases, as a switch to a red signal would then not cause the loading point to be blocked by queued vehicles. When the rule should be applied, the two following cases are then considered (see Figure 8.3):

- When a red signal is currently displayed at the stop line, a request is generated to switch the signal heads to a green display at the latest time that will allow queued vehicles to clear the loading point just before the transit vehicle is expected to reach it.
- If a green signal indication is currently displayed, a request to maintain the indication is generated until the remaining number of vehicles ahead of the transit vehicle can all be stored downstream of the transit stop.

The *Bring Transit Vehicle to Stop Line* rule complements the above rule. It calls for a green signal indication when a transit vehicle is projected to reach the intersection stop line. Similar to the *Begin Serving Platoon* requests, the time associated with the request also takes into account the time needed to dissipate queues of vehicles that may have formed at the stop line ahead of the transit vehicle. In this case, a call for a green signal indication is placed just soon enough to allow any leading vehicles to clear the intersection before the approaching transit vehicle crosses the stop line.

8.2. Rule-Based Decision-Making Process

While the rules introduced in Section 8.1 define reactions to key traffic events, they do not provide any indication as to which events should be served in priority when competing requests for green or red signal indications are generated. This section describes the rule-based decision-making process used to solve these conflicts and develop signal control decisions reflecting both current traffic conditions and control objectives. Specifically, this section describes:

- The prioritization of rules determining the relative importance of each request,
- The generation of request lists,
- The signal-switching decision process,
- The handling of phase sequence restrictions, and
- The handling of timing restrictions.

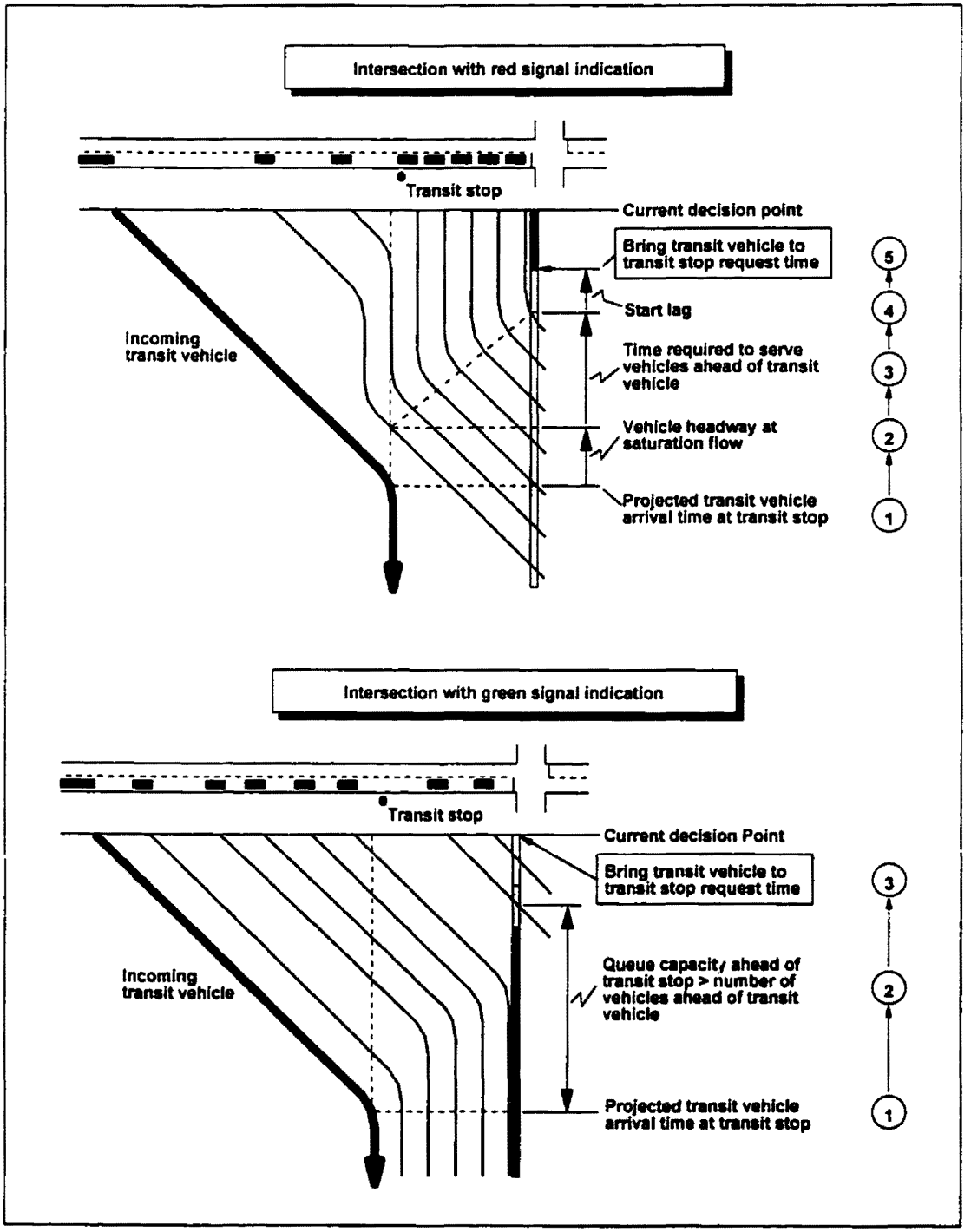


Figure 8.3 - Response to a Transit Vehicle Approaching a Transit Stop

8.2.1. Rule Prioritization

In order to determine the relative importance of each possible event, the user is required to prioritize the rules used to generate signal-switching decisions, by specifying a priority table for each individual link in the system. The priorities are link-specific to reflect the fact that responses to specific events might not carry the same importance on all links. For each link, the table must be filled with numerical values ranging from zero to 1000, where zero indicates that the request should not be considered in the optimization process and 1000 provides the highest possible priority level.

Table 8.1 indicates how the user-specified values are converted into priority levels. In most cases the provided values are directly converted into priorities. The only exceptions are for the rules dealing with queues and platoons. For these rules, the priorities are a function of the square of the size of the queue or platoon to serve to reflect the growing importance of serving longer queues and larger platoons in the signal optimization process. Consequently, the user is also required to provide maximum priority values to prevent SPPORT from assigning excessively high priorities to requests associated with long queues and large platoons. If required, fixed priorities can also be imposed on the need to serve queues and platoons. This can be done by specifying identical lower and upper priority limits for all requests for which two such parameters must be provided.

Table 8.1 - Request Priority Evaluation Functions

Rule	Priority Level
Begin Serving Queue	Min [User priority · (Queue size)^2, User max priority]
Continue Serving Queue	Min [User priority · (Green onset queue size)^2, User max priority]
Begin Serving Upstream Spillback	User priority
Continue Serving Upstream Spillback	User priority
Do not Feed Downstream Spillback	User priority
Begin Serving Platoon	Min [User priority · (Platoon size)^2, User max priority]
Continue Serving Platoon	Min [User priority · (Platoon size)^2, User max priority]
Continue Serving High Volume	User priority
Bring Transit Vehicle to Transit Stop	User priority
Bring Transit Vehicle to Stop Line	User priority
Begin Serving Excessive Wait	User priority

There are currently no specific rules regarding the assignation of priority values to the various rules. It is simply recommended that the priority levels be established based on sound engineering judgement, and possibly by conducting a series of simulation tests. For example, in the scenarios that were developed to evaluate the SPPORT model and that will be described in Chapter 9, the priority levels that were assigned to the different rules were determined using a two-step process. In the first step, the relative importance of the various traffic events was qualitatively assessed. In the second step, a set of numerical values was determined and refined through a detailed analysis of the behavior of the model in various simulated scenarios.

8.2.2. Generation of Request Lists

At each intersection, the signal-switching decision process starts with the evaluation of current traffic conditions on each approach to the intersection under control. This evaluation then leads to the generation of green and red signal indication requests reflecting the needs of current traffic demands. In this process, requests are generated on the basis of the traffic conditions that are projected to exist on each approach at the moment vehicles would stop entering the intersection on an intergreen period on the approaches on which a green signal indication is currently displayed if a decision to switch the signal indication to red would be taken immediately. Figure 8.4 illustrates graphically the moment at which traffic conditions are evaluated at each decision point for the generation of signal requests. At each point, vehicles crossing the stop line on a green signal indication prior to the request evaluation time are not considered as part of the current demand, but as part of past demands, since these vehicles would enter the intersection under control regardless of the signal-switching decision made at the current decision point.

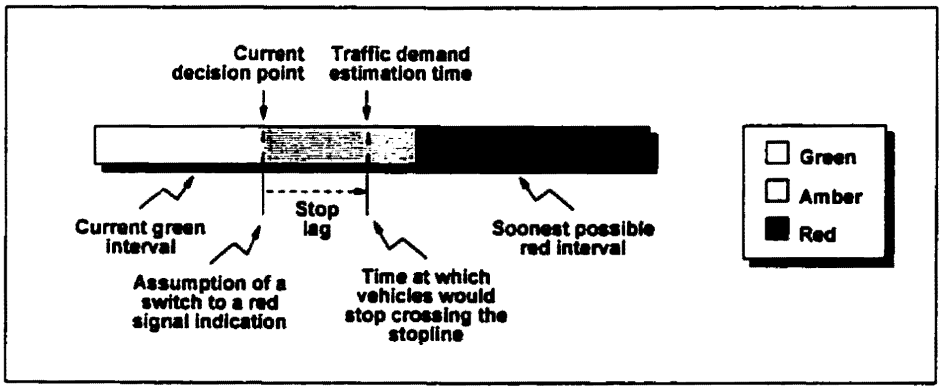


Figure 8.4 - Traffic Conditions Considered when Generating Signal Requests

For each intersection, requests for green and red signal indications are generated in turn for each approach link. For each request, the number, priority level, link for which a green or red signal display is requested and the time at which the signal indication is to be displayed are recorded. After all the requests have been generated, they are separated into active and inactive lists. As it will be explained in the next section, these two lists are used to evaluate the ability of each candidate phase to satisfy current traffic demands.

A request is classified as active if, as a result of minimum green interval and intergreen period constraints, none of the phases that might be used to service the request can be effectively displayed before the requested time. All other requests are considered as inactive. Figure 8.5 illustrates how the classification is made on each intersection approach. Figure 8.5a illustrates how requests are classified as active and inactive on an approach on which the current signal indication is red, while Figure 8.5b illustrates the classification for an approach with a green signal indication.

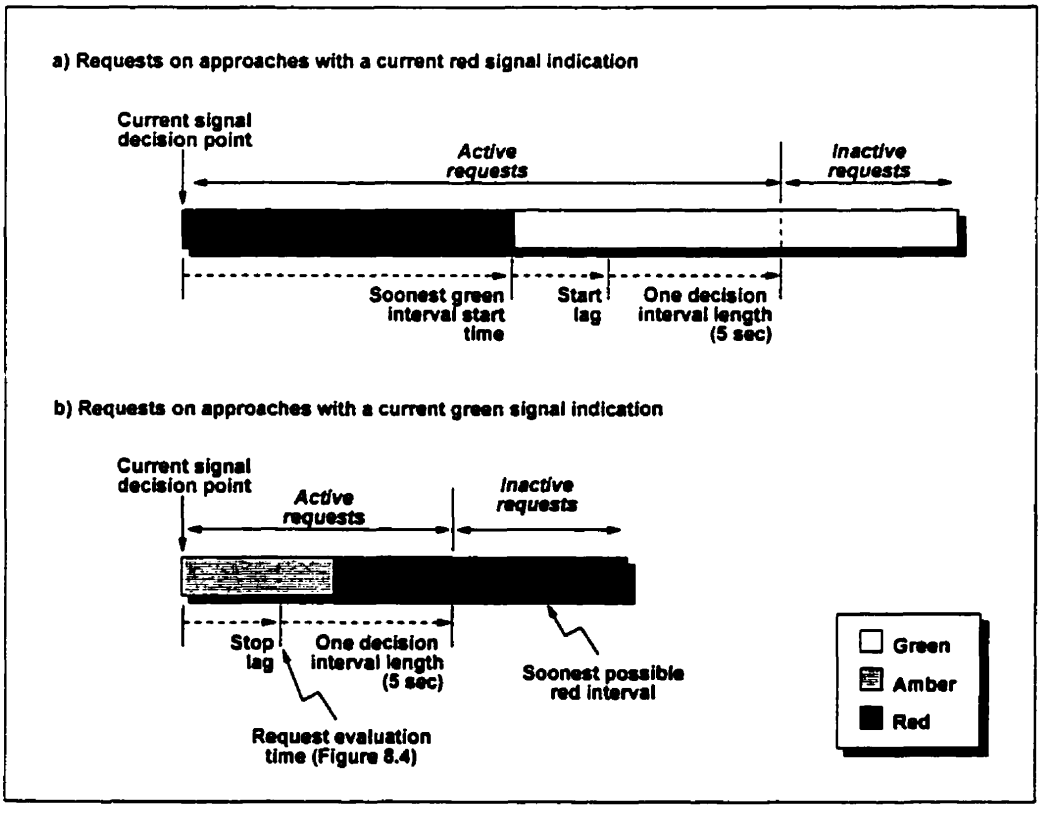


Figure 8.5 - Active and Inactive Requests

In the case of Figure 8.5a, all requests calling for a green display before the soonest time at which such display can be effectively initiated are classified as active since these requests cannot obviously be served in time. The requests calling for a green signal initiation within one decision interval after the soonest possible effective green interval start time are also considered as active. This is to reflect the fact that these requests, which would be otherwise inactive, would become active at the next decision point if a decision to switch the signal heads to a green display is not made at the current decision point.

To illustrate this point, consider an approach on which the signal optimization is done every five seconds and where ten seconds are required to make the switch from a red signal display to a green display. Also, consider that a request to serve an incoming platoon is calling for a green signal indication 12 seconds after the current decision point. In this case, if a decision to switch the signal to green is made at the current decision point, a green interval can then be initiated in ten seconds, that is, two seconds ahead of the projected platoon arrival time. However, if the signal switch is delayed five seconds up to the next decision point, a green interval can then only be initiated when the clock will mark 15 seconds. In this case, the green signal onset would occur three seconds too late to serve the platoon in time and would cause additional stops and delays to the incoming vehicles. As a result, by ensuring that all requests that would become active at the next decision point are considered as such at the current decision point, it thus becomes possible for SPPORT to better plan signal switches as a function of near-future traffic arrivals.

In the case of Figure 8.5b, it is observed that the interval over which requests are considered as active is much shorter. This is to reflect the fact that SPPORT evaluates requests for green signal extensions on an interval-by-interval basis. In most situations, requests calling for the continuation of a given green signal indication have a call time corresponding to the current decision point. As a result, these requests are automatically considered as active since any switch to a red signal indication at the current decision point would prevent them to be served in time. Similar to approaches on which a red signal is currently displayed, all the requests that would become active on other approach links at the next decision point if a switch to a red signal would be initiated at the current decision point are considered active to improve the ability to plan efficient signal switches.

8.2.3. Switching Decision Logic

The main goal of the signal-switching decision process is to serve the highest number of high-priority conditions requiring immediate service while delaying service to the least number of high-priority conditions requiring service at some point into the future. To fulfill this goal, SPPORT evaluates all candidate phases that can be implemented at a given time and selects for implementation the one having the highest rating. At any given time the selection of the current phase as the best phase indicates that service to that phase should be continued for another plan interval. On the other hand, the selection of a phase other than the current one indicates that the signal switching procedure between the current phase and the selected one should start immediately.

Candidate phases are evaluated by compiling the benefits and future costs of implementing that phase at the current decision point. The immediate benefit of selecting a phase is taken as the sum of the priorities of all the active requests that would be served by switching the signal displays to that phase. The future cost of that phase is calculated as the sum of the priorities of all the inactive requests to which service would be delayed if a switch to the phase is initiated immediately. In the later case, delayed service to an inactive request is determined by analyzing if switching to the selected phase would cause the inactive request to immediately become active. In all cases, the benefit and cost of a particular phase are calculated by adding together the priority level of all active or inactive requests in a simple linear way to reflect the cumulative importance of serving more than one request at a time.

In assessing whether service to an inactive request would be delayed, the user has the option of specifying an *allowable delay* for each request. This parameter adds flexibility to the signal-switching decision process by permitting switches that may result in slight delays to high priority inactive requests. For instance, specifying a five-second allowable delay for the *Begin Serving Queue* requests indicates to the signal optimizers that no costs should be calculated for the selection of phases that may delay service to the existing queues by less than five seconds.

8.2.4. Handling of Phase Sequence Restrictions

In making a switching decision, it is possible to arrive at a situation in which one or more of the phases being considered cannot legally follow the current one as a result of phase sequence restrictions. When such situations are encountered, a logical decision appears to be the immediate dismissal of the phase.

However, limiting the switching decision process to those phases that can legally follow the current one may be overly restrictive in many scenarios. In order to keep as many options open as possible, switching decisions requiring the display of a single intermediate phase between the current phase and the desired one are considered in the switching decision process. In this case, the model assumes that the intergreen period between the current phase and the one being considered consists of the shortest possible intermediate phase between them.

If a phase requiring an intermediate phase is selected as a result of the decision-making process, there may be more than one possible intermediate phase than can be implemented. When such situations occur, the intermediate phase with the highest rating is used. However, since it is initially assumed that the shortest possible intermediate phase would be inserted between the current phase and the selected one, it results from this assumption that the selection of longer intermediate phases is not as desirable as the selection of shorter ones. Consequently, the rating of all candidate intermediate phases is reduced in proportion to the minimum amount of time required by them before a signal switching decision is made.

In some situations, the phases that can legally follow the current phase and those requiring the display of a single intermediate phase may not account for all the phases that can be implemented at an intersection. For instance, there could be phases requiring the display of more than one intermediate phase. While such phases could be considered in the signal optimization process to broaden the range of possible decisions, it is unlikely that their use would provide any practical benefits. In this case, the time spent serving the intermediate phases and the increased lost time resulting from the implementation of additional intergreen periods would likely make serving multiple intermediate phases impractical. For this reason, the search process at any decision point is currently limited to those phases that do not require more than one intermediate phase.

8.2.5. Handling of Signal Timing Constraints

In addition to phase sequence restrictions, the SPPORT model allows the user to specify minimum and maximum duration constraints on all defined phases. Minimum green interval duration constraints are usually imposed by traffic engineers to ensure that a minimum service time is provided by a given phase, either to allow a certain number of vehicles to cross the intersection or to provide enough time for pedestrians to cross the intersection. On the other side, maximum constraints are imposed to ensure

that traffic conditions on a given approach will not result in the implementation of excessively long green intervals on that approach, and consequently, of excessively long red intervals on other approaches. Minimum green interval duration constraints are enforced by SPPORT by locking the signal displays from the beginning of a given green interval until the end of the imposed minimum duration, while maximum constraints are enforced by attributing a very high cost to phases that have elapsed their maximum allowed time. In the later case, the imposition of a high cost to the phase having exceeded its maximum interval results in a very poor evaluation for that phase and causes the phase to be rejected as a potential best candidate in the signal-switching evaluation process.

8.3. Signal Coordination Logic

Signal coordination between adjacent intersections is implemented within SPPORT through the three following channels:

- Rule for handling of incoming platoons,
- Rules for handling queue spillbacks, and
- Request delaying logic.

8.3.1. Handling of Incoming Platoons

The first aspect of the signal coordination logic followed by SPPORT is the handling of incoming platoons on intersection approaches through the *Begin Serving Platoon* and *Continue Serving Platoon* rules. On each approach, the combined effect of these two rules is to promote the implementation of green intervals at the times platoons of vehicles are projected to reach the intersection stop line. Since incoming arrival patterns are functions of the timing strategies implemented at upstream intersections, any attempt to avoid stopping incoming platoons of vehicle at an intersection result in an implicit coordination of the operation of that with the operation of other upstream intersections.

The type of coordination that results from the application of the *Begin Serving Platoon* and *Continue Serving Platoon* rules is similar to the maximization of green bands along urban arterials in fixed-time traffic signal control systems. In these systems, a green band is a window of green signal indications that allows vehicle to travel uninterrupted along a set of adjacent intersections. In SPPORT, there is a similar attempt to align the periods of green signal indication of successive intersections to avoid stopping the progression of groups of vehicles. However, there is no explicit green band calculation,

and therefore, no explicit green band maximization. Signal timings are directly adjusted according to the needs of the traffic demands between pairs of adjacent intersections. In that respect, the type of coordination performed by SPPORT is more similar to the maximization of progression opportunities in TRANSYT-7F (Wallace and Courage, 1982; Hadi and Wallace, 1992), in which a progression opportunity is simply seen as the ability for a vehicle leaving an intersection at a given time to travel across the next intersection without being stopped.

8.3.2. Handling of Queue Spillbacks

A second aspect of the signal coordination logic is the handling of queue spillbacks through the *Begin Serving Upstream Spillback*, *Continue Serving Upstream Spillback* and *Stop Serving Downstream Spillback* rules. Here, the main objective is not to provide uninterrupted progression across signalized intersections but simply to prevent the queues produced by the signal operation from blocking adjacent intersections. While the first two rules accomplish this task by preventing signal controllers from implementing control strategies that could cause the blockage of upstream intersections, the third rule allows the controllers to react to the blockage of downstream links.

8.3.3. Request Time Delaying Logic

The most important aspect of the signal coordination logic are the rules used to strategically adjust the times at which green signal indications are requested on each intersection approach following the identification of traffic events potentially commanding a signal switch. In the initial SPPORT model designed for the control of isolated intersections, the time associated with each request was always the call time initially calculated by the corresponding rule. For instance, a request to serve a queue of vehicles always resulted in a call for an immediate green signal indication. In another example, the green signal indication call time associated with a request to serve an incoming platoon always corresponded to the expected platoon arrival time at the intersection stop line minus the time to serve all vehicles that were traveling or queued ahead of the platoon.

In the above model, traffic conditions downstream of the intersection under control were completely ignored. This assumption is typically made in isolated signalized intersection control problems. In a network setting, however, where intersections are relatively closely spaced, a totally different situation exists. In this case, the closeness of intersections puts a limit on the storage capacity of individual

links. As a result, the operation of adjacent signalized intersections must be coordinated to avoid situations in which traffic is allowed to leave an intersection while there is not enough remaining storage capacity on the downstream link to receive the incoming vehicles. In addition, it serves no purpose in such situations of having drivers held at one intersection watching a green signal indication at a downstream intersection, only to arrive at that intersection just when the signal turns red.

Within the SPPORT model, it is primarily the role of the platoon rules to promote good signal coordination between adjacent intersections, and the role of the queue spillback rules to avoid situations in which queues of vehicles may threaten to spill across signalized intersections. However, appropriate signal coordination between adjacent intersections can also be promoted by directly attempting to adjust the times at which green signal indications are requested on each intersection approach as a function of the traffic conditions on the major downstream links onto which traffic is flowing. In many situations, unnecessary stops and delays could be avoided by slightly delaying the times at which green signal indications are requested. Ideally, requests call time could be delayed up to a moment in time that would allow vehicles to travel uninterrupted along major downstream links and across the next intersection.

8.4. Calculation of Ideal Request Times

This section finalizes and illustrates how intelligent, experienced traffic cops might operate when they are educated in traffic engineering principles, looking in all directions simultaneously and communicating with one another. The fully distributed network version of SPPORT has the potential to accomplish this, but without requiring human "super cops". For each rule, the calculation of ideal request times in the request delaying process is a function of the following five elements:

- Implemented and projected signal timings at downstream intersections,
- Traffic conditions on downstream links,
- Transit interference on downstream links,
- Implemented and projected signal timings at upstream intersections, and
- Potential queue spillback at upstream intersections.

Before explaining how ideal request times are calculated, it must be pointed out that there might be situations in which request delaying is not desired. For instance, it might be desired that traffic from a given approach always be allowed to proceed towards the next intersection at the soonest possible

time, regardless of the fact that vehicles may not be able to go across the next intersection without stopping. In another example, it might also be required that platoon requests be never delayed to preserve existing progression patterns with upstream intersections. In order to account for these situations, the model has been structured to enable the user to specify for each link the requests that should and should not be subjected to the request delaying logic.

8.4.1. Coordination with Downstream Timings

The first task performed by the request delaying logic is to adjust the times at which green signal indications are requested on each approach as a function of both the current and projected signal timings at adjacent downstream intersections. The main objective of this adjustment is to delay initiation of green intervals at a given intersection in such a way that vehicles released from that intersection would reach the next intersection during a scheduled green interval.

The above adjustment is done on a request-by-request basis. The ideal green display time is calculated in turn for each request calling for a green signal indication that has been generated during the demand evaluation process. No calculation is performed for the requests calling for a red indication. For each request, the calculation starts by assuming that vehicles would leave the controlled intersection at the initial time at which a green signal indication is requested and that they would travel uninterrupted up to the stop line of the next intersection. At this point, the amount of time by which each request should be delayed to allow traffic to go across the downstream intersection without being stopped is determined on the basis of the time at which the first released vehicle from the controlled intersection is expected to reach the downstream intersection (see Figure 8.6):

- If the first vehicle is projected to reach the intersection during a scheduled green interval, no adjustment is necessary as good coordination already exists between the intersection under control and the downstream intersection
- If the first vehicle is projected to reach the intersection during a red interval while a future local green interval has already been scheduled for implementation (Figure 8.6a), the request is delayed by an amount of time that would allow the vehicle to reach the downstream intersection just at the beginning of the scheduled downstream effective green interval.

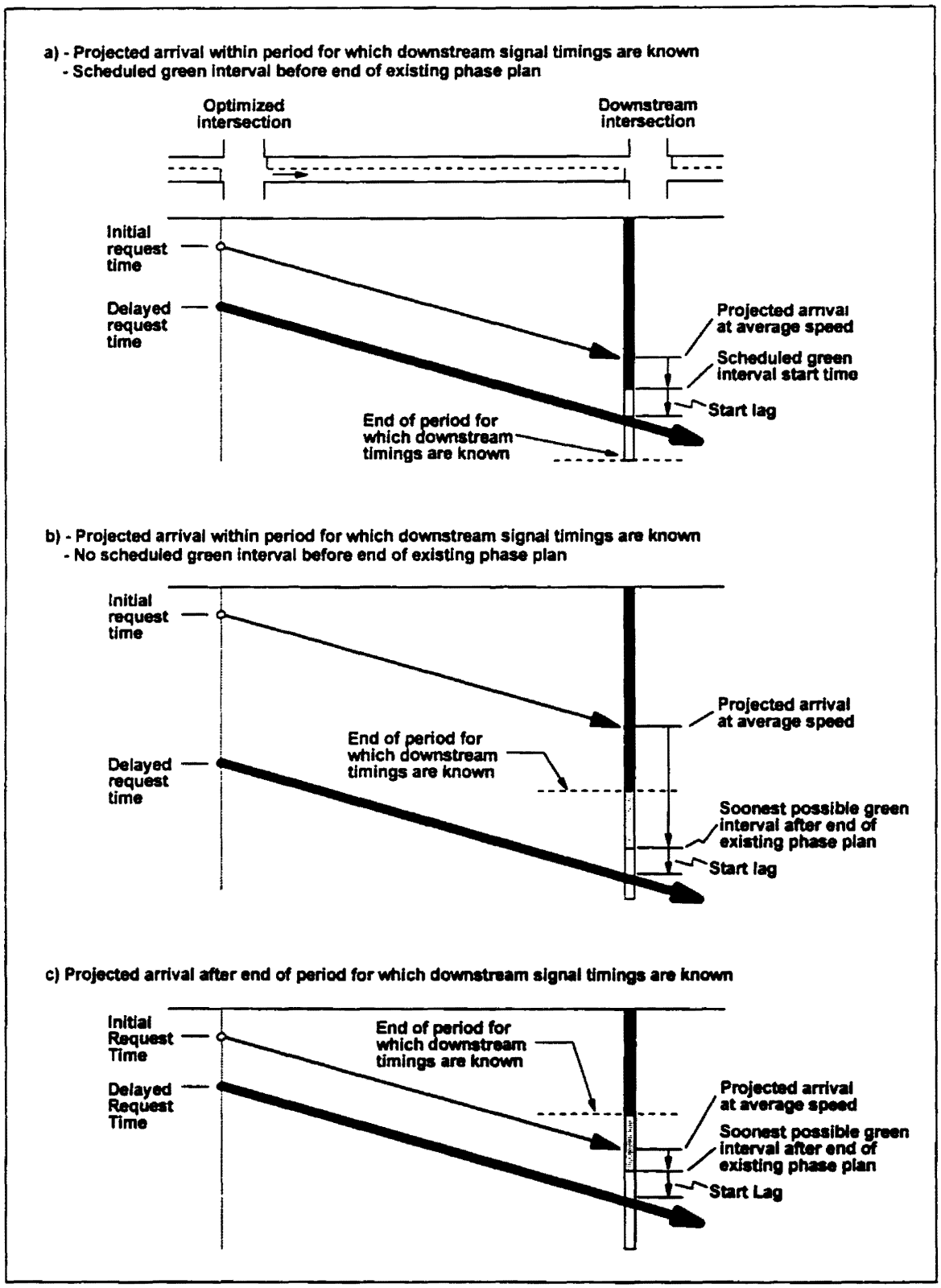


Figure 8.6 - Coordination of Signal Operation with Downstream Signal Timings within Request Delaying logic

- If the first vehicle is projected to reach the downstream intersection during a scheduled red phase while no future green interval has been scheduled (Figure 8.6b), or after the period for which projected signal timings are known (Figure 8.6c), the request is delayed as a function of the soonest time at which a green signal indication can be displayed at the downstream intersection.

At many urban intersections, it is not always required to coordinate the signal operation with the timings of all adjacent intersections. For example, it is usually not desired to coordinate the timings of the intersections along an arterial with the timings of adjacent intersections on minor cross-streets. In order to account for these situations, SPPORT requires that the user specifies with which exit link coordination should be attempted. These links are labeled as *main exit links* and are identified in the setup file of each signal optimizer. If more than one main exit link is specified for the traffic exiting an intersection from a given approach, the ideal delayed request time is calculated as a weighted average of all the individual offsets that were calculated for each link. This average is calculated according to Equation 8.5 using the turning percentages provided by the user when setting up the simulation model.

$$RT_{del} = RT_o + \frac{\sum_{j'=1}^{N_{exit\,j}} Turn_{j,j'} \left(RT_{del\,j} - RT_o\right)}{\sum_{j'=1}^{N_{exit\,j}} Turn_{j,j'}} \qquad [8.5]$$

where: RT_{del} = Delayed request time (seconds).

$RT_{del\,j}$ = Delayed request time based on traffic conditions on exit link j (seconds).

RT_o = Initial request time (seconds).

$Turn_{j,j'}$ = Turning proportion from approach link j to exit link j'.

$N_{exit\,j}$ = Number of exit links from approach link j.

8.4.2. Coordination when Downstream Queues Exist

When coordinating the operation of an intersection with its downstream neighbors, it is equally important to consider the interference caused by vehicles queued at the stop line of downstream intersections as it is to consider the projected timings for these intersections. The time-space diagram of Figure 8.7 illustrates the importance of such consideration. In this diagram, it is apparent that all vehicles queued at the downstream intersection do not immediately start to move at the beginning of

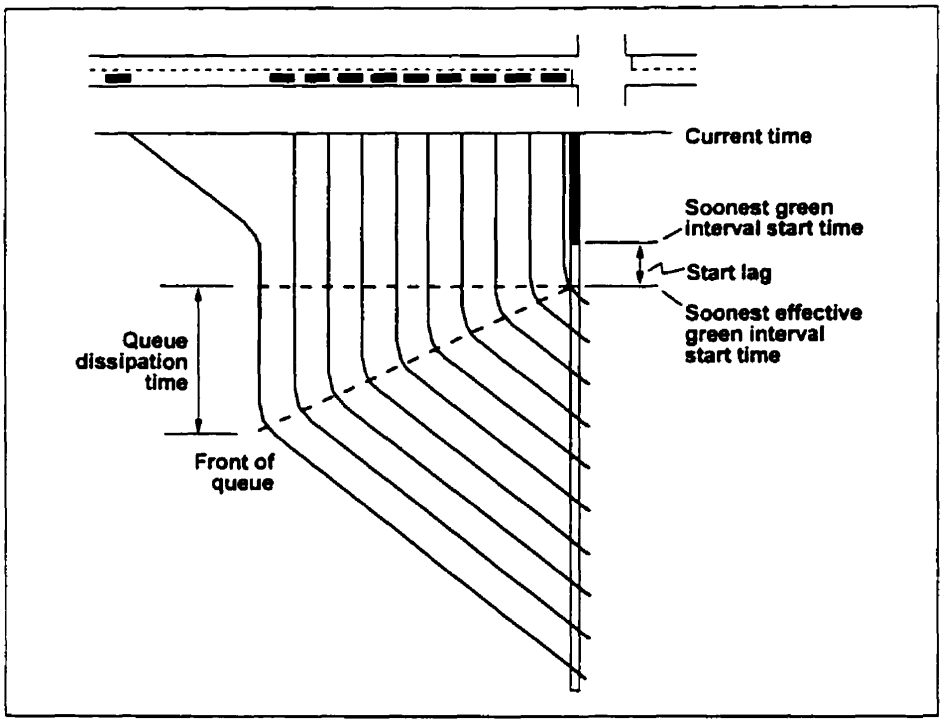


Figure 8.7 - Stop Line Queue Dissipation Time at the Beginning of a Green Interval

the green interval. On the contrary, there is a finite time interval between the moment the green signal indication is initiated and the moment the last vehicle in queue starts to move. As a result, any vehicle projected to arrive at the downstream intersection before the complete dissipation of the stop line queue will be required to stop at the tail of the remaining queue.

In order to avoid unnecessary stops and delays, traffic from upstream intersections should arrive at an intersection after the complete dissipation of the queue that has been generated during the red interval. To promote such arrangement, SPPORT attempts to adjust the times at which green signal indication requests are placed on each approach to an intersection as a function of the queuing conditions on the main exit links from that intersection. Figure 8.8 illustrates the basic impact of downstream queues on request times. The objective of the adjustment performed is to delay the initiation of the green interval at the optimized intersection so that vehicles leaving the intersection will travel uninterrupted along the links leading to the downstream intersections with which coordination is attempted.

The offset calculation starts at the stop line of the downstream intersection and then works its way back upstream, segment by segment, up to the stop line of the approaches to the intersection being

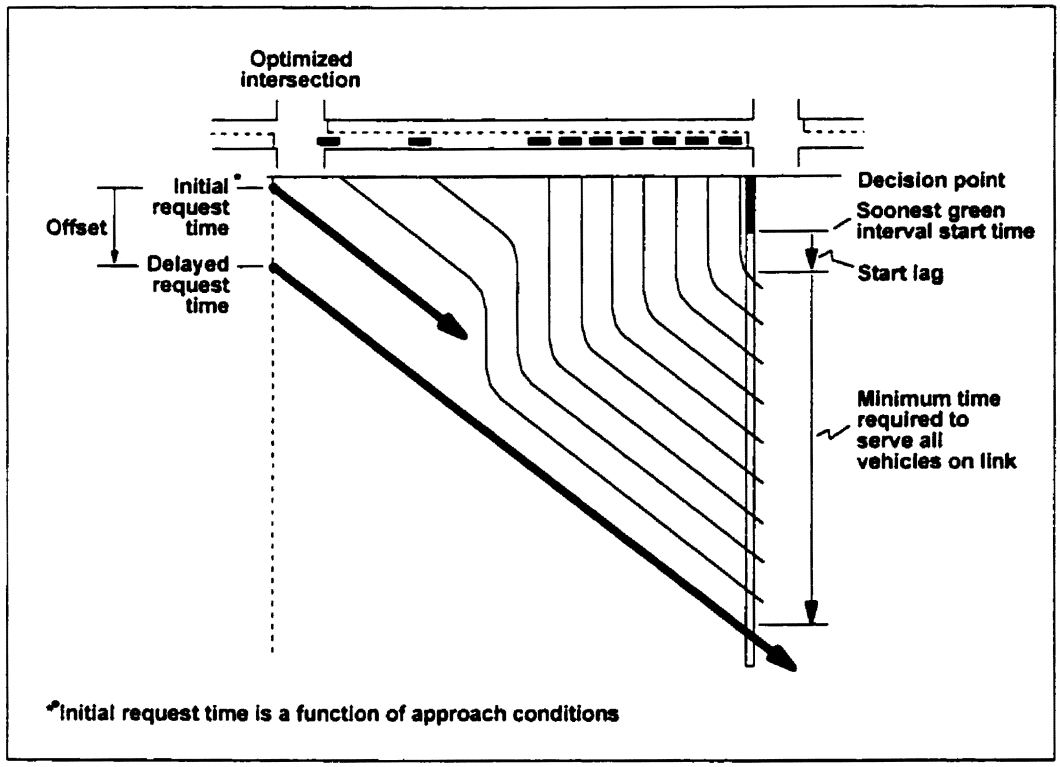


Figure 8.8 - Request Delaying in Response to Downstream Queuing

optimized. For each segment, the soonest time at which vehicles can enter at its upstream end and travel its full length without being forced to stop by other vehicles is computed using Equations 8.6 and 8.7:

$$\Delta RT_{queue\,m} = \frac{q_{veh\,m}}{q_{max\,m}} - \frac{L_m}{V_m} \qquad [8.6]$$

$$q_{max\,m} = min\left[q_{sat\,m}, q_{max\,m-1}\right] \qquad [8.7]$$

where: $\Delta RT_{queue\,m}$ = Change in ideal request time attributed to queuing conditions on segment *m* (seconds).

$q_{max\,m}$ = Maximum exit flow rate from segment m (passenger car units/second).

$q_{sat\,m}$ = Saturation flow of segment *m* (passenger car units/second).

$q_{veh\,m}$ = Number of vehicles on segment *m* (passenger car units).

L_m = Length of segment *m* (meters).

V_m = Average free-flow speed on segment *m* (meters/second).

The first term of Equation 8.6 reflects the shortest time needed to serve all vehicles currently present on the segment, and thus, the period during which any new vehicle reaching the stop line would be forced to stop before being able to exit the segment. The second term reflects the average time needed to travel along the segment. As shown in Equation 8.7, the parameter $q_{max\ m}$, which represents the maximum rate at which vehicles can exit segment *i*, is based on the maximum rate at which vehicles can flow across the segment ($q_{sat\ m}$) and the maximum rate at which these vehicles can enter the following segment when there are no temporary capacity reductions other than those caused by the signal operation ($q_{max\ m-1}$). The effects of temporary capacity reductions caused by transit vehicles dwelling in the right-of-way on queue dissipation time are considered separately in the next section. For all temporary capacity reductions caused by other traffic events, SPPORT relies on the automatic adjustment capacity of its rolling horizon process to update every few seconds its estimate of the time required to serve vehicles traveling on each segment based on the current maximum flow rates on each one of these links.

At the beginning of the calculation, the initial ideal request time is assumed to correspond to the soonest time at which vehicles can start to go across the downstream intersection. When the signal display at this intersection is currently red, this initial request time is based on the existing signal displays, signal timing and phasing constraints, and drivers' reaction time at the beginning of the next possible green interval. If the display is green, the soonest time at which vehicles can start to enter the downstream intersection is based on the last time a vehicle crossed the stop line and the minimum time headway at which vehicles can enter the intersection.

To obtain the ideal time at which a green interval should be initiated at an upstream intersection on the basis of downstream traffic conditions, $\Delta RT_{queue\ m}$, $q_{max\ m}$ and $RT_{queue\ m}$ are computed for each segment constituting a link using Equation 8.8, beginning at the the downstream intersection and moving upstream to the approach on which a green signal indication is requested.

$$RT_{queue\ m} = max\left[t_{\Delta t}\ ,\ RT_{queue\ m-1} + \Delta RT_{queue\ m}\right] \qquad [8.8]$$

where: $RT_{queue\ m}$ = Delayed request time after consideration of queuing conditions on segment *m* (seconds).

$t_{\Delta t}$ = Time of current decision point (seconds).

These calculations result in the ideal request time to be advanced or delayed after each segment depending on the number of vehicles present on the segment. This incremental adjustment allows SPPORT to truly evaluate how traffic conditions on downstream links will affect the progression of vehicles leaving the intersection for which signal-switching decisions are currently being made.

Figure 8.9 and 8.10 illustrate the application of these calculations to a link under two different queuing conditions. In the first example (Figure 8.9), the stop line queue at the downstream intersection dissipates completely before vehicles from the upstream queue reaches the stop line. As a result, only the dissipation time of the upstream queue affects the best time at which vehicles should be released from the upstream intersection to avoid any unnecessary stops and delays. The calculations performed by SPPORT over the entire length of the link correctly estimate the time when no queued vehicles will affect the progression of upstream traffic. The computed offset would be applied to any request generated at the current time. In the second example (Figure 8.10), both upstream and downstream queues affect the progression of upstream traffic. This can be observed by the fact that the ideal request time never comes back to the current decision time between the two queues. While this scenario represents a more complex condition, Figure 8.10 indicates that the proposed method correctly computes the appropriate offset.

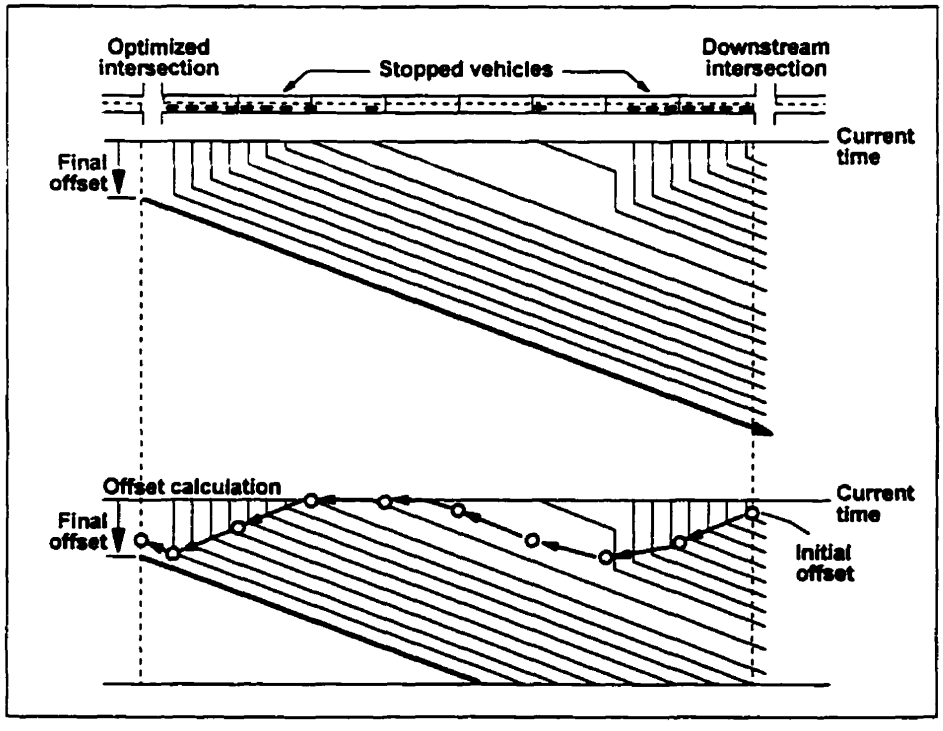


Figure 8.9 - Signal Offset Calculation Example 1

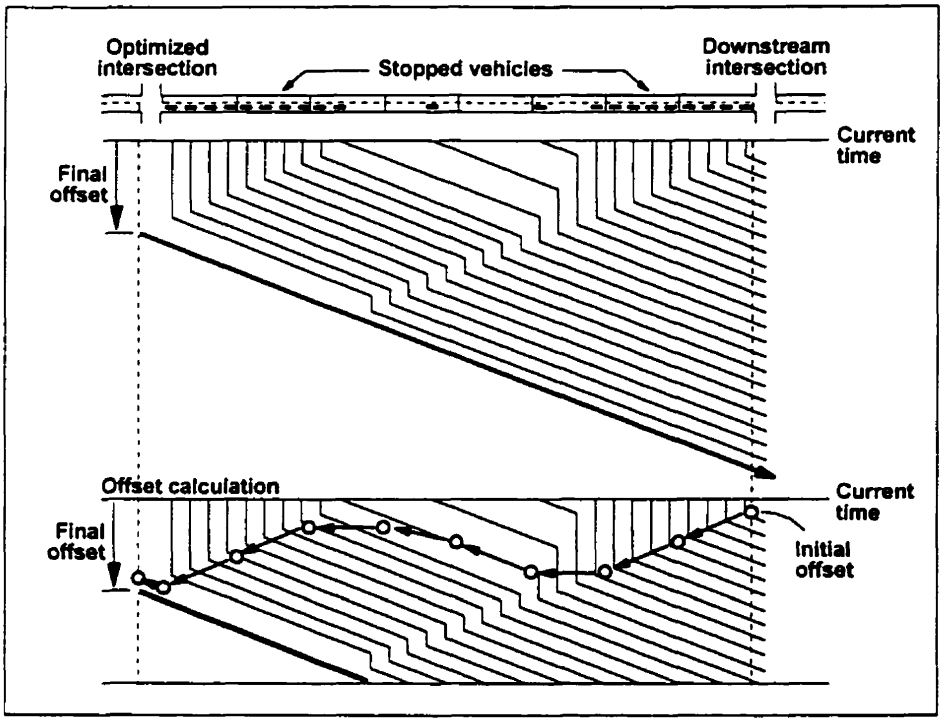


Figure 8.10 - Signal Offset Calculation Example 2

In the calculation of offset, it must also be observed that the maximum rate at which vehicles are assumed to be able to exit from a given segment ($q_{max\ m}$) corresponds to the lowest saturation flow rate of all the segments previously considered in the calculation (i.e., all downstream segments). This characteristic reflects the constraining effect that physical bottlenecks along intersection approaches have on the maximum upstream flow rate. For example, if the number of lanes on an approach drops from three to two at a given location, the maximum rate at which vehicles can flow across the section with three lanes is dictated by the rate at which vehicles that can pass through the two-lane section.

8.4.3. Coordination with Downstream Transit Activities

Another important element of the request delaying logic is the interference to the progression of general traffic caused by transit vehicles stopping in the right-of-way to board and discharge passengers. The primary goal of these adjustments is similar to the goal dealing with queuing conditions. This goal is to delay, by a minimum amount of time, the initiation of the green interval at an upstream intersection to ensure that vehicles released from that intersection would avoid the queuing or blockage caused on the downstream link by a transit vehicle dwelling in the right-of-way.

The extent to which a request is delayed is a function of the maximum rate at which vehicles can flow pass a dwelling transit vehicle. If the reduced rate is below a user-defined threshold, referred to as the *transit interference critical flow*, the request delaying calculations are carried out so that vehicles released from the upstream intersection will arrive at the transit stop just after the transit vehicle has left the stop. If the flow rate is greater than the user-defined threshold, the initiation of the green signal indication is delayed so that vehicles will reach the transit stop just after the queue that is expected to form behind the dwelling transit vehicle will have completely dissipated.

Equation 8.9 is used to calculate the additional amount of time that green signal indication requests should be delayed at an upstream intersection in order to consider the temporary capacity reductions caused by transit vehicles dwelling in the right-of-way.

$$\Delta RT_{transit} = t_{interf\ dwell}\left(1 - f_{transit}\frac{q_{sat\ m}}{q_{sat\ m-1}}\right) \qquad [8.9]$$

where: $\Delta RT_{transit}$ = Delay in request time accounting for transit interference (seconds).

$t_{interf\ dwell}$ = Interfering dwell time, as defined in Figures 8.11 and 8.12 (seconds).

$f_{transit}$ = Proportion of saturation flow available on segment *m* during dwell time.

$q_{sat\ m}$ = Saturation flow on transit stop segment *m* when no transit vehicle is dwelling (passenger car units/second).

$q_{sat\ m-1}$ = Saturation flow on segment *m-1* immediately upstream of transit stop segment *m* (passenger car units/second)

The request delay calculated by Equation 8.9 ($\Delta RT_{transit}$) corresponds to the additional time required to serve the vehicles upstream of the transit stop as a result of the temporary bottleneck created by a transit vehicle partially or completely blocking traffic during all or part of its dwell time. To illustrate, consider a queue of ten vehicles behind a loading transit vehicle. If there were no interference from the transit vehicle, only ten seconds would be required to serve the ten vehicles at a rate of one vehicle per second. However, 20 seconds would be required to dissipate the same queue if the transit vehicle temporarily reduces by half the maximum rate at which vehicle can pass the transit stop. In this case, ten seconds of flow capacity are lost during the dwelling process. This, therefore, causes SPPORT to delay all calls for a green signal indication at the upstream intersection by an additional ten seconds.

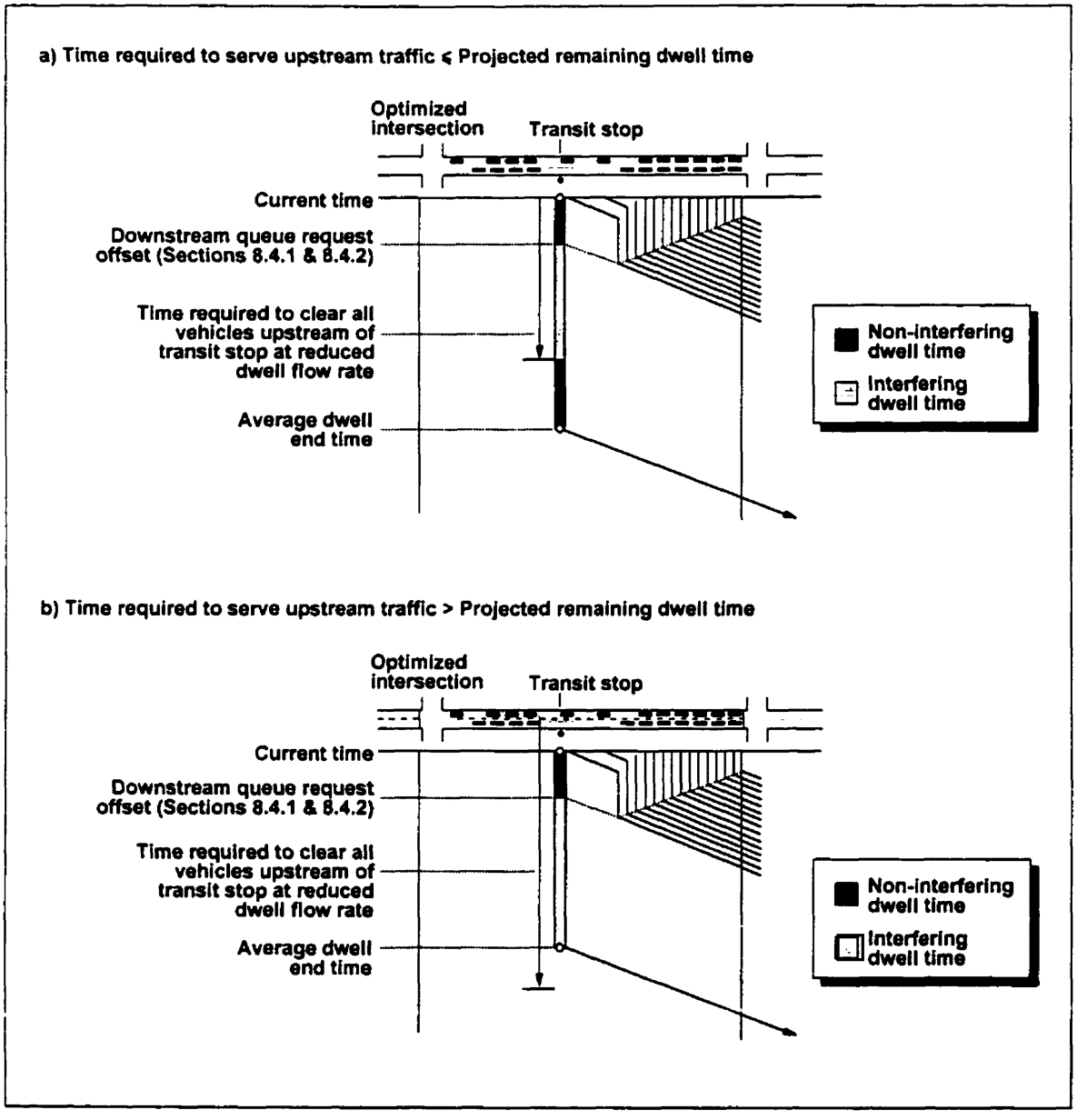


Figure 8.11 - Interfering Dwell Time on Passenger Car Traffic for Dwelling Transit Vehicles

In Equation 8.9, the interfering dwell time ($t_{interf\ dwell}$) is the portion of the total time spent by transit vehicles loading up and unloading passengers during which the vehicle truly interferes with the progression of upstream traffic. This distinction must be made, as the interfering dwell time is often less than the total observed dwell time. For example, transit vehicles stopped in the right-of-way do not interfere with the progression of other traffic if they are loading and unloading passengers while stopped within a queue of vehicles. From a coordination point of view, no significant interference also

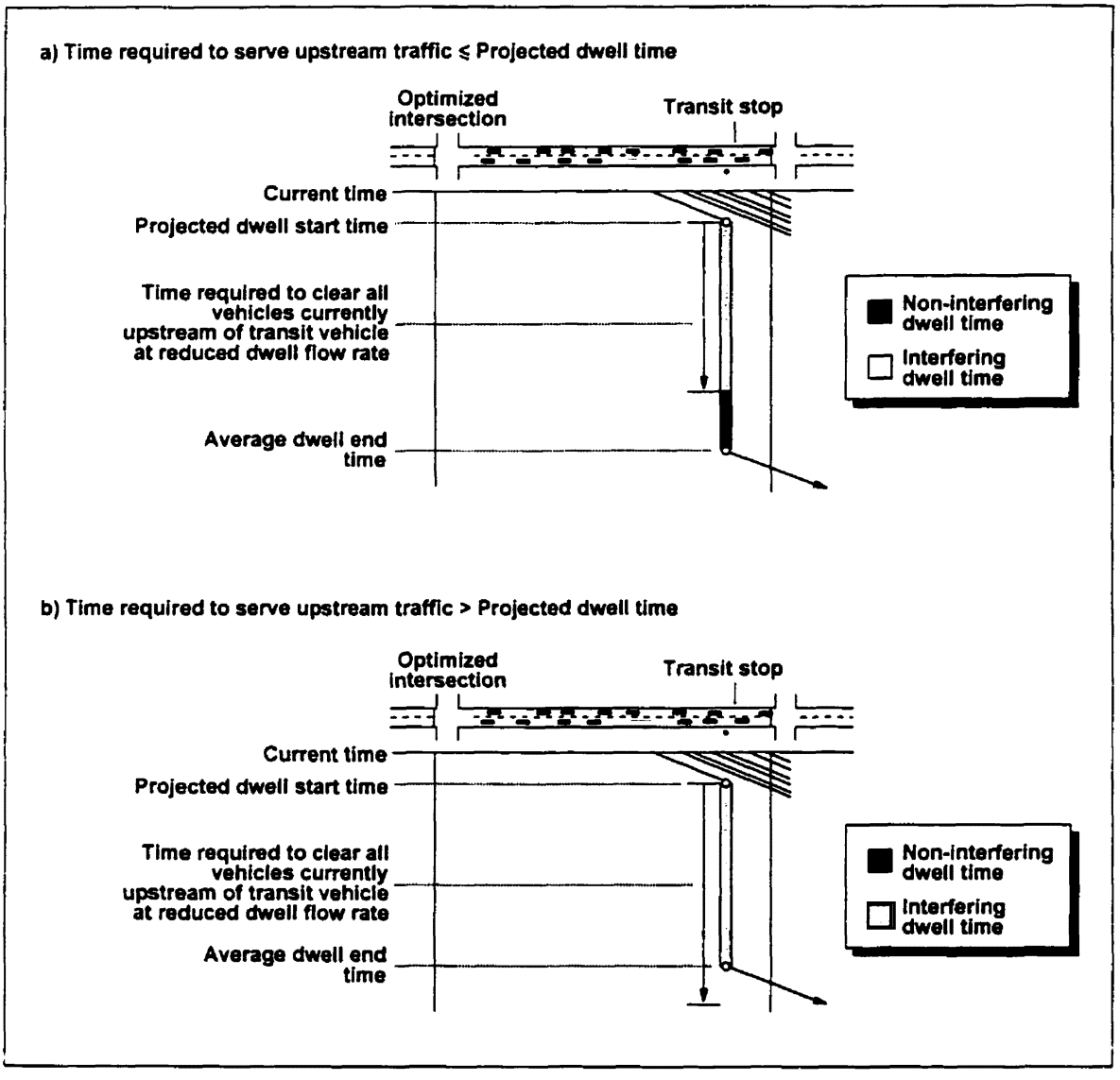


Figure 8.12 - Interfering Dwell Time on Passenger Car Traffic for Approaching Transit Vehicles

exists if the disruptions caused by a dwelling transit vehicle do not prevent the disrupted vehicles from joining the tail of an existing queue before they reach the downstream end of the link.

Figure 8.11 illustrates how the interfering dwell time is determined for the scenarios in which the transit vehicle causing interference is currently loading and unloading passengers. In both Figures 8.11a and 8.11b, the interfering dwell time starts at soonest time at which vehicles leaving the transit stop can expect to travel without interruption along the remaining portion of the link. This start time is determined using the procedures described in Sections 8.4.1 and 8.4.2. The end of the interfering dwell time depends on the whether or not the vehicles behind the transit vehicle can all be expected to

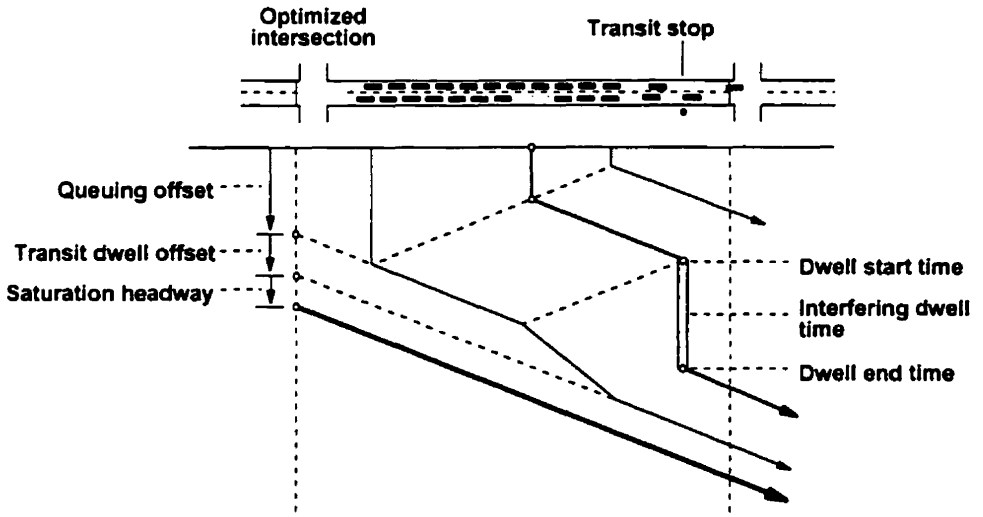
pass the transit stop before the end of the loading and unloading period. To determine which case prevails, the time that would be required to serve all the vehicles currently upstream of the transit stop under the reduced flow rate is determined. If the service time is less than the transit dwell time (Figure 8.11a), the interfering dwell corresponds to the service time. If the service time is greater than the transit dwell time (Figure 8.11b), it can be expected that the transit vehicles will affect traffic conditions until it departs from its loading point. In this case, the interfering dwell time corresponds to the remaining dwell time.

Figure 8.12 illustrates how the interfering dwell time is determined for scenarios in which an approaching transit vehicle has not yet reached its loading point. In this case, the interfering dwell time is assumed to start at the time the transit vehicle is projected to reach its loading point. Similarly to Figure 8.11, the end of the interfering dwell time is calculated by determining the time required for serving all the vehicles currently behind the transit vehicle. Depending on whether all these vehicles could be served before the transit vehicle finishes loading and unloading passengers, the interfering dwell time will either correspond to the calculated traffic service time (Figure 8.12a) or the expected average dwell time (Figure 8.12b).

In each case, it is assumed that transit vehicles will only remained stopped for a interval corresponding to their average observed dwell time. The exact time a transit vehicle is stopped while loading and unloading passengers cannot be directly used in the calculations, as the exact duration of this period is not known ahead of time. In order to provide automatic adjustments to longer or shorter dwell times, traffic detectors capable of selectively detecting transit vehicles could be installed just downstream of transit stops. By monitoring when transit vehicles cross these detectors, messages could be sent to the signal optimizers to inform them that a transit vehicle has left the transit stop earlier than expected, or that a transit vehicle is still loading after its expected departure time. Since traffic signals are re-optimized every five to ten seconds, this new information could then be used to make appropriate signal timing adjustments.

To illustrate the effect of the request delaying principles described in this section, Figure 8.13 presents two application examples dealing with a transit vehicle waiting in queue to reach its loading point. Figure 8.13a illustrates the case in which the time required to serve the traffic upstream of the transit vehicle under the reduced dwell time saturation flow is projected to be less than the time required by transit vehicle to board and discharge passengers. Figure 8.13b presents the opposite case. In both

a) Approaching transit vehicle within queue with projected dwell time greater than time required to serve vehicles upstream of transit vehicle



a) Approaching transit vehicle within queue with projected dwell time greater than time required to serve vehicles upstream of transit vehicle

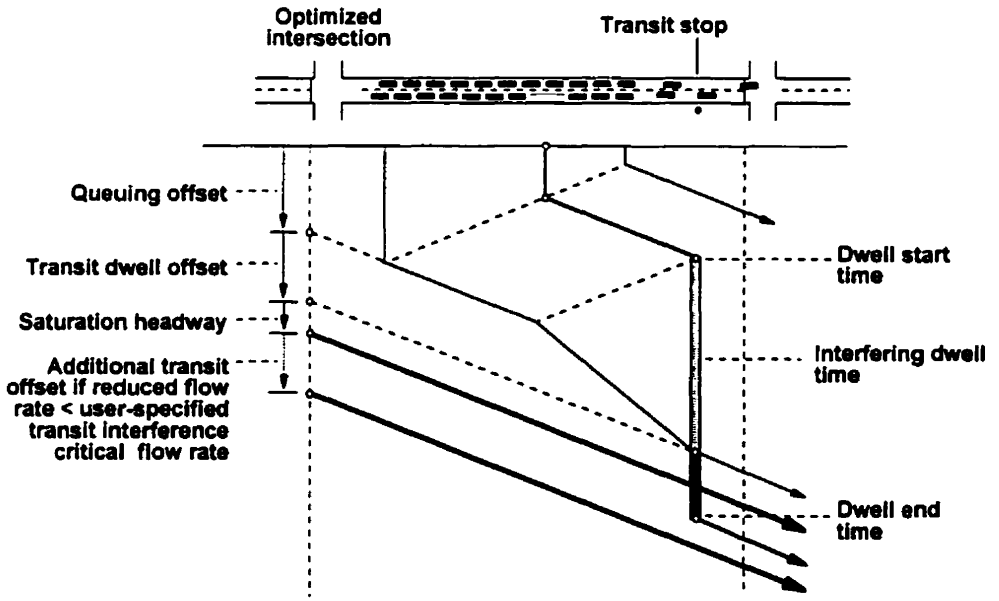


Figure 8.13 - Examples of Request Delaying in Response to Transit Activities

examples, the contribution of the queuing and transit offsets in the total delay imposed to the requests at the intersection being optimized are very apparent. While the queue offset deals with queue blockages, the transit offset deal with the additional constraints imposed by the dwelling transit vehicle. The example of Figure 8.13b also illustrates the request offset if the reduced flow rate during the transit dwell time is less than the user-defined transit interference critical flow rate.

8.4.4. Coordination with Upstream Timings

While the request delaying rules described in the three previous sections are designed to adjust the signal operation of a given intersection as a function of traffic conditions on downstream links, they do not attempt to promote coordination with upstream intersections. As a result, the application of these rules could lead to the establishment of very poor offsets with adjacent upstream intersections. To illustrate this point, consider the example in Figure 8.14. In this example, the presence of a large queue of vehicles on the downstream link B-C could cause all the requests calling for a green signal indication at the intersection B to be delayed in time. Such delays could then likely cause the initiation of the green interval at intersection B to be delayed. While a delayed green signal onset would allow vehicles to travel uninterrupted along the downstream link B-C, it would also force all the vehicles in the incoming platoon on the upstream link A-B to stop at intersection B when all these vehicles could have been allowed to proceed uninterrupted across intersection B and stored on the downstream link B-C. As a result, a good progression pattern between intersections A and B is broken on the sole basis that these vehicles could not achieve good progression between intersections B and C.

Another important problem with the above delaying logic is the possibility that it may cause undesirable control loops to occur. Referring again to the example of Figure 8.14, consider that the initiation of the green signal indication at intersection B is delayed by a few seconds to avoid unnecessary stops and delays on the downstream link B-C. If such change occurs, vehicles will then arrival later at intersection C. As a result, the optimizer in charge of intersection C might decide to change its projected signal control strategy and to delay the time at which a green signal indication would be displayed on link B-C. This change could in turn result in further delaying of the green signal indication requests and green interval onset at intersection B. At this point, the control loop causing the delaying of green intervals at both intersections B and C could start again if there is no criterion to stop it.

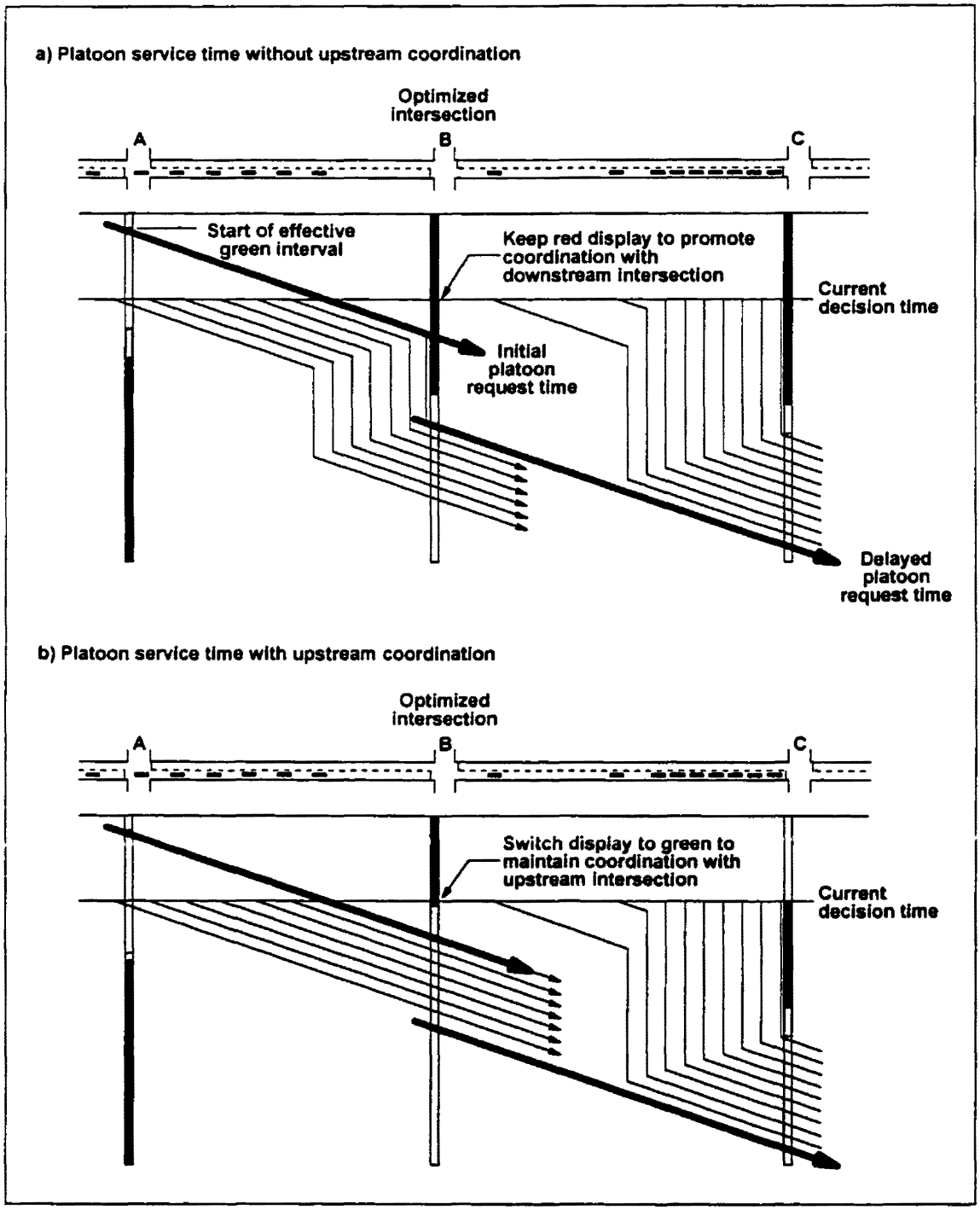


Figure 8.14 - Effect of Signal Coordination with Upstream Signalized Intersections

In order to preserve existing coordination patterns with upstream intersections and avoid falling into undesirable control loops, a limit is imposed on the maximum delay that can be imposed on a request on each approach link as a result of the application of the request delaying principles of Sections 8.41, 8.4.2 and 8.4.3. This limit is only applied on links for which the user has identified an approach link to the upstream intersection with which coordination should be attempted. The limit is also not fixed nor the same for every link. On each link, the amount of request delay allowed is a function of both the traffic conditions along the link and the signal timings at the intersection controlling traffic at its upstream end (or downstream end of the upstream link with which coordination is requested).

Figure 8.15 indicates how the maximum request delay is determined on each link. The figure distinguishes two main cases. Figure 8.15a illustrate the case in which an alignment of green intervals already exists between the intersection being optimized and the intersection at the upstream end of the approach under consideration. In this case, no request delay is allowed. Figure 8.15b illustrates the case in which no progressive pattern currently exists. In this case, requests can be delayed up to the furthest point in time at which a signal display switch from red to green can be implemented without interfering with the progression of the vehicles coming from the upstream intersection.

To identify the furthest point in time at which requests can be delayed, SPPORT first determines the time at which vehicles were or are expected to be released from the upstream intersection under the implemented and/or projected timings (Point #1 in Figure 8.15b). The time of arrival at the intersection being optimized of the first vehicle released from the upstream intersection is then determined (Point #2) assuming that the vehicle can travel without interruption along the full length of the link joining the two intersections. The furthest point in time to which a request can be delayed (Point #3) is finally determined by subtracting from the arrival time identified in the previous step the time required to serve all the vehicles currently on the approach to the optimized intersection. This subtraction is made with the objective of avoiding situations in which the arriving traffic from upstream intersections is forced to stop being a queue that has not finished to dissipate. In this calculation, the time required to serve all the vehicles on the approach is determined while assuming that these vehicles can cross the intersection stop line at the saturation flow rate.

As a final note, it must be mentioned that SPPORT does not attempt to modify the requests which initially call for a signal switch beyond the maximum delayed request time. This is to preserve the fact that these requests are associated with traffic events scheduled to occur farther in time.

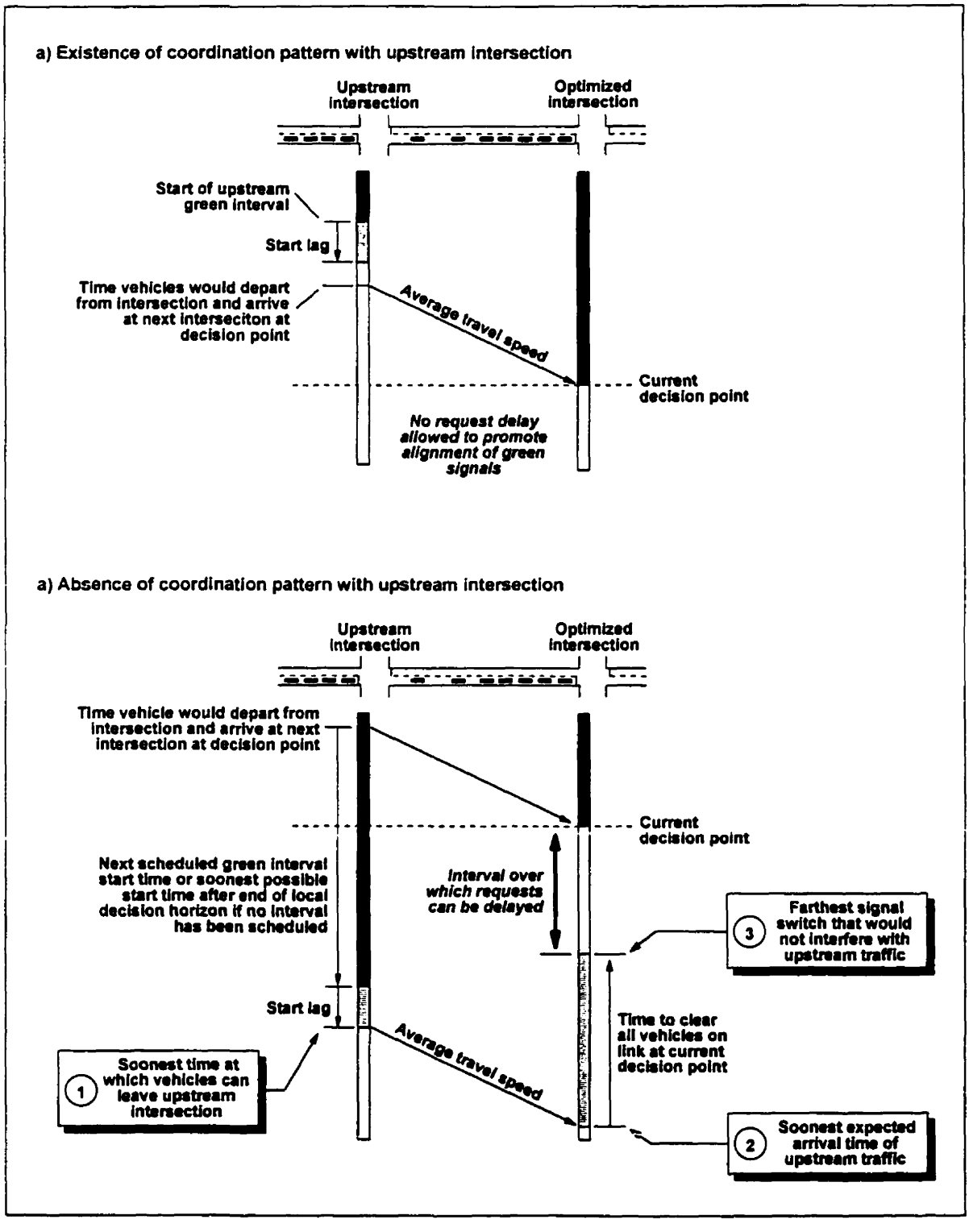


Figure 8.15 - Calculation of Maximum Permitted Request Delay

8.4.5. Coordination with Upstream Queue Spillbacks

Under normal traffic conditions, the objective of the signal optimization is typically to minimize driver's real and perceive delays. This is generally accomplished by minimizing the stops and delays incurred by vehicles at signalized intersections as a result of the signal operation. However, this control objective is generally not appropriate when one or more movements at the intersection being optimized become oversaturated. When queue threaten to spill across upstream intersections, it becomes more important to serve the existing queues to avoid the blockage of upstream intersections. As a result, the objective of minimizing stops and delays is often replaced with the objective of maximizing the number of vehicles that can cross the controlled intersection from one or more approach in a given amount of time.

In order to provide a similar shift in signal control strategy, special coordination rules were developed to allow SPPORT to reduce the scope of the coordination with downstream traffic conditions on intersection approaches on which a queue threaten to spill across the upstream intersection. Figure 8.16 illustrates how traffic conditions on an exit link can affect the application of the request delaying logic on intersection approaches in the presence and absence of queues on these approaches threatening to spill across the upstream intersection. In this Figure, intersection B is the one being optimized in all three examples.

Figure 8.16a illustrates a scenario in which there is not currently any danger of a queue on link A-B spilling back through intersection A. In this case, the main objective of the signal optimization remains the minimization of stops and delays on all approaches to intersection B. Consequently, the request delaying logic is applied without restriction on all approaches.

Figure 8.16b illustrates a scenario in which a longer queue exists on link A-B. In this case, the tail of the queue extends farther upstream than the user-defined critical queue reach. On the basis of the signal control rules defined in Section 8.1, this condition results in an immediate call for service to the link A-B to start dissipating the queue and avoid further problems around intersection A. This call is then subjected to the request delaying principles of Sections 8.4.1 to 8.4.4. In this case, a large delay in the initiation of the green interval on link A-B could result from the need to coordinate the signal operation with traffic conditions along the entire length of link B-C. Such delay could in turn cause the queue on link A-B to grow to such extent as to spill across intersection A and could therefore

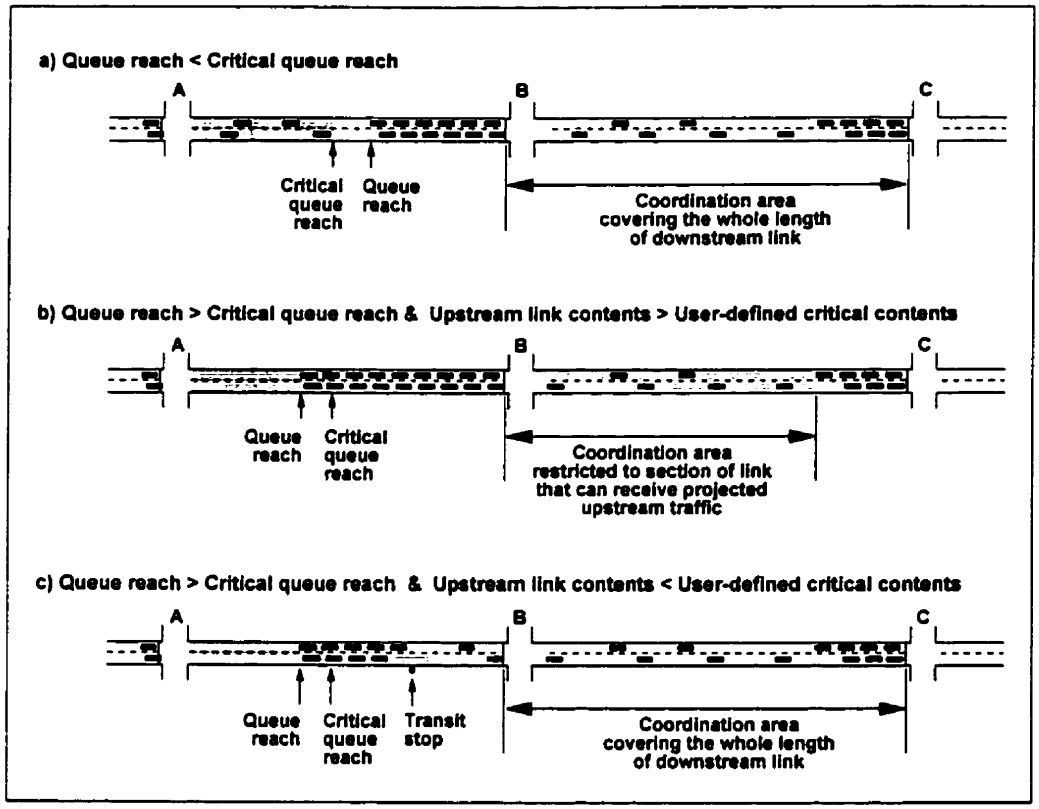


Figure 8.16 - Coordination Area on Intersection Exit Links within Request Delaying Logic

completely counteract the goal of generating a call for an immediate service to the link. In this situation, in order to delay the initiation of green signal indications as little as possible on link A-B and still attempt to minimize unnecessary stops and delays on the downstream link B-C, the request delaying only considers the traffic conditions on the upstream portion of link B-C that can accommodate the number of vehicles that each request is projecting to send across the intersection. For example, only the upstream portion of link B-C on which 10 additional vehicles can be stored would be considered by a request calling for an immediate green signal indication to serve a stop line queue of 10 vehicles.

Figure 8.16c illustrates a scenario in which the queue on link A-B exceeds the critical queue reach, but in which the discharge rate of vehicles from the link is controlled by something other than the signal (in this case, a transit vehicle stopped in the right-of-way). In this case, no special action is taken to reduce the scope of the coordination with link B-C. In SPPORT, an intersection approach link with a potential queue spillback problem (link A-B in this case) is only assumed to require an immediate green signal indication whenever the number of vehicles it currently contains exceeds half its total

queuing capacity or its queuing capacity downstream of the critical queue reach. The rationale for this criterion is that no immediate signal changes need to be made at an intersection if there is still enough queuing capacity available on its approach to receive, and store, incoming traffic from upstream intersections.

Table indicates for the scenario of Figure 8.16b the number of vehicles that each request is attempting to send across intersection B. These numbers define the demand for each request and help SPPORT in determining over which portion of link B-C traffic conditions should be considered when coordinating the signal operation of intersection B with downstream traffic conditions. Since these number represent the number of vehicles expecting to cross the intersection stop line, the number of vehicles expected to enter each exit link is obtained by multiplying the number given by Table 8.2 with the turning percentage between the approach and the exit links under consideration.

Table 8.2 - Evaluation of Traffic Demand Associated with each Request

Rule	Intersection Exit Link Demand
Begin Serving Excessive Wait	• Current stop line queue size
Begin Serving Queue	• Current stop line queue size
Begin Serving Upstream Spillback	• For existing spillbacks: Approach link contents currently exceeding the user-defined allowed contents; • For projected spillbacks: Approach link contents currently exceeding the user-defined allowed contents + vehicles projected to cross the stop line between the current time and the projected spillback occurrence.
Begin Serving Platoon	• Platoon size + number of vehicles projected to cross the stop line ahead of the first vehicle in the platoon
Continue Serving Queue	• Stop line queue size at the onset of a green interval minus the number of vehicles that have been projected cross the stop line since then.
Continue Serving Platoon	• Platoon size + number of vehicles projected to cross the stop line ahead of platoon
Continue Serving Upstream Spillback	• Projected stop line volume during the next five seconds
Continue Serving High Volume	• Projected stop line volume during the next five seconds
Bring Transit Vehicle to Transit Stop	• Total number of vehicles downstream of transit vehicle minus queue capacity downstream of the transit stop.
Bring Transit Vehicle to Stop Line	• Transit vehicle plus number of vehicles projected to cross the stop line ahead of it.
Do not Feed Downstream Spillback	• No demand, as this requests call for red signal display

Depending on the objective of each request, two main principles govern the determination of the demand associated with each request in Table 8.2:

- For requests calling for the continuation of the current phase, the demand corresponds to the minimum between the number of vehicles requesting service on the approach link and the maximum number of vehicles that can flow across the stop line between the current and next decision points (typically five seconds).
- For requests calling for the beginning of a new green interval, the demand corresponds to the minimum between the number of vehicles requesting service (number of vehicles in queue, in platoon, etc.) and the maximum number of vehicles that can flow across the stop line during a period corresponding to the longest minimum green interval of all the possible phases that can be implemented following the current one.

For each exit link, the portion of the link that can receive the projected traffic is determined by evaluating the storage capacity currently available downstream of the critical queue reach. The storage capacity upstream of the critical queue reach is not considered to avoid generating potential spillback conditions on the intersection exit link. On each link, the search for the coordination area to consider starts at the most upstream segment and ends on the first segment on which the condition of Equation 8.10 is met.

$$R_d \leq \sum_{i=1}^{N_{up\,seg\,jm}} (Q_{tot\,m} - q_{veh\,m}) \qquad [8.10]$$

where: R_d = Request demand (passenger car units).

$N_{up\,seg\,jm}$ = Number of segments between upstream end of link j and segment m.

$Q_{tot\,m}$ = Queuing capacity of segment m, taken to be zero if the segment is located upstream of the critical queue reach (passenger car units).

$q_{veh\,m}$ = Number of vehicles on segment m (passenger car units).

If there is not enough storage capacity currently available on the link being analyzed, the search algorithm automatically stops at the stop line of the downstream intersection. In this case, traffic conditions on the entire downstream link are be considered by the request delaying logic.

As a final note, it must be mentioned that it is possible that the calculation for the ideal request time does not start at the stop line of the downstream intersection. It is possible that the most downstream

segment included in the calculation is not the first segment upstream of the signal segment, but a segment located at some upstream distance from the intersection. When this situation occurs, the same calculations as described above are performed. The only difference is related to how the initial request time is determined. In this case, since there is no traffic signal at the exit of the most downstream segment included in the coordination area, the ideal request time is only a function of the soonest time a vehicle can enter the next segment as determined from the maximum rate at which vehicles can enter this segment assuming that no queue blocks its access.

8.5. Signal Optimization Process

Figure 8.17 illustrates the overall process followed by each signal optimizer to generate near-optimal timing plans for actual on-street implementations. This process can be summarized as follows:

1. At the beginning of the optimization period, detector and phase change information collected since the last phase plan had been generated are introduced into the intersection model tracking the state of the real-world intersection (*Model 1*).
2. The state of the intersection model is updated by simulating traffic behavior up to the current time.
3. A copy of the signal optimizer's intersection model is made for use in the signal optimization process (*Model 2*).
4. Projected arrivals at selected detector stations and projected timings from adjacent intersections are introduced into *Model 2* so that their effect can be included in future simulations.
5. Each phase plan generation module receives a new copy of the intersection modeling (*Model 3*).
6. Each phase plan generation module is prompted to generate a new phase plan using the following steps:
    - 6.1. All previous disutility measures recorded in *Model 3* are cleared (total travel time, total delay, total stops and terminal cost).
    - 6.2. Identification of the end of the decision horizon.
    - 6.3. Identification of the current decision point.
    - 6.4. Simulation of *Model 3* up the identified decision point.
    - 6.5. Determination of the end of the look-ahead period over which traffic demand will be evaluated.

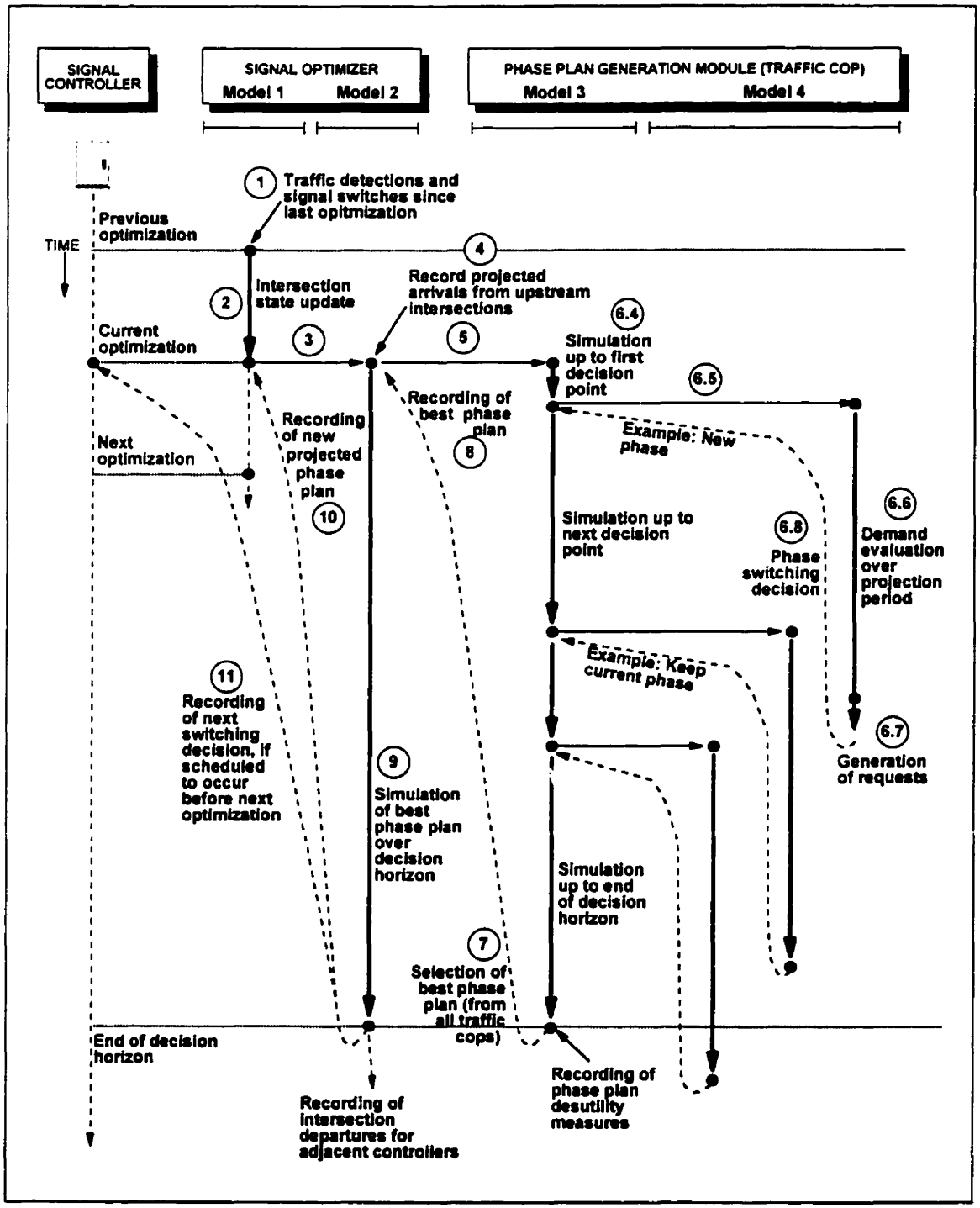


Figure 8.17 - Signal Optimization Process

6.6. Projection of future stop line and transit stop arrivals by simulating a copy of the intersection model received from the signal optimizer (*Model 4*) over the full duration of the look-ahead period assuming that the green signal indication can be returned as soon as possible on all links.

6.7. Using current queue sizes and projected stop line demands, generation of requests for every approach link in the intersection model.

6.8. Evaluation of requests and selection of best phase to implement at the current decision point based on the cost/benefit analysis described in section 8.5.

6.9. Registration into *Model 3* of any scheduled phase change.

6.10. Repetition of steps 6.3 to 6.9 until the end of the decision is reached.

7. Using the disutility measures compiled over the decision horizon by each phase plan generation module while simulating *Model 3 in steps 6.1 to 6.10*, the signal optimizer evaluates each candidate phase plan based on the objective function provided by the user and selects for implementation the one yielding the least overall cost.
8. The new best phase plan is stored into *Model 2*.
9. *Model 2* is simulated over the full duration of the decision horizon to record the projected intersection departures that will be later used by adjacent controllers.
10. The new best phase plan is stored into *Model 2*.
11. If the new best phase plan calls for a signal switch before the next optimization interval, appropriate signal changes are scheduled and sent to the signal controller for actual implementation.

9. Evaluation of the SPPORT Model

This chapter evaluates the ability of the SPPORT model to provide efficient and suitable real-time traffic signal control under various traffic conditions. The chapter begins by describing the test procedure that was used to simulate the operation of a SPPORT-controlled traffic signal control system. The next section describes the test scenarios that were developed to analyze the behavior of the model under a range of traffic conditions. The remainder of the chapter presents the results obtained from these tests. These results cover the following topics:

- Effectiveness of SPPORT in controlling traffic at isolated intersections;
- Effectiveness of SPPORT in controlling traffic along urban arterials;
- Effectiveness of signal coordination logic;
- Ability to provide effective transit priority on a conditional basis;
- Effectiveness of multi-objective signal optimization;
- Sensitivity of selected signal control parameters;
- Real-time applicability of the SPPORT model.

9.1. Special-Purpose Test Program

In order to test the SPPORT model, a special purpose computer model was developed. This model emulates the operation of a fully decentralized traffic control system and was constructed to allow the tests to be made in an environment in which traffic simulation and signal control parameters could be easily modified. Within the model, a user-configured traffic simulation model represents the real-world network, as shown in Figure 9.1. The role of this simulation model in the test procedure is twofold. First, it is to generate the traffic detector and phase change information that would normally be provided by real-world traffic detectors and traffic signal controllers. Second, it is to simulate the effects on real-world traffic of the signal-switching decisions selected for field implementation by SPPORT.

Along with the real-word simulation model, the test program creates a signal controller for each signalized intersection within the modeled traffic network. Depending on the setup parameters

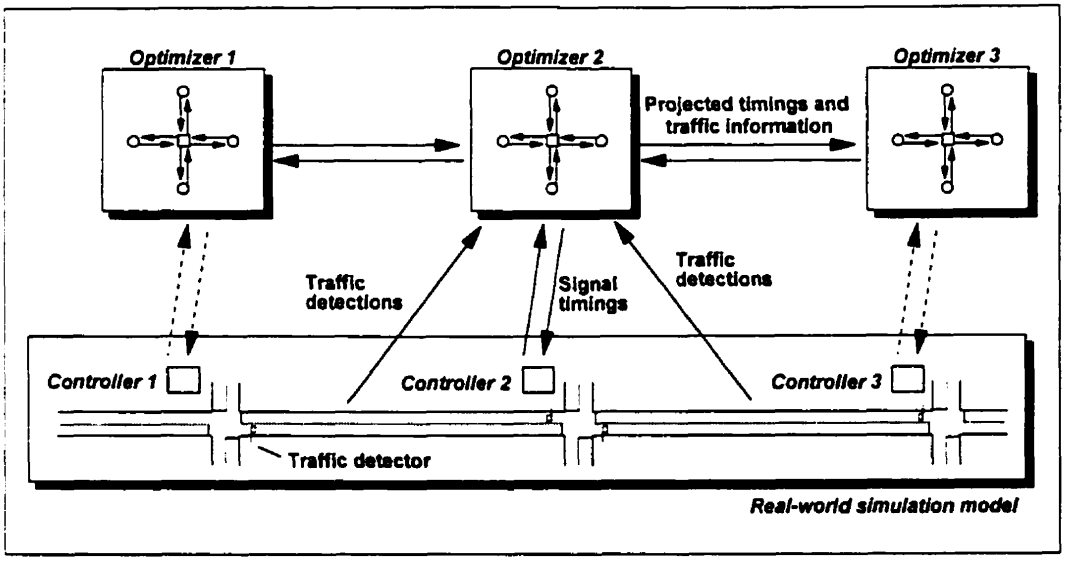


Figure 9.1 - Structure of Special-Purpose Test Program

provided by the user, each one of these controllers can either implement signal switches on a fixed-time or a real-time basis. In the fixed-time mode, signal switches are implemented on the basis of the cycle time, green allocation and offset information provided by the user. In the real-time mode, the controller receives signal control instructions from an associated signal optimizer every few seconds. This optimizer, which emulates the implementation of the SPPORT software in the controller, generates signal-switching decisions on the basis of the signal timing constraints and prioritized lists of events provided by the user. In this case, signal optimizations are performed using the traffic demand estimation and rule-base decision-making process described in Chapter 7 and Chapter 8 respectively.

In order to allow each signal optimizer to predict future traffic conditions and evaluate the potential impacts of candidate signal timing decisions, each optimizer is provided with a modeling of the links surrounding the intersection under their control. To perform the simulations, the same simulation model used to estimate real-world traffic behavior is used within the optimizer.

Ideally, different simulation models should have been used to simulate real-world traffic behavior and perform traffic projections within the signal optimizers. Several reasons can be identified that prevented the use of a different simulation model for representing reality. First, the use of the same simulation model ensures that the operation of the SPPORT model will not be affected by the reception of traffic information that does not perfectly describe real-world traffic conditions. This treatment is desirable in the fact that it allows a more objective evaluation of the abilities of the SPPORT model to

control urban traffic in real-time through the elimination of a certain number of variable elements affecting the operation of the signal optimizers. Improved evaluation then arises from the fact that eliminating variable elements that are external to SPPORT enhances the effects of the switching decisions taken by the signal optimizers. Second, it was required to use a simulation model capable of considering transit interference on other traffic. A survey of available simulation models indicated that there was no model capable of performing such task that could be easily integrated with SPPORT. For example, the TRAF-NETSIM microscopic traffic simulation model could have been used to simulate real-world traffic behavior and evaluate the SPPORT model. This model is currently often used to evaluate traffic signal control systems. However, to use it in a real-time control environment would have implied the development of a complex software interface to allow data transfer to occur between this model and the simulation model used by SPPORT.

It has been anticipated that a new traffic simulation model specifically designed for the evaluation of real-time, traffic responsive signal control systems, named RT-CORSIM (Federal Highway Administration, 1994, 1998; KLD Associates, 1994), would be used to evaluate the SPPORT model. This model, under development in the United States by the Federal Highway Administration, is a real-time extension of the CORSIM simulation model, which uses the NETSIM microscopic traffic simulation program to simulate traffic conditions along urban streets. It is designed to simulate changing traffic conditions and receive, through the development of proper interface procedures, the timings determined by an independent real-time, traffic responsive signal control model. However, delays in the development of an operational program prevented its use in the current research project. In fact, efforts are still under way to verify, validate, and provide calibration methods for the RT-CORSIM model (FHWA, 1998).

The final element of the computer program developed to evaluate the SPPORT model is the modeling of the communication system allowing individual controllers to obtain traffic and signal information from neighboring controllers. To model the communication system, the program assumes that each controller is directly linked with its immediate neighbors and that data queries at adjacent controllers can be performed directly by each controller without the help of a central computer. In the program, it is also assumed that all information exchanges occur at the end of each decision step, once all signal controllers have finished updating their control strategy.

9.2. Test Scenarios

This section describes the main characteristics of the test scenarios that were developed to test and evaluate the SPPORT model. The description is divided into three main sections. The first section presents the networks that were modeled to conduct the evaluations. The second section describes the traffic demand scenarios assigned to each one of the test networks. The last section describes the main parameters directing the operation of the SPPORT model in all scenarios.

9.2.1. Test Networks

Two test configurations were developed to evaluate the behavior of the SPPORT model in different control environments. The first configuration, illustrated in Figure 9.2, represents a single urban intersection, while the second configuration, illustrated in Figure 9.3, represents a five-intersection arterial with closely spaced intersections.

These two test networks were developed with specific objectives in mind. The isolated intersection test network was developed with the objective of validating the ability of the SPPORT model to

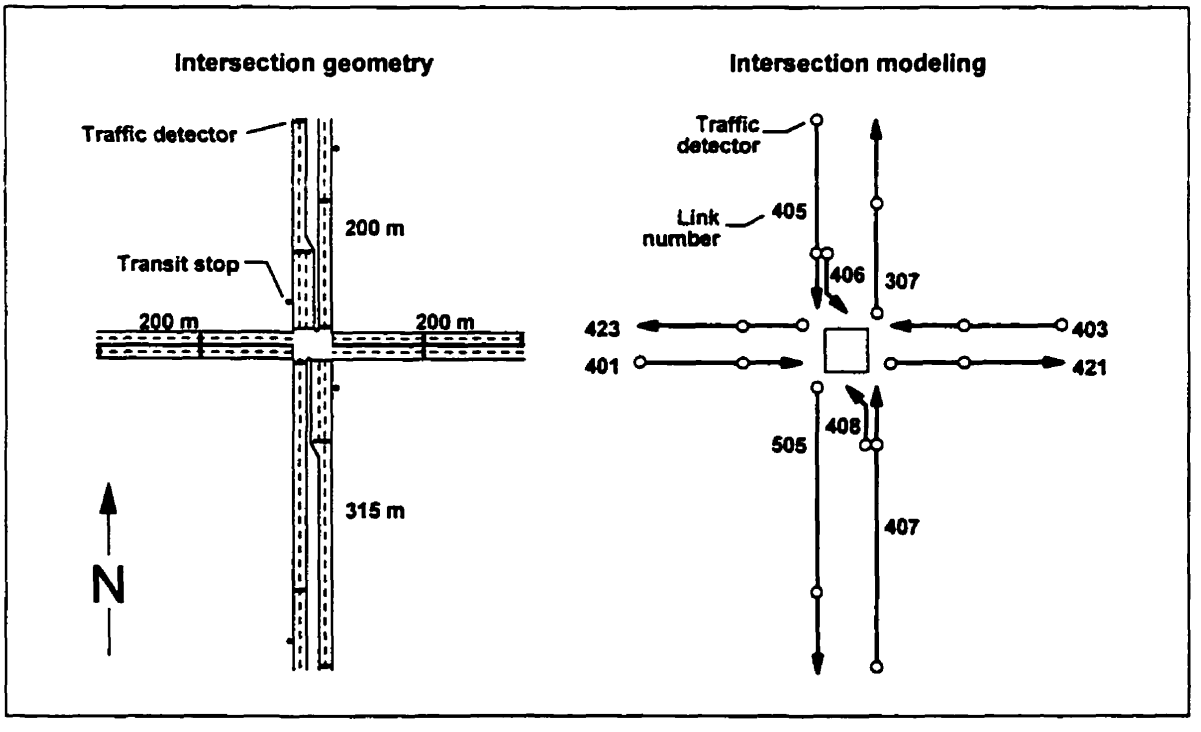


Figure 9.2 - Isolated Intersection Test Network Configuration

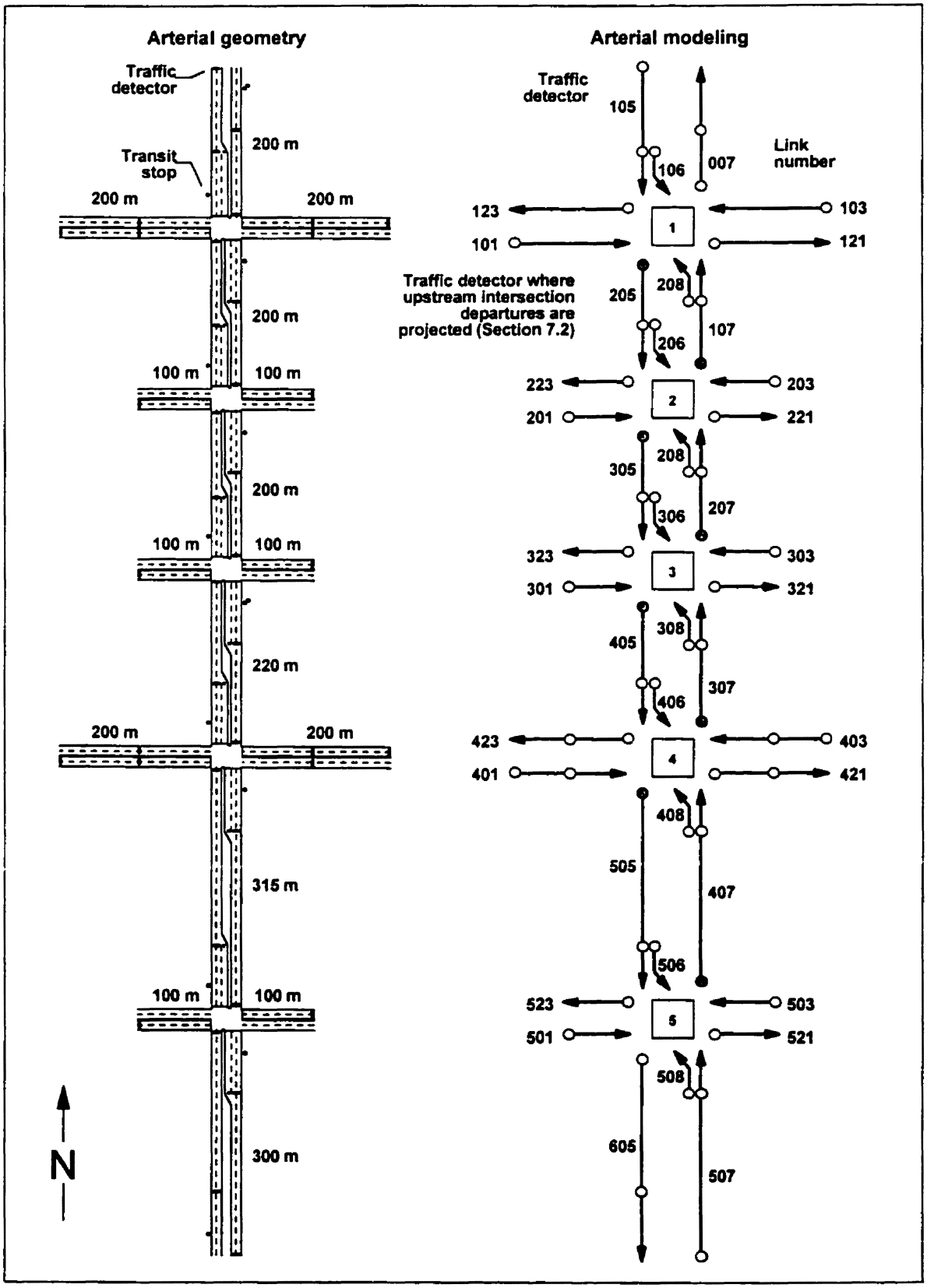


Figure 9.3 - Arterial Test Network Configuration

respond to occurring traffic events using its rule-based decision-making process without considering the additional challenge of providing coordination. The arterial test network was developed with the objective of evaluating the effectiveness of the signal coordination logic introduced in the model as part of this thesis.

9.2.2. Traffic Demand Scenarios

On the basis of the above two networks, 24 test scenarios were developed. Table 9.1 describes the main characteristic of each of these scenarios. The 24 scenarios are differentiated on the basis of four characteristics, namely whether or not the demand is temporally constant, the magnitude of traffic demand, whether or not transit priority is provided, and the network used.

Two types of traffic demands are simulated: constant demand and peaking demand. In fixed-demand scenarios, the average rates at which vehicles are simulated to enter the control area from each entry link remain constant over time. This is the type of traffic demand usually assumed to exist by fixed-time signal optimization methods. In the peaking-demand scenarios, the rates at which vehicles are

Table 9.1 - Test Scenarios

Scenario	Peaking demand	Demand level	Transit priority	Network configuration
1	No	Low	No	Isolated
1p	Yes			
2	No	Medium		
2p	Yes			
3	No	High		
3p	Yes			
4	No	Low	Yes	
4p	Yes			
5	No	Medium		
5p	Yes			
6	No	High		
6p	Yes			
7	No	Low	No	Arterial
7p	Yes			
8	No	Medium		
8p	Yes			
9	No	High		
9p	Yes			
10	No	Low	Yes	
10p	Yes			
11	No	Medium		
11p	Yes			
12	No	High		
12p	Yes			

assumed to enter the control area at the northern and southern boundary of the network vary over time, while the arrival rates at the eastern and western boundaries remain constant. Figure 9.4 illustrates the variation in traffic demands over a one-hour control period. Rates are expressed in relation to the overall average arrival rate to indicate the fact that the overall one-hour demand for all peaking scenarios is the same as the overall demand in the corresponding fixed-demand scenario. The same overall demands are used to allow comparisons to be made between test runs based on each type of traffic demand pattern.

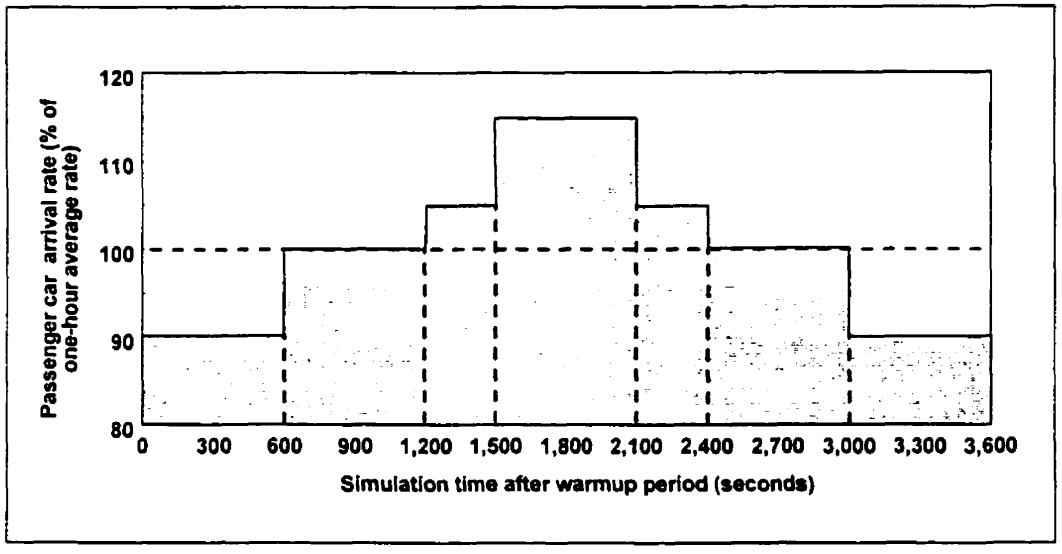


Figure 9.4 - Temporal Variations in Traffic Demand for Peaking Demand Scenarios

For each scenario, three levels of traffic demands are defined: low, medium and high. In the low demand scenarios, arrival rates of passenger cars are 25 percent lower than in the medium demand scenarios. Similarly, there is a 25 percent increase in traffic arrival between the medium and high demand scenarios. In all cases, however, transit arrival rates remain constant. These various scenarios allow SPPORT to be evaluated under a fair range of traffic conditions. The low demand scenarios allow the effectiveness of the model to be tested in situations in which minimum green interval constraints greatly affect the signal operation while the high demand scenarios allow evaluations to be made in networks operating at or near capacity.

In each simulation, the only source of randomness is the path chosen by individual vehicles at points along intersection approaches where traffic splits into through and turning movements. For example, on southbound and northbound approaches of all isolated intersection and arterial scenarios, it is

assumed that five percent of the arriving traffic chooses to turn left at the entrance of the left turn bay, and that five percent of the remaining traffic (4.75 percent of total arriving traffic) chooses to turn right at the stop line. These selections are randomly made according to a probability distribution reflecting the turning percentage provided. All other simulation parameters, such as vehicle inter-arrival times, vehicle speeds, and transit dwell times are all assumed constant.

To evaluate the degree of variability associated with the test simulations that would be conducted to evaluate the SPPORT model, 20 replications were made of a typical fixed-time and a typical SPPORT traffic signal operation using different seed numbers for the random process determining the path of individual vehicles across controlled intersections. Figure 9.5 illustrates the resulting observed variations to the total person delay. While these results do not conclusively quantify the within model variance under all conditions, they do indicate that the within model variance is quite small with respect to the mean. Based on these results, a high degree of confidence can be placed on the simulation results. In this case, what is of most interest is the coefficient of variation, which expresses the magnitude of the standard deviation in relation to the magnitude mean. As it can be observed, both the fixed-time and SPPORT operations showed in this case a very low coefficient of variation.

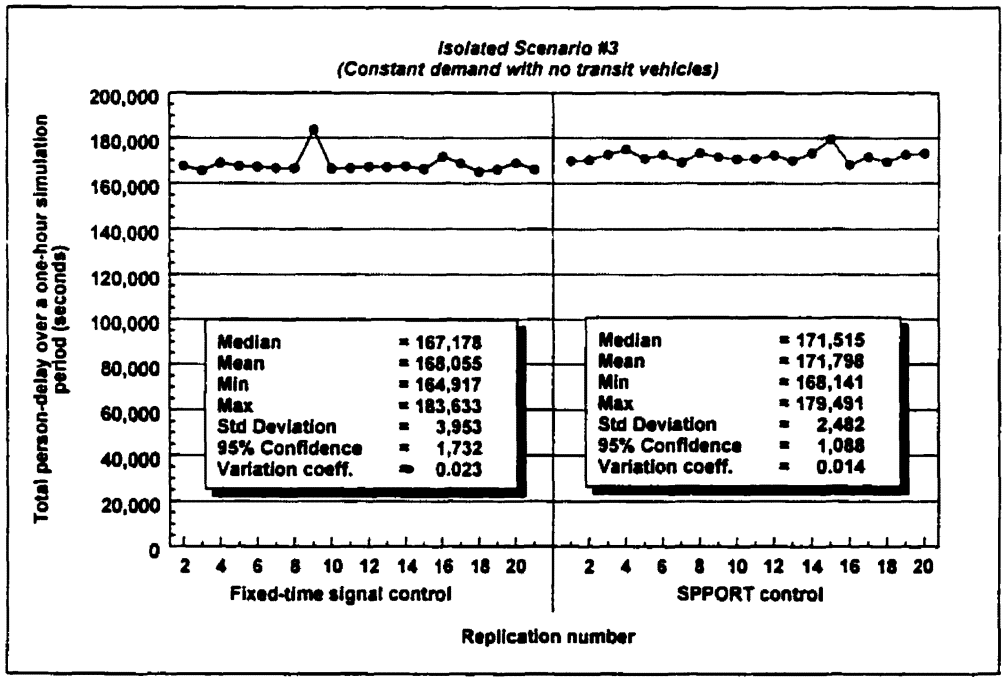


Figure 9.5 - Variance of Simulated Fixed-Time and SPPORT Operations over 20 Replications

9.2.3. Signal Control Parameters

In each scenario, the three following phases are defined:

- Phase 1: Protected left turn interval, serving northbound and southbound left-turners;
- Phase 2: Main-street green interval, serving all northbound and southbound traffic;
- Phase 3: Cross-street green interval, serving all eastbound and westbound traffic.

While the SPPORT model has the ability to choose which phase should follow the current one when given a set of possible choices, this feature is not used in the evaluations conducted in this thesis. The phase selection feature is not considered on the basis of a desire to determine the range of potential benefits that can be obtained from the SPPORT model in an incremental manner. Before evaluating the benefits provided by the simultaneous use of all the features of the model, the benefits provided by each individual component must first be assessed and understood. This means starting the evaluation with a simple control structure and progressively evaluating the benefits of increasingly complex structures. For this thesis, it has been determined that evaluating the benefits of using variable phase sequences was outside the scope of the current research project. Consequently, in all scenarios signal controllers are instructed to follow a fixed sequence of phases. The sequence starts with Phase 1 described above, and then respectively follows up with Phase 2 and Phase 3, before returning to Phase 1 and starting a new signal cycle.

As a result of the above choice, the rule-based decision-making process implemented in SPPORT is only used to determine the best duration of each green interval in the imposed sequence on the basis of current traffic conditions and given signal control parameters. While imposing a fixed sequence of phases restricts the freedom of decision of signal optimizers, this operation is also more representative of the type of signal control preferred by traffic engineers. Fixed phasing sequences, with only possibilities for skipping phases for which there is currently no demand, are often preferred since they create repetitive signal patterns and give drivers the impression that the traffic signal system operates correctly. In many cases, knowing which phase comes next gives more patience to the drivers to wait for the desired phase and leads to reductions in the number of complaints issued about the control system. In addition, the use of fixed phase sequences simplifies coordination with adjacent intersections by increasing the predictability of the signal operation.

The information presented in Table 9.2 describes the different *traffic cops* used by SPPORT to perform the signal optimizations. The traffic cops are the objects within the signal optimization module holding the prioritized lists of events and generating phase plans on the basis of these lists. A traffic cop is generated by SPPORT for each prioritized list of events provided by the user. In total, six different cops were defined. They differ from each other in the way that incoming platoons and transit vehicles are handled on individual approaches. The first two traffic cops, labeled *1np* and *1tp*, consider approaching platoons from the northbound approaches only. The next two cops, labeled *2np* and *2tp*, consider approaching platoons from the southbound approach only. The last two cops, *3np* and *3tp*, consider incoming platoons on both the northbound and southbound approaches. The labels *1tp*, *2tp*, and *3tp* refer to cops providing priority to approaching transit vehicles on both the northbound or southbound approaches, while the other cops (*1np*, *2np*, and *3np*) do not provide transit priority.

Depending on the scenario being simulated, different combinations of traffic cops are used to generate signal-switching decisions. In the scenarios with no transit vehicles, only the traffic cops labeled *1np*, *2np* and *3np* are used. In the scenarios in which transit activities are simulated, the traffic cops that provide transit priority are included along with the cops that do not provide transit priority, which results in the use of a total of six traffic cops.

In each traffic cop, constant priority values are assigned to the rules directing their operation, except for the *Begin Serving Platoon*, *Begin Serving Queue*, and *Continue Serving Queue* rules. For the *Begin Serving Platoon*, the priority level assigned to the requests generated by this rule is a function of the size of the incoming platoon for which service is requested. For the *Begin Serving Queue* requests, the priority level is a function of the size of the existing stop line queue, while the stop line queue that existed at the green interval onset is used to determine the priority level of the requests generated by the *Continue Serving Queue* rule.

These variable priorities are assigned to reflect the greater importance of serving larger queues and to allow large platoons to proceed uninterrupted across the controlled intersection. For example, if it is considered that a minimum of 10 vehicles constitutes a valid platoon for signal control purposes, a request generated by the *Begin Serving Platoon* rule will never be assigned a priority of less than 4.0. However, based on the limit specified in Table 9.2, the priority level assigned to these requests will also never exceed 40. For the *Begin Serving Queue* rule, each vehicle in queue at the stop line

Table 9.2 - Rule Control Parameters for Test Scenarios

Rules described in Chapter 8	Priority Level	Allowable delay (sec)	Request Delaying allowed	Traffic Cop					
				1np	1tp	2np	2tp	3np	3tp
Do not Feed Downstream Spillback	300 - 750 [1]	0	---	All	All	All	All	All	All
Begin Serving Upstream Spillback	200 - 500 [2]	0	Yes	All	All	All	All	All	All
Continue Serving Upstream Spillback	200 - 500 [2]	0	No	All	All	All	All	All	All
Bring Transit Vehicle to Transit Stop	100	5	No	--	NS	--	NS	--	NS
Bring Transit Vehicle to Stop line	100	5	No	--	NS	--	NS	--	NS
Begin Serving Excessive Wait	50	0	Yes	All	All	All	All	All	All
Continue Serving High Volume	40	0	No	XL	XL	XL	XL	XL	XL
Continue Serving Platoon	40	0	No	N	N	S	S	NS	NS
Begin Serving Platoon	0.04 * (Platoon Size)^2 Maximum: 40	5	Yes	N	N	S	S	NS	NS
Continue Serving Queue	0.03 * (Onset Queue Size)^2 Maximum: 30	0	No	All	All	All	All	All	All
Begin Serving Queue	0.01 * (Queue Size)^2 Maximum: 20	0	Yes	All	All	All	All	All	All

Notes: N = Northbound approaches only
S = Southbound approaches only
NS = Both northbound and southbound approaches
All = All approach links
XL = All approach link, except left turn bays.
[1] 750 on approaches on which platoons are considered, 300 otherwise.
[2] 500 on approaches on which platoons are considered, 200 otherwise.

increases the priority level of the corresponding request by 0.03, while the maximum allowed priority level is 30. For the *Continue Serving Queue* rule, priority levels are only allowed to vary between 0.01 and 20.

In Table 9.2, an *allowable delay* of zero second is assigned to each rule, except for those dealing with incoming platoons and transit vehicles. For these rules, an allowable delay of five seconds is used. This value indicates to the signal optimizers that is it acceptable to select a phase that may prevent the corresponding requests to be served on time, but only if service to the incoming platoon or transit vehicle will not be delayed at the intersection by more than five seconds. This provides the signal optimizers with an additional margin of maneuver by allowing them to implement signal changes that may have been otherwise rejected on the basis of only one or two seconds of delay caused to future incoming vehicles for which predicted stop line arrival times are still subject to variations.

The column labeled *Request delaying allowed* indicates whether the requests generated by the associated rule can have their call time modified by the request delaying logic of Section 8.4. In table 9.2, it can be observed that only the requests calling for the beginning of a new green interval can be delayed. The requests calling for a continuation of an existing phase cannot be delayed, as a delaying may result in the premature ending of the phase. Requests calling for the beginning of a red interval are also not subjected to the request delaying logic. In this case, the reason for not delaying the requests is associated with the fact that these requests are generated to stop flows of vehicles that might cause downstream queues to spill across the controlled intersection. The delaying of the requests would then contradict the objective of the rule by allowing vehicles to continue to cross the intersection despite the threat of downstream queue spillbacks.

Tables 9.3 to 9.6 list the main traffic signal control parameters used by SPPORT in its signal optimization process:

- Table 9.3 lists the general parameters constraining the operation of the traffic signals at individual intersections and defining the portion of the green interval truly used by vehicles. These parameters are the allowed minimum and maximum duration for each green interval, the duration of all amber intervals and all-red periods, and the amount of time lost at the beginning and end of every green interval. They also define the effective green interval of each phase. This interval corresponds to total duration of the scheduled green and amber intervals, minus the start-up and clearance lost times.

Table 9.3 - Signal Timing Parameters

Phase	Minimum Green (sec)	Maximum Green (sec)	Amber Interval (sec)	All Red Period (sec)	Start-up lost time (sec)	Clearance lost time (sec)
1 (Protected left -turn)	5	10	2	0	2	0
2 (Main green)	10	60 / 90[1]	5	0	2	2
3 (Cross-street green)	10	60	5	0	2	2

(1) 60 for low and medium demands, 90 for high demand

Table 9.4 - Link Control Parameters

Link Parameter	Value
Queue threshold	1 vehicle
Platoon threshold	10 vehicles
Critical queue reach	75% of link length
Critical exit flow	66.7% of link saturation flow
Transit interference critical flow	45% of link saturation flow

Table 9.5 - Coordinated Links for Arterial Test Configuration

Link	Coordinated Links	
	Upstream	Downstream
Southbound direction		
105	---	205
205	105	305
305	205	405
405	305	505
505	405	605
605	505	---
Northbound direction		
507	---	407
407	507	307
307	407	207
207	307	107
107	207	007
007	107	---
Eastbound direction		
No coordination		
Westbound direction		
No coordination		

Table 9.6 - Signal Optimization Parameters

Optimization Parameter	Value
Decision horizon	60 seconds
Decision interval	5 seconds
Commitment period	5 seconds
Look-ahead period	45 seconds

(See Figures 5.7 and 7.4)

- Table 9.4 presents the parameters used by SPPORT to control traffic conditions along intersections during a signal optimization. These parameters include the minimum number of queued vehicles and platoon size for which green signal indication requests can be generated, the *critical queue reach*, the *critical exit flow rate* assigned to every link, and the *transit interference critical flow* (see Sections 8.1.1, 8.1.3 and 8.4.3).
- Table 9.5 indicates with which upstream and downstream links each intersection approach should be coordinated within the arterial test scenarios. The numbers that are indicated in the table are the link numbers shown in Figure 9.3. As it can be observed, signal coordination is only attempted for the arterial links.
- Table 9.6 presents the parameters defining the operation of the SPPORT model. The information listed in the table indicates that in each scenario signal timings are to be generated for the next 60 seconds of signal operation, with signal switches possible every five seconds only. In this process, signal-switching decisions at each decision point within the decision horizon are to be based on 45-second traffic projections. The information presented in the table also indicates that for each 60-second timing plan generated, SPPORT only commits to the decisions taken in the first five seconds.

Further information regarding the traffic simulation parameters used in the various test scenarios can be found in Appendices B and C.

9.3. Effectiveness of SPPORT Signal Control at Isolated Intersections

To evaluate the ability of the SPPORT model to effectively control traffic in real time, the performance of the model is first compared against the performance of a fixed-time traffic signal operation at an isolated intersection. To perform the test, the delay minimization principles that were established by Webster (1959) are used to determine the optimal fixed signal timings. These principles are currently the most widely used by traffic engineers to perform signal optimizations for isolated intersections. They state that the total delay incurred by vehicle passengers at a single intersection can be minimized by allocating the total available effective green time during each signal cycle in proportion to the demand for each phase, where the demand is determined by the highest volume to saturation flow ratio of all the traffic movements simultaneously served by each phase. A more detailed presentation of Webster's delay minimization principles is provided in Appendix D.

Since an optimization following Webster's delay minimization principle primarily attempts to minimize the delays incurred by vehicles at signalized intersections, a similar type of optimization must be used within SPPORT to maintain a common base of comparison. To fulfill this objective, Table 9.7 lists the values that were attributed to the various parameters of Equations 5.1 and 5.2 with the objective of defining a performance function based on delay only. In this table, a value of 0.0 is assigned to both the stop coefficient k_s and travel time coefficient k_{TT} to cancel the effects of the estimated number of stops and estimated total travel time within the controlled network in the performance function used by SPPORT. A value of 1.0 is maintained for both the delay coefficient k_d and terminal cost coefficient k_{TC}. The terminal cost is kept in the performance function, as this element attempts to evaluate the total delay incurred by all vehicles beyond the end of the decision horizon as a result of the signal-switching decisions taken during the horizon. In this case, a value of 0.0 is assigned to the red signal display coefficient α_R of each approach link to indicate that only the vehicles left waiting in front of a red signal at the end of the decision horizon should be considered in the evaluation.

Table 9.7 - Performance Function Parameters for the Isolated Intersection Scenarios

Parameter (Equations 5.1 and 5.2)	Value
k_d	1.0
k_s	0.0
k_{TT}	0.0
k_{TC}	1.0
α_R	0.0
$o_{v\ private}$	1.5
$o_{v\ transit}$	60.0

To reflect differences in occupancy between various types of vehicles, a weighting coefficient of 1.5 is assigned to all passenger cars, while a coefficient of 60 is assigned to all transit vehicles. These values allow SPPORT to compile performance measures on a person basis and to provide transit priority on the basis that one transit stop is equivalent to 40 passenger car stops. While Webster's delay minimization principles only consider vehicles, and not vehicle passengers, the base of comparison between Webster's timings and SPPORT signal operation is maintained for all the scenarios in which no transit vehicles are simulated. Since these scenarios only consider one type of vehicles, all vehicles thus have the same weight with respect to the signal optimization, which is equivalent to providing a weighting coefficient of 1.0. For the scenario involving transit priority, a perfect common base of comparison cannot be maintained since Webster's delay minimization principles do not consider transit

priority. Person-based statistics must be used to evaluate the effectiveness of the transit priority treatments implemented by SPPORT since the implementation of these treatments schemes is based on the recognition that transit vehicles weight more than passenger cars.

Theoretically, the cycle time and green allocation obtained with the equations derived by Webster should minimize the delays caused by the signal operation. However, the optimality of Webster's timings is not guaranteed in the tests conducted in this thesis due to the use of a discrete-event simulation model. While Webster's delay minimization equations are based on a continuous mathematical formulation of the vehicle arrival and departure processes that take place at the stop line of each approach, the simulation model used in this thesis to evaluate the various signal control strategies uses a discrete formulation of the same processes. While Webster's formulation assumes that it is possible for fractions of vehicles to enter an intersection when there is not enough time to allow a complete vehicle to do so, vehicles only enter an intersection in the discrete event simulation model used in this thesis when there is sufficient time. As a result, the delays that are estimated by the traffic simulation model used to perform the tests might not correspond to the delays estimated by Webster.

The effect of discretization is illustrated in Figure 9.6. In this diagram, instead of obtaining a smooth continuous curve similar to those illustrated in Figure 5.9, significant oscillations are observed in the amount of delay estimated from one signal cycle to the other, especially for short signal cycles. An analysis of the simulation results that were used to produce the diagram clearly indicates that these oscillations are caused by the discrete nature of the simulation model used to perform the analysis. This analysis indicates that each spike in the amount of estimated delay when reading the curve from left to right is caused by an increase in green interval duration that does not result in an additional vehicle being able to cross the intersection during any of displayed phases. For these cases, the increased green time resulting from the increased cycle time is entirely converted into a loss time. Consequently, each drop in the estimated delay is the result of a green interval increase allowing at least one additional vehicle to cross the controlled intersection during each signal cycle.

In addition to the above effect, the only source of randomness in the simulations carried out to evaluate both the fixed-time and real-time traffic signal operations is associated with the selection of the path to follow at intersection approaches where traffic splits into through and turning movements. As explained in Section 9.2, travel speeds, transit dwell times, and inter-arrival times are assumed to

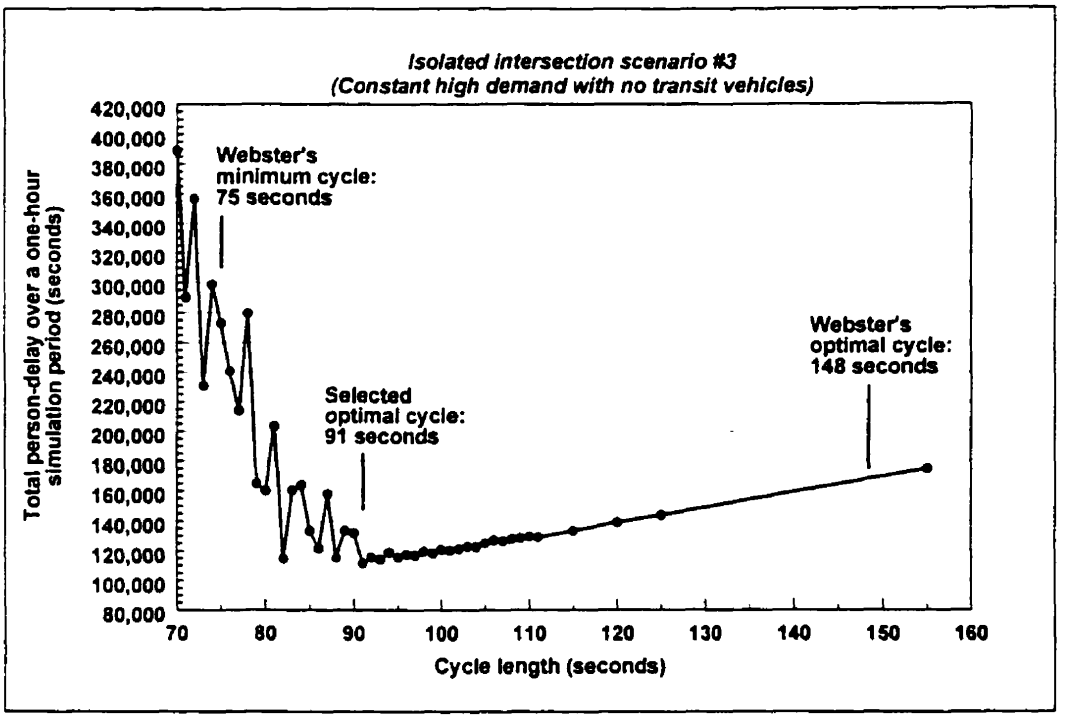


Figure 9.6 - Examination of Delay versus Cycle Time Relationship for a Simulated One-Hour Fixed-Time Operation

remain constant within each control period. While the average arrival rate may vary over time in the peaking-demand scenarios, the inter-arrival times remain constant within each sub-period. Webster's delay formulation assumes Poisson arrivals, while all arrivals in the test scenarios have constant inter-arrival times. As a result, the delays estimated by the traffic simulation model used in this thesis may not correspond to the delays assumed in Webster's delay minimization process. Given the reduced variability of simulated arrivals, the simulated delays are expected to be lower than the delays estimated by Webster's expressions.

To ensure that the fixed timings used to evaluate the effectiveness of the SPPORT model are the best that can be obtained given the above sources of variations, a two-step optimization process was used to determined them. In this process, the green allocation was first calculated for a series of cycle times using Webster's delay minimization principle. The cycle time/green allocation combination yielding the lowest total passenger delays was retained as the best fixed-time solution. For instance, consider again the example of Figure 9.6. In this case, a cycle time of 91 seconds is assumed to be optimal as this cycle time results in the lowest amount of delay incurred by all vehicle passengers. For the same example, the optimal cycle time computed using Webster's equation is 148 seconds, while the

minimum cycle time computed on the basis of the same equations is 75 seconds. It can be observed that the optimum cycle time resulting from the simulation falls between the minimum and optimum cycle times obtained from Webster's equations. These results provide a measure of confidence in the validity of the chosen cycle time, especially when the previously mentioned sources of variation are considered.

The detailed results of the evaluations that were conducted can be found in Appendix E. The main results are summarized in this section in Figures 9.7 and 9.8. Figure 9.7 illustrates the changes in total passenger delays that were achieved by SPPORT with respect to a fixed-time operation for the three scenarios of Table 9.1 with no transit vehicles, while Figure 9.8 illustrates the same changes for scenarios with transit vehicles. In each case, the comparison is made against the fixed optimal timings determined through simulation, as these timings produce better results in each case than those obtained with the application of Webster's equations only.

In both series of tests, the results shown on the left-hand side of the diagram are for the low, medium and high constant-demand levels, while the results shown on the right-hand side illustrate the observed changes in overall passenger delay for scenarios in which the demand is assumed to follow the peaking pattern of Figure 9.4. In all cases, simulation results are based on a one-hour simulation of a SPPORT operation. In order to allow the evaluation process to start with a realistic set of initial traffic and

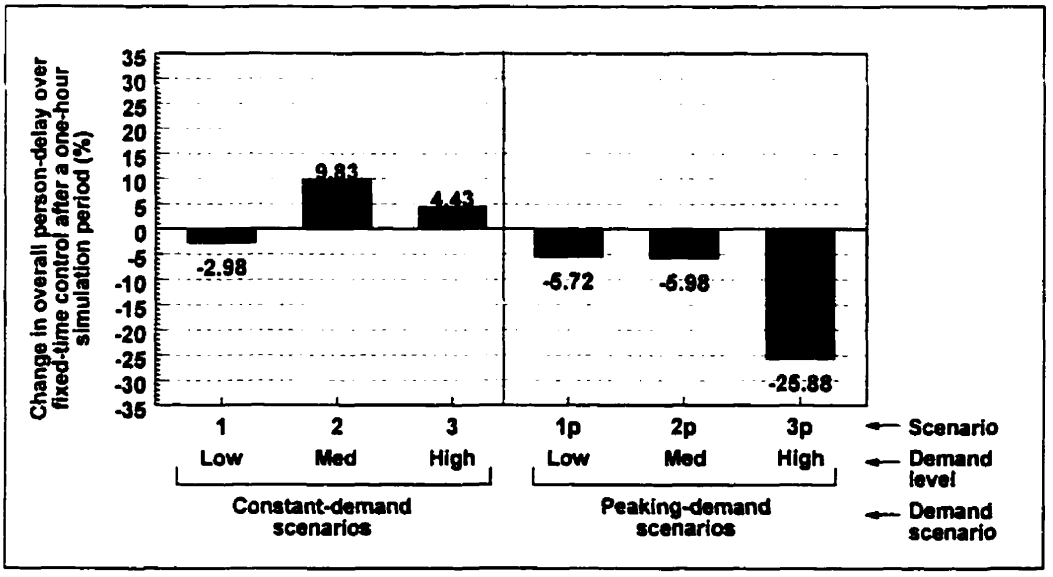


Figure 9.7 - SPPORT Effectiveness over a One-Hour Control Period for Isolated Intersection Scenarios without Transit Vehicles

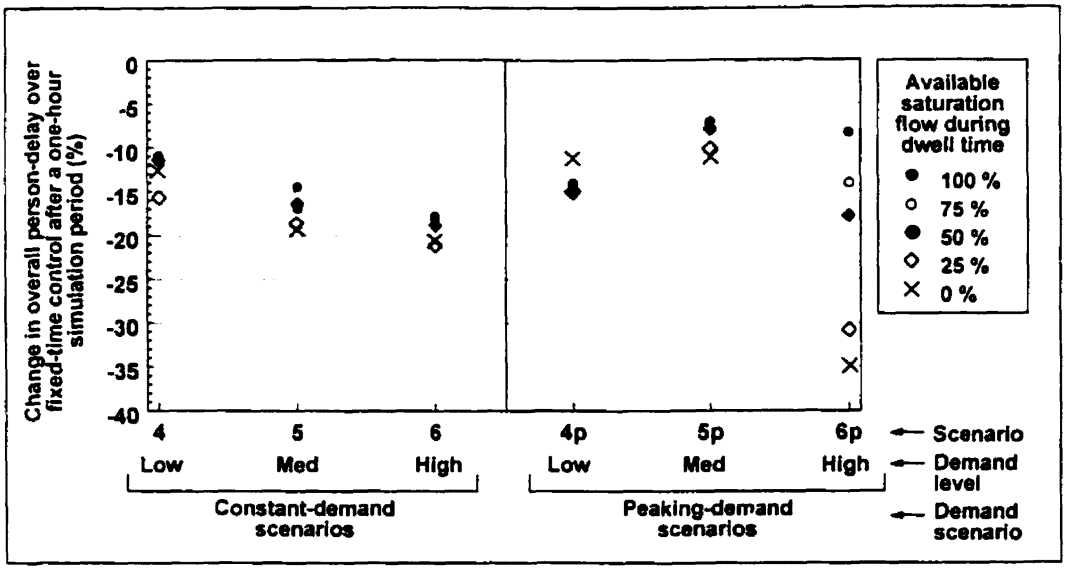


Figure 9.8 - SPPORT Effectiveness over a One-Hour Control Period for Isolated Intersection Scenarios with Transit Vehicles

queuing conditions, each simulation was first run for a five-minute warm-up period before starting to compile the delays incurred by vehicle passengers. This means that each simulation was run for a total of 65 minutes in order to obtain performance measures over a one-hour control period.

The results of Figure 9.7 indicate that SPPORT was able to reduce delays over the optimal fixed-time strategies by as much as 25 percent. However, reductions in delay were not obtained for all scenarios. SPPORT was unable to improve traffic conditions under all simulated levels of traffic demand in cases with constant vehicle arrivals and no transit vehicles. This increase in delay for three scenarios is not completely surprising given the fact that the fixed-time signal timings have been explicitly optimized for these traffic conditions.

Several reasons can be identified explain why the SPPORT model is unable to reduce delay for all the conditions examined. One reason is associated with the priorities assigned to the various events directing the operation of the model. While the assigned priorities were selected on the basis of engineering judgment and test simulations, there is no theoretical proof that these priorities lead to optimal signal control decisions. Assigning truly optimal priorities to the requests generated by each rule is currently very difficult due to the complex interactions that exist between traffic events. The process used by SPPORT to evaluate the conflicting needs of current and future traffic events in the phase selection process also adds significant complexity to the problem of selecting adequate priority

levels for each rule of the signal control logic. Since the priorities of various events are added together to evaluate the ability, or inability, of a given phase to serve the needs of the existing demand, there is no guarantee that the request with the highest priority will determine the next course of action. In many cases, a phase serving a certain number of medium-priority requests may be selected instead of a phase providing only service to the request with the highest priority level. As a result, the selection of an appropriate priority level for each rule is often subject to compromise between the need to allow the highest-priority requests to direct the signal operation and the need to serve lower priority requests.

Another possible explanation for the observed delay increases could be the fact that SPPORT is currently constrained to make signal-switching decisions at fixed five-second intervals. It must be remembered that the SPPORT model currently only allows green intervals to start at a decision point within the decision horizon (see Figure 5.7). As a result, SPPORT may not be able to implement truly optimal signal timings. For example, consider a 60-second decision horizon with a five-second interval between each decision point. In this case, the control parameter defining the operation of the model only provide 12 possible points within the decision horizon at which new phases can be scheduled to start. If a one-second interval were used, there would then be 60 possible switching points within the horizon.

The decision to impose a five-second decision interval has not been made on the basis of improved signal control performance, but rather on concerns that allowing signal optimizations to be conducted at intervals of less than five seconds may require too much CPU time for a computer to optimize the signal timings in a reasonable time (time requirements for the use of the model are explored in more details in Section 9.9). An indication of the effects of the above constraint can be seen in Figure 9.9, where it is observed that the duration of the green intervals implemented by SPPORT tend to oscillate around the optimal duration of the corresponding phase in the fixed-time operation.

To evaluate the impact of these oscillations on the performance of the model, an additional simulation, with a one-second decision interval instead of a five-second interval, was carried out for the same example as Figure 9.9. The results of this simulation are shown in Figure 9.10. In this figure, it is noted that much less variability exists in the duration of the cross-street green intervals. It is also observed that the magnitudes of the variations associated with the main-street green intervals are reduced, but that there is some significant variability remaining. The same behavior is observed for the cycle times implemented by SPPORT. In this case, however, the average cycle time is less than 50

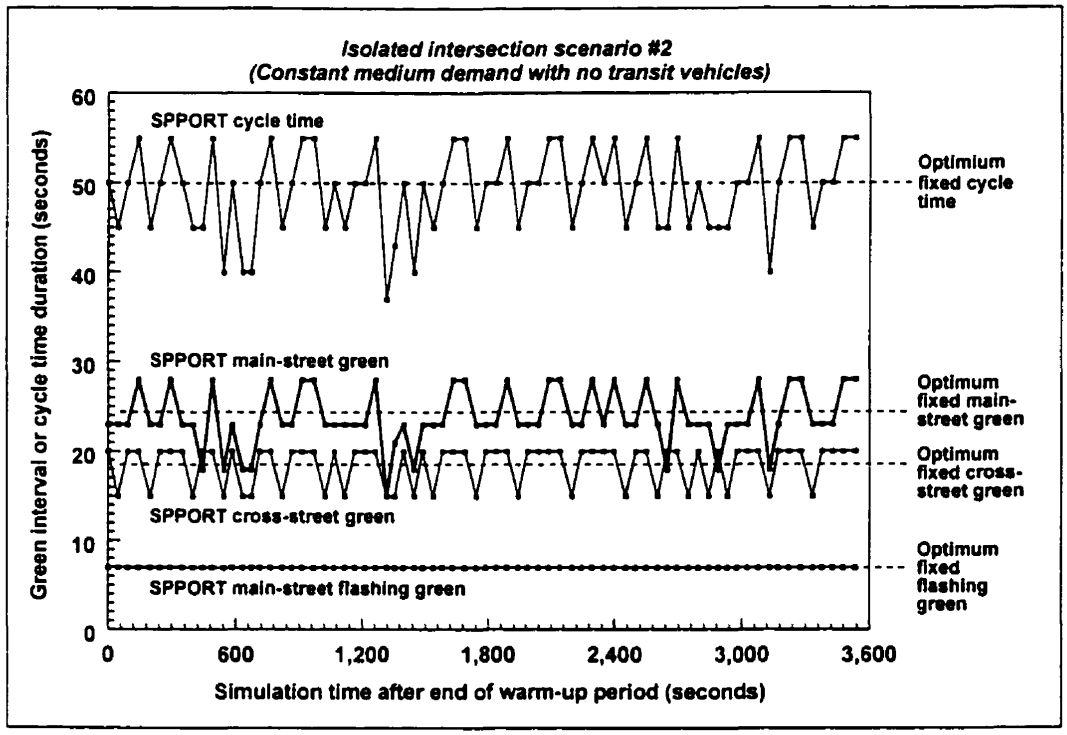


Figure 9.9 - SPPORT Signal Timings at an Isolated Intersection under Constant Demand while Using Five-Second Decision Intervals

seconds. In the example of Figure 9.9, the average SPPORT cycle time was at about 50 seconds. This reduction in average cycle time is attributed to the ability of the SPPORT model to implement shorter green intervals on both the main-street and cross-street approaches when a one-second decision interval is used in the signal optimization process.

The variations in signal timings observed in Figures 9.9 and 9.10 are attributable to the sensitivity of the rule-based decision logic to the existing traffic conditions. As previously explained, while the rate at which vehicles enter the controlled area is constant, there is a certain degree of randomness in the stop line demand being simulated due to the random process used to determine whether a particular vehicle approaching an intersection will go straight across it, make a right turn or turn left. In many cases, a change in only one vehicle in the estimation of the size of queues and incoming platoon might be sufficient to prevent a request from being generated, to change the call time of such a request, or to modify the priority level assigned to it. Such changes could then affect the phase selection process and lead to the delaying of a signal switch or to the unexpected truncation of the current phase. For

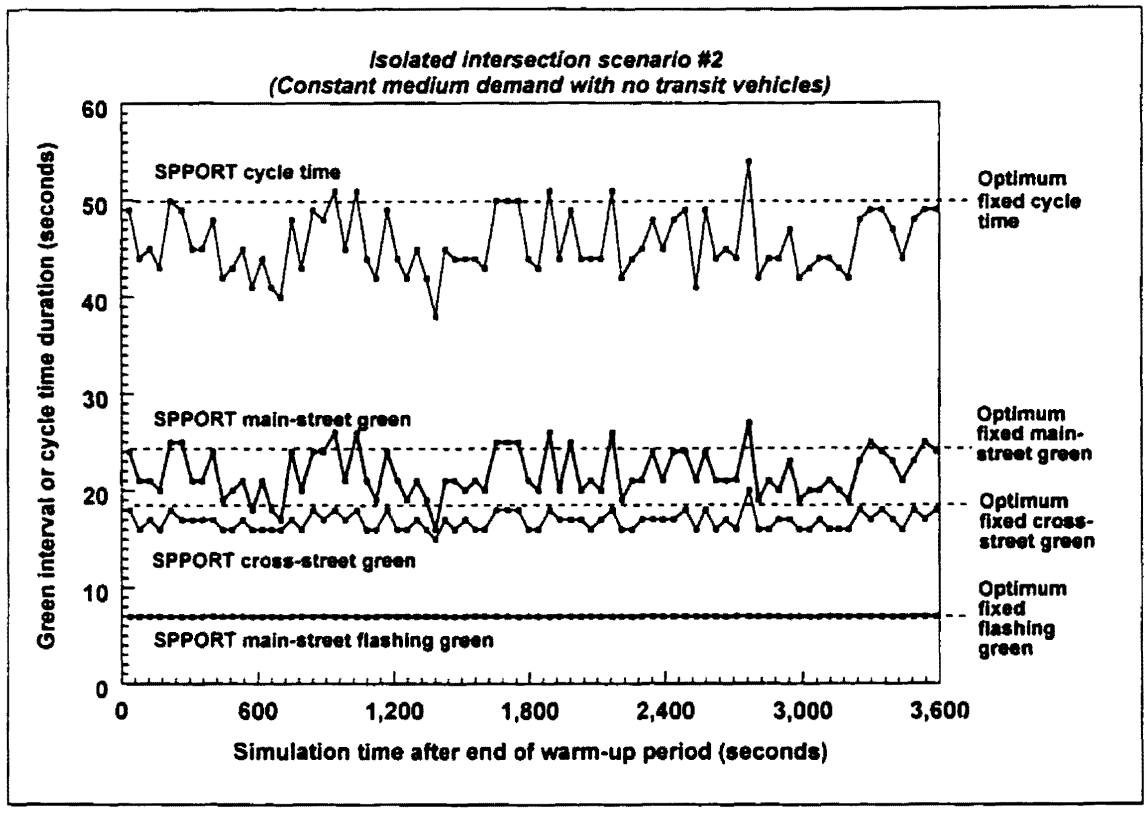


Figure 9.10 - SPPORT Signal Timings at an Isolated Intersection under Constant Demand while Using One-Second Decision Intervals

example, in order to continue serving a platoon, it is essential, in the absence of other requests, that the priority level assigned to the *Continue Serving Platoon* rule exceeds the sum of the priorities assigned to the requests generated by the *Begin Serving Queue* rule on each cross-street. In the first case, the priority level is a function of the estimated platoon size, while the priorities in the second case are functions of the estimated queue size on each cross-street approach.

For the test scenario illustrated in Figure 9.9, the SPPORT model performed better with the use of a five-second decision interval than with the use of a one-second interval. In this case, the use of a one-second decision interval results in an 8.5 percent increase in delay incurred by vehicle passengers. However, the use of a five-second interval does not result in less delay for all scenarios examined. A detailed analysis of simulation results indicates that the non-optimality of the priorities assigned to each rule might be the main factor at work behind these variations.

Figure 9.7 also illustrates the results obtained for the scenarios in which the traffic demand is assumed to peak within the control period. In this case, delay reductions over an optimal fixed-time operation ranging between 7.7 and 25.9 percent were obtained in all scenarios. These results illustrate the ability of the SPPORT model to adjust the signal timings automatically to newly observed traffic conditions.

As an example, consider Figure 9.11, which illustrates the timings implemented both by SPPORT and the fixed-time signal operation in a scenario involving medium traffic demand, peaking arrival patterns, and transit vehicles requesting priority of passage. In this example, it is observed that SPPORT automatically lengthens and shortens the duration of both the main-street and cross-street green intervals in response to increases and decreases in traffic demands. In this case, the main-street green is varied in direct relation to changes in the rate of arriving traffic, while the cross-street green is mostly varied in response to the need to serve the longer or shorter queues that are produced on these streets as a result of varying cross-street red intervals (it must be remembered here that all peaking-demand scenarios feature constant arrivals on the cross-streets). Some short time variations are still

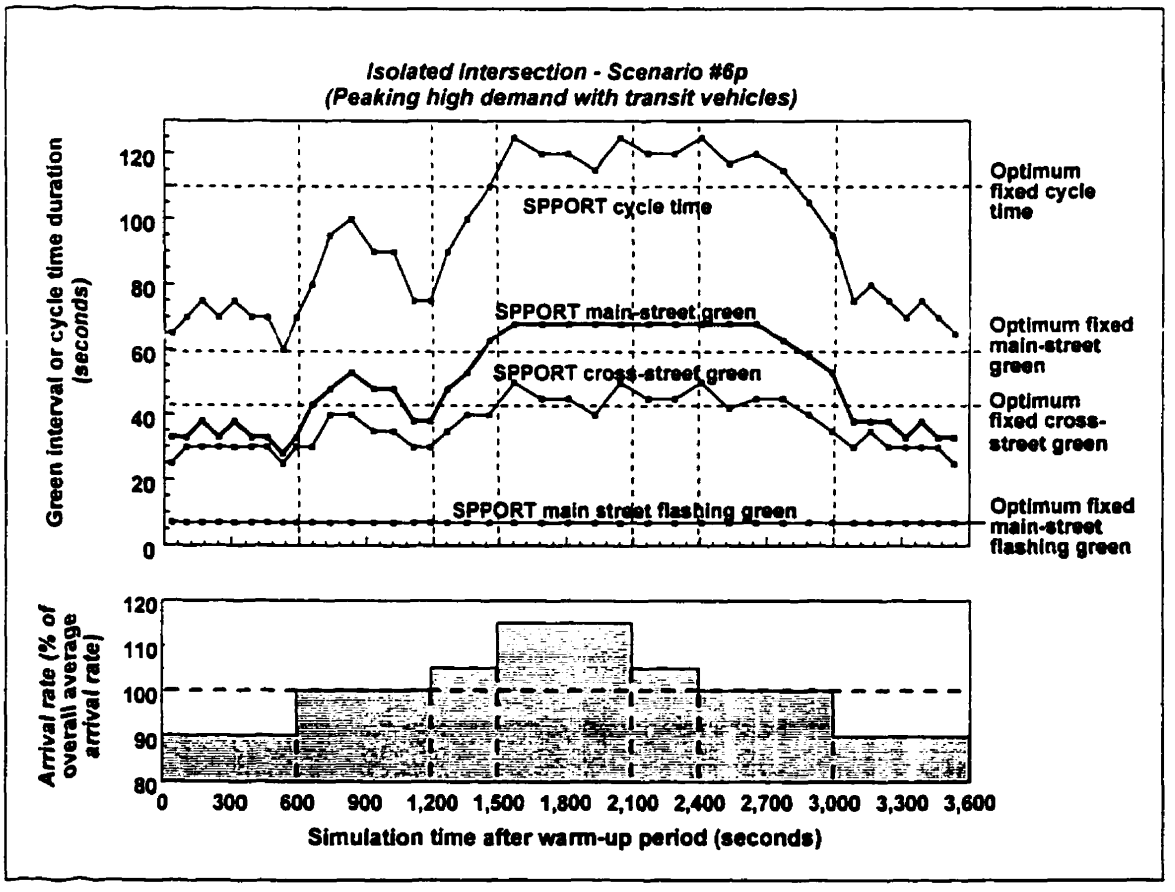


Figure 9.11 - Variation of SPPORT Timings at an Isolated Intersection under Peaking Demand

observed, but as explained earlier, these variations might be caused by the randomness of projected stop line arrivals. The larger variations occurring between 600 and 1200 seconds within the simulation could for their part have many causes. For instance, the need to provide priority of passage to successively approaching transit vehicles in the northbound and southbound approaches of the intersection could cause the sudden increase in main-street green interval duration over a period of three signal cycles. In such case, the extension in the duration of the cross-street green interval would be explained once more by a need to serve longer cross-street queues. The reduction in main street green interval duration occurring between 900 and 1200 seconds would then be explained by the sudden absence of transit vehicles requesting priority of passage and the gradual dissipation of the queues caused by the previously implemented priority treatments. In the simulation, only the duration of the main-street flashing green interval remains the same, as there is never enough demand for that phase to justify the implementation of a phase longer than the user-imposed minimum interval.

Figure 9.11 also clearly indicates the inability of the fixed-time traffic signal operation to follow the changes in traffic demand. While the fixed timings are designed to satisfy the overall average demand, it is clear that they provide excessively long green intervals on each approach at the beginning and end of the simulated period, and maybe not enough in the middle of the period when the arrivals are peaking. At the beginning and end of the control period, unnecessary delays may then be incurred by drivers as a result of the imposition of longer than necessary red intervals. In the middle of the period, additional delays might result from the inability of the signal operation to completely dissipate queues of vehicle during each signal cycle, causing some drivers to wait a full cycle of signal operation before being able to cross the intersection.

Figure 9.8 illustrates in a different way the ability of the SPPORT model to react to sudden changes in traffic demands. This diagram shows the changes in total person delay resulting from the utilization of SPPORT as a function of the level of interference caused by transit vehicles stopping in the right of way to board and discharge passengers. In this case, better traffic signal control is obtained with SPPORT in all simulated scenarios. Depending on the scenario considered, the use of the model resulted in delay reductions over an optimal fixed-time operation ranging between 7 and 35 percent. However, what is of more interest is the observed general tendency of improved benefits with increasing levels of transit interference. Since transit activities introduce additional variability in arriving traffic, these results show again the ability of the SPPORT model to operate effectively in variable traffic demand conditions. In this case, the results show more specifically the ability of the

model to react to sudden drops in arriving traffic on approaches currently having a green signal indication and to move the green indication to approaches on which traffic could be served at a higher rate.

A detailed analysis of the results of Figure 9.8 also indicates that the illustrated benefits are equally attributable to the ability of the SPPORT model to provide priority to transit vehicles on an active basis and to respond to the consequences of implementing such special treatment. This ability is shown in the example of Figure 9.12. In this example, six distinct spikes in the curve representing the cycle time are observed within the simulated one-hour period. Each one of these spikes corresponds to an increase in the main-street green interval that is implemented to accommodate an incoming transit vehicle. Following each main-street green interval extension, it can also be observed that SPPORT temporarily extends the cross-street green. This extension is awarded to serve the longer queues that have formed on these streets as a result of the previous main-street green interval extension.

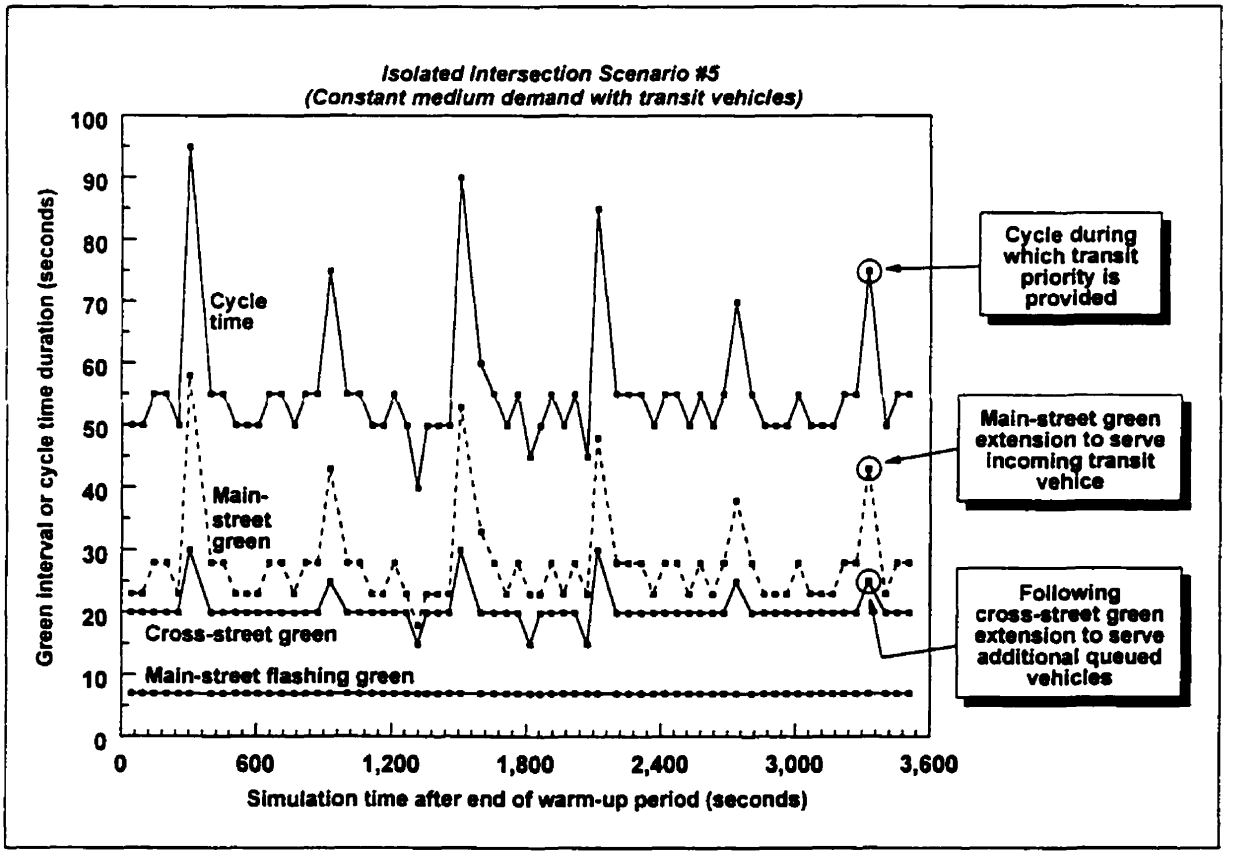


Figure 9.12 - Variation of SPPORT Signal Timings at an Isolated Intersection in Response to Transit Priority Requests

While the information presented in all the above figures are based on an evaluation of the delays incurred by all vehicle passengers, similar delay increases and reductions are observed when the simulation results are compiled on a simple vehicular basis. The results of the above simulations are also consistent with the results of experiments reported in Conrad et al. (1998). In these experiments, an earlier version of the SPPORT model was applied to an isolated intersection with constant passenger car and transit vehicle arrivals. In this case, transit vehicles were assumed to travel on only one approach, with a constant arrival rate of one vehicle every three minutes. It was also assumed that each transit vehicle would cause a complete blockage of the approach on which they travel for 30 seconds when they stop near the intersection stop line to board and discharge passengers. In these experiments, the application of the SPPORT model resulted in significant improvements over an optimal fixed-time operation based on Webster's signal control principles. Evaluations were made by estimating a performance index considering both the stops and delays incurred by all simulated vehicles. Depending on the level of traffic simulated, the use of the SPPORT model resulted in a performance index decrease ranging between 8.1 and 28.5 percent, with the highest benefits obtained with the highest level of traffic demand. Given the fact that severe transit interference, and thus highly variable traffic conditions, were simulated, these benefits again clearly showed the ability of the SPPORT to quickly respond to observed changes in traffic demand, and the inability of traditional fixed-time traffic signal operation to do the same.

9.4. Effectiveness of SPPORT Signal Control on Urban Arterials

The second series of tests that were conducted to evaluate the ability of the SPPORT model to efficiently control traffic in real-time is concerned with the control of a typical five-intersection urban arterial. In these tests, the performance of the model is compared with a fixed-time operation using signal timings optimized with TRANSYT-7F, Release 7.1 (Wallace and Courage, 1992). This software was chosen to perform the analyses as it is one of the most widely used in North America for developing optimal timing plans for urban street networks controlled in fixed-time.

Similar to SPPORT, TRANSYT-7F combines the use of traffic simulation and signal optimization routines to generate optimum signal timing plans. Depending on the performance function chosen by the user, this model may be used to estimate the number of stops and the amount of delay incurred by vehicles at signalized intersections, the maximum stop line queue size on each approach during each signal cycle, or the total fuel consumed by vehicles within the controlled network. However, contrary

to SPPORT, TRANSYT-7F does not simulate the progression of individual vehicles. Instead, it projects second-by-second histograms of average flow rates from one intersection stop line to the next, starting from the upstream boundary of the network, and then going downstream from one intersection to the other. While transit activities can also be simulated, it is assumed that transit vehicles stopping to board and discharge passengers never interfere with the progression of other traffic.

When instructed to generate new optimum signal timings, TRANSYT-7F starts by determining the optimum common cycle time that should be applied to all intersections in the controlled area, and then simultaneously determines the optimum green allocation and signal offset combination at individual intersections. The common cycle time is determined using a trial-and-error process. The best cycle time is determined by performing a quick green allocation and signal offset optimization for all possible cycle times in the range provided by the user, and then by choosing the cycle time yielding the best performance index. In the second part of the optimization, the optimum green allocation and signal offset for each intersection are determined using an iterative, gradient search algorithm. This algorithm searches for the optimal green allocation and signal offset by testing a series of small changes in the values of these parameters and only retaining those changes that result in an improvement in the performance function.

In order to maintain a common base of comparison between TRANSYT-7F and SPPORT, the timings produced by both models were evaluated using the discrete-event simulation model described in Chapter 6. Both models were also instructed to use a similar performance function to evaluate the effectiveness of their proposed signal timings when making signal timing decisions. Signal optimizations were carried out using an objective function that linearly combines the total number of stops and the total amount of delay incurred by all vehicle passengers within the controlled network. Table 9.8 indicates the values that were attributed to the various parameters of Equations 5.1 and 5.2 in

Table 9.8 - Performance Function Parameters for the Arterial Scenarios

Parameter (Equations 5.1 and 5.2)	Value
k_d	1.0
k_s	20.0
k_{TT}	0.0
k_{TC}	1.0
α_R	0.0
$\alpha_{v\ private}$	1.5
$\alpha_{v\ transit}$	60.0

order to define the above-mentioned performance function in SPPORT, while Equation 9.1 defines the performance function minimized by TRANSYT-7F.

$$PI_{T7F} = \sum_{j=1}^{N_{links\ T7F}} w_j \left(d_j + k_s\, s_{avg\ j} \right) \qquad [9.1]$$

where: PI_{T7F} = TRANSYT-7F performance index function.

j = Subscript identifying the link number (j = 1, 2, ..., $N_{link\ T7F}$).

$N_{links\ T7F}$ = Total number of links in the networks.

d_j = Total delay on link j (vehicle-hours).

$s_{avg\ j}$ = Average number of stops on link j per unit of time (stops/second).

k_s = User-defined coefficient expressing the importance of stops relative to delay (k_s = 20.0 seconds/stop).

w_j = Link-specific stop and delay coefficient (w_j = 1 for links carrying passenger cars and 40 for links carrying transit vehicles).

As indicated in Table 9.8 and Equation 9.1, it is assumed in the performance function of each model that each stop carries the same importance as a 20-second delay. In both cases, a value of 20.0 has been assigned to the stop coefficient k_d expressing the relative importance of stops over delays. This value was chosen based on a recommendation made in the user's manual of TRANSYT-7F, Release 6 (Federal Highway Administration, 1988). In this guide, it is indicated that a stop coefficient in the range of 20 to 50 has been shown to produce signal timings providing a good balance between stops and delays. The manual also indicates that a value falling in the above range tends to minimize the total fuel consumed by vehicles traveling across the controlled area.

To verify the validity of the chosen value for the stop coefficient k_d, a series of optimizations runs were made with the SPPORT model in which the relative weight of stops in relation to delays was varied over a predefined range. Figure 9.13 illustrates the results of these simulations for the five-intersection arterial configuration with medium demand levels, peaking arrivals and transit vehicles (scenario 8 in Table 9.1). The upper diagram of Figure 9.13 illustrates the variation of total person delay as a function of the stop coefficient. These results indicate that while the use of a stop coefficient of 20 does not result in the absolute minimum delays, the delays incurred by all vehicle passengers still appears to be very close to the minimum levels. In the lower diagram of Figure 9.13, the same coefficient is associated with the lowest number of stops incurred by vehicle passengers. When the

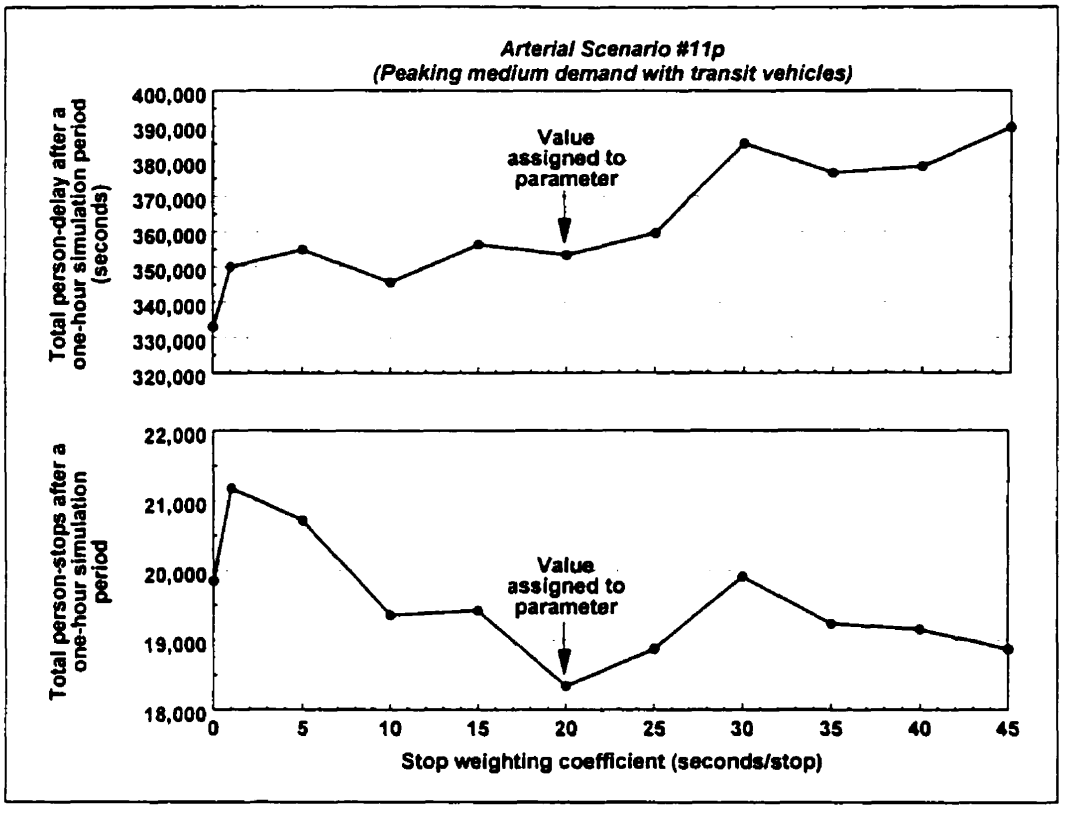


Figure 9.13 - Effect of Stop Weighting Coefficient Value on Traffic Signal Performance

results of the two diagrams are combined, the choice of a stop coefficient of 20 for this example appears to be a near optimal decision.

In Equation 9.1, a link-specific stop and delay coefficient is used to allow TRANSYT-7F to compile performance measures on a person basis rather than on a vehicle basis. This coefficient is also used to allow the model to systematically bias the optimization in favor of transit movements and in this way provides priority to transit vehicles on a passive basis. Contrary to the simulation model used by SPPORT, TRANSYT-7F models transit operations along urban streets by assigning separate links to the transit vehicles. These "transit links" may be entirely separate to simulate traffic lanes used exclusively by transit vehicles, or links that share a stop line with the links used by passenger cars. In the later case, a shared stop line simulation is made to consider the interference caused to the progression of transit vehicles by passenger cars waiting in queue at the stop line of signalized intersections. Since TRANSYT-7F only compiles stops and delays on a vehicular basis, providing a coefficient that increases the estimated number of stops and delays on transit links indirectly allows the model to consider person-based statistics. For example, if it is assumed that a transit vehicle holds on

average 60 persons and a passenger car 1.5 person, a stop and delay coefficient of 40 assigned to the transit links will allow TRANSYT-7F to consider that a transit vehicle holds on average 40 times more passengers than a passenger car when optimizing the signal timings.

When optimizing signal timings with TRANSYT-7F, the platoon dispersion feature provided with the model was not used. Platoon dispersion usually occurs as a result of the different speeds at which individual vehicles travel within a platoon. In typical situations, these differing speeds cause compact groups of vehicles to spread in space as these groups travels away from a signalized intersection. In this thesis, no platoon dispersion is assumed to occur as a result of the fact that the discrete-event simulation model used to evaluate both fixed-time and real-time signal operation does not model such traffic behavior. Another reason is that SPPORT also does not consider platoon dispersion, as it uses the same simulation model as the one used to perform the evaluations. In this case, the assumption of no dispersion is not unrealistic since simulations are carried out for an urban arterial featuring relatively closely spaced intersections. As shown in the diagram of Figure 9.3, the longest distance between two successive intersections in the arterial configuration is 315 metres. Consequently, while some platoon dispersion will occur, it can be expected that platoons leaving each intersection would retain much of their compact form in a real-world network.

The detailed results of the various simulations that were conducted to evaluate the ability of the SPPORT model to effectively control traffic in real time along urban arterials can be found in Appendix F, while Figures 9.14 and 9.15 summarize the main results. Figure 9.14 illustrates the effect of the SPPORT timing on the selected performance function for the three arterial scenarios of Table 9.1 with no transit vehicles. Figure 9.15 illustrates the same effect for the three scenarios with transit vehicles. In both diagrams, the results shown on the left-hand side are for the low, medium and high levels of traffic demands in the scenarios with constant arrivals, while the results shown on the right-hand side are for the scenarios in which traffic demand is assumed to peak within the control period according to the pattern illustrated in Figure 9.4. Similar to the tests performed with the isolated intersection test configuration, the results reported in Figures 9.14 and 9.15 are for a one-hour traffic signal control simulation. Again, all simulations were first run for a five-minute period before starting to record performance measures.

Similar to the isolated intersection scenarios, the results that are reported in Figure 9.14 show a certain inability for the SPPORT model to improve traffic conditions in scenarios involving no transit vehicles.

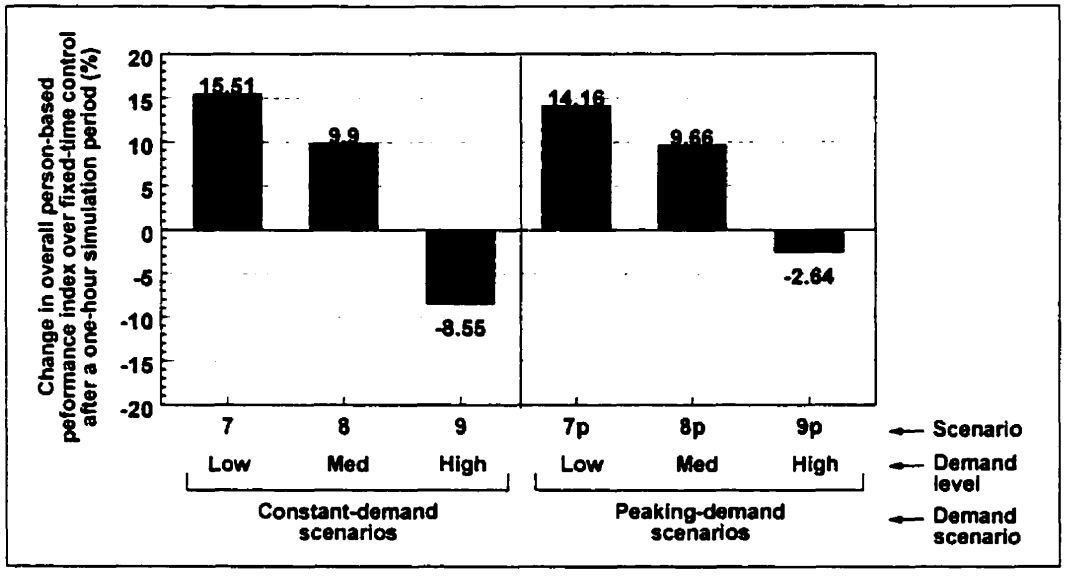


Figure 9.14 - Effectiveness of SPPORT Traffic Signal Operation over a One-Hour Control Period in Urban Arterial Scenarios with no Transit Vehicles

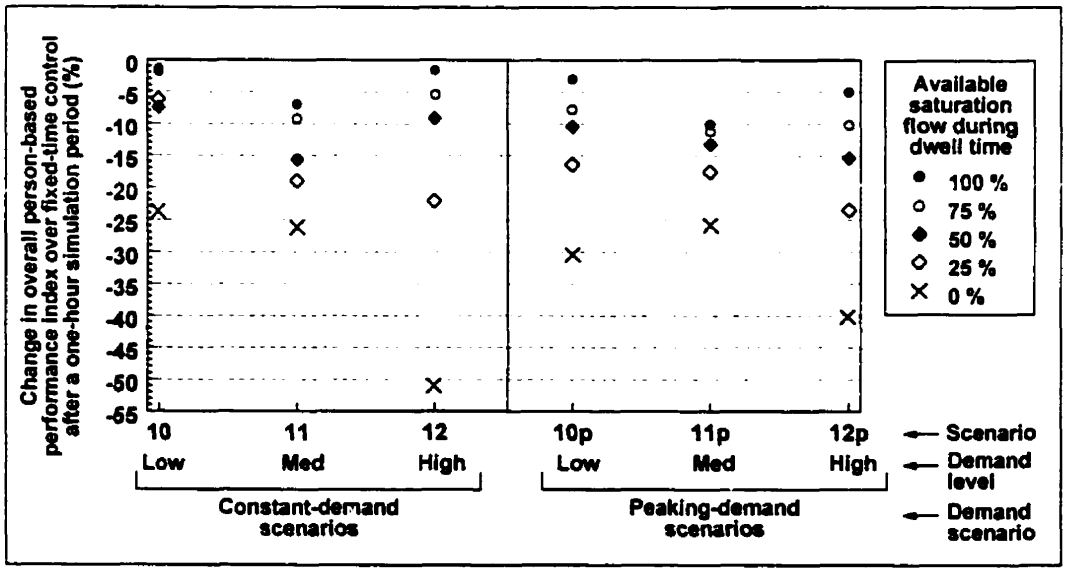


Figure 9.15 - Effectiveness of SPPORT Traffic Signal Operation over a One-Hour Control Period in Urban Arterial Scenarios with Transit Vehicles

In these results, it is found that TRANSYT-7F provides better signal control for both the low and medium demand scenarios. For the scenarios with low demand, the difference in the estimated performance index is 15.5 percent in favor of TRANSYT for the case with constant arrivals, and 14.5 percent for the case with peaking arrivals. For the scenarios with medium demands, the differences in the performance index are reduced to 9.9 and 8.7 percent, respectively. Improvements in favor of SPPORT are only observed with the high-demand scenarios. In this case, it is found that SPPORT improved traffic performance over TRANSYT-7F by 8.5 and 2.8 percent, respectively, for the cases with constant and peaking arrival patterns.

The results of Figure 9.14 are not surprising given that TRANSYT-7F has been explicitly designed to generate optimum signal timings for urban networks with constant traffic demands. As for the isolated intersection scenarios, the current inability of the SPPORT model to improve traffic conditions might be attributable to the use of five-second decision intervals or the possible non-optimality of the priorities assigned to each rule.

Despite the inability of the SPPORT model to consistently improve traffic conditions over TRANSYT-7F timings, it is observed that the benefits associated with the use of the SPPORT model increases with every increase in the level of traffic demand. When the high demand scenario is reached under both the constant and peaking demand arrival processes, the SPPORT model outperforms TRANSYT-7F. In this case, the ability of the SPPORT to produce better timings might be attributable to its queue spillback prevention features. In TRANSYT-7F, the desire to avoid queue spillbacks is only considered through the imposition of a penalty factor for queues exceeding the capacity of a given link. In SPPORT, queue control is dynamic. As it was illustrated in Figure 8.16, when queues threaten to spill across an intersection, SPPORT simultaneously attempts to switch the signal at the upstream intersection to red to prevent the queue from spilling back across the intersection, and to switch the downstream signal display to green to promote queue dissipation. This treatment results in a temporary change of signal control strategy. Instead of minimizing stops and delays, the objective becomes the preservation of intersection capacity and the maximization of traffic throughput.

Similar to the results of the analyses conducted with the isolated intersection scenarios, the results of Figure 9.15 present a totally different evaluation picture. In this case, improvements were obtained with SPPORT in all scenarios considering transit activities along the simulated arterial. Depending on the scenario being simulated and on the level of transit interference, these improvements range from

one percent to 50 percent over TRANSYT-7F timings. Again, higher benefits were also obtained with the simulations with higher degrees of transit interference. These results confirm the ability of the SPPORT model to react automatically to short-term variations in traffic demands.

9.5. Effectiveness of Signal Coordination Logic

While the results of Section 9.4 focused attention on the ability of the SPPORT model to improve traffic conditions on an urban arterial with respect to an optimal fixed-time operation, the results presented in this section specifically attempt to evaluate the effectiveness of the signal coordination logic that has been developed as part of this thesis. In order to evaluate the effectiveness of the various components of the logic, a series of simulations were made with different combinations of active and inactive elements within the signal coordination logic. For each of the arterial scenarios described in Table 9.1, the following five signal optimizations were performed:

1. Optimization with no data exchange between adjacent signal controllers and no request delaying (as if each intersection would be controlled in isolation of other intersections);
2. Optimization with exchange of projected signal timings between adjacent upstream and downstream signal controllers and no request delaying;
3. Optimization with exchange of projected near-future stop line departures between adjacent signal controllers and no request delaying;
4. Optimization with exchange of both projected near-future departures and signal control strategy and no request delaying logic;
5. Optimization with exchange of both projected near-future departures and signal control strategy and with the request delaying logic active;

In each case, simulations were performed assuming that transit vehicles stopping in the right-of-way to board and discharge passengers would completely close one of two available traffic lanes for the entire duration of their dwell time. Similar to the tests performed in the previous sections, simulations were first allowed to run for five minutes before starting to compile performance measures over a one-hour control period. Simulations were also carried out using the same traffic cops, decision horizon, look-ahead period, decision interval, and performance function as for the previous arterial simulations.

Figure 9.16 illustrates the overall benefits of the signal coordination logic implemented in SPPORT. More detailed results are provided in Appendix F. For each scenario, the diagram compares the

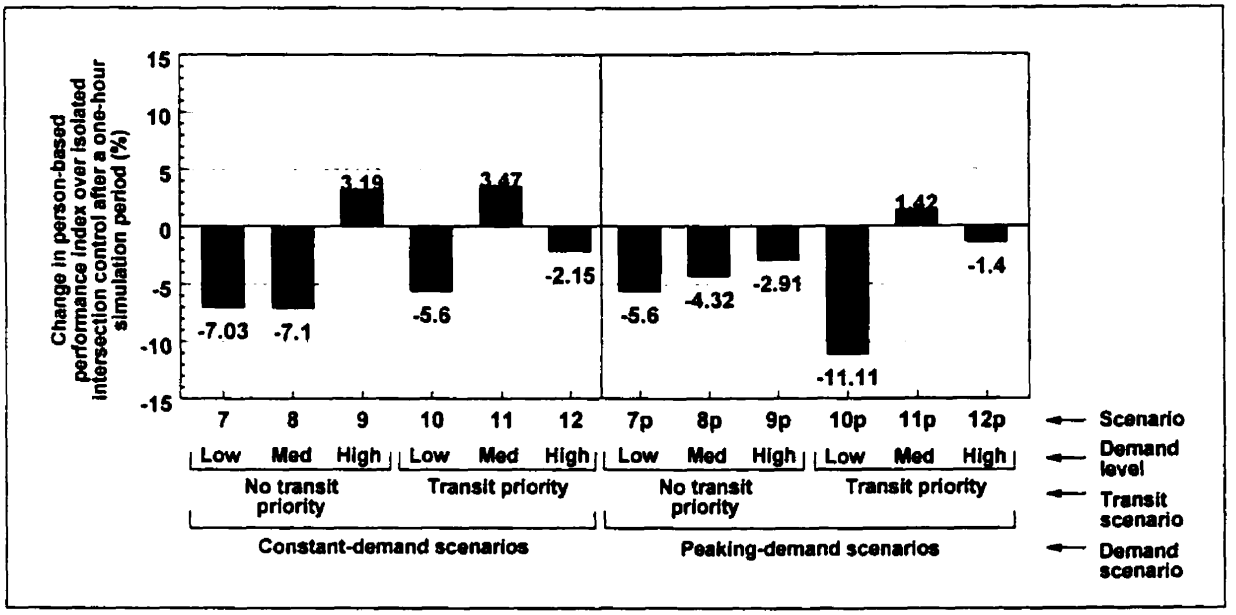


Figure 9.16 - Effect of Signal Coordination Elements on Arterial Traffic Control Performance

performance index of the signal operation considering no coordination elements (optimization case #1) with the index of the operation implementing all data transfers between adjacent intersections and the request delaying logic (optimization case #5). The comparison indicates an overall positive impact of the signal coordination logic. As it can be noted, increases in the estimated performance index as a result of the use of the coordination logic were only obtained in three of the twelve scenarios. In these three cases, performance index increases ranged between 1.4 to 3.5 percent. In all other cases, decreases in the value of the estimated performance index ranged between 1.4 to 11.1 percent.

To determine the source of the three observed increases in the estimated performance index, the results of Figure 9.16 were recompiled to show the specific contribution of the request delaying logic in the observed traffic performance changes. The resulting diagram is given in Figure 9.17. For each scenario, the diagram compares the performance of the signal operation implementing all data exchange processes between adjacent intersections, but not the request delaying logic (optimization case #4), with the performance of the operation implementing both the data exchanges processes and the request delaying logic (optimization case #5). The results of this compilation indicate that the request delaying logic had a negative impact on traffic performance in only one scenario. This scenario is the one with the medium demand, constant arrivals and no transit vehicles. In this case, the impact of the request delaying logic is a 3.29 percent increase in the estimated performance index. For all other scenarios, reductions ranging up to 5.91 percent are observed.

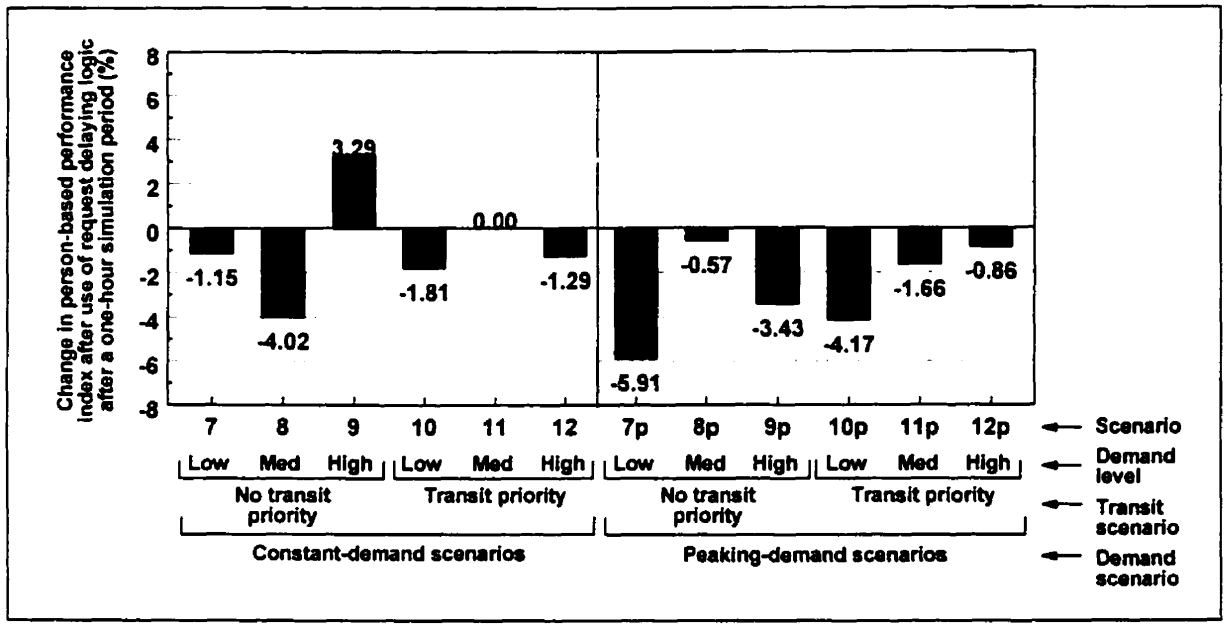


Figure 9.17 - Effect of Request Delaying Logic on Arterial Traffic Control Performance

Figure 9.18 provides a more detailed examination of the specific effects of each element of the signal coordination logic for the scenario of Figure 9.17 for which the use of the request delaying logic did not result in an overall traffic performance improvement (constant high demand with no transit vehicles). The main conclusion that can be drawn from the information contained in that diagram is that the use of projected signal switches by adjacent controllers and projected departures from adjacent upstream intersections does not necessarily lead to better control strategies. For instance, while the use of projected signal switches at adjacent intersections (optimization case #2) leads to a reduction in the number of stops along the controlled arterial, the use of this information also leads to an overall increase in delay. Instead, a delay reduction is observed when only projected departures from adjacent intersections are used (optimization case #3). However, the increase in the number of stops that accompany this delay reduction is sufficient in this case to completely offset the benefits of the delay reduction and produce a performance index that is even higher than the one obtained with the operation considering no data exchange. The best traffic conditions are still obtained when both projected signal timings and vehicle departures are considered; however, the resulting improvement over the non-coordinated scenario is rather small.

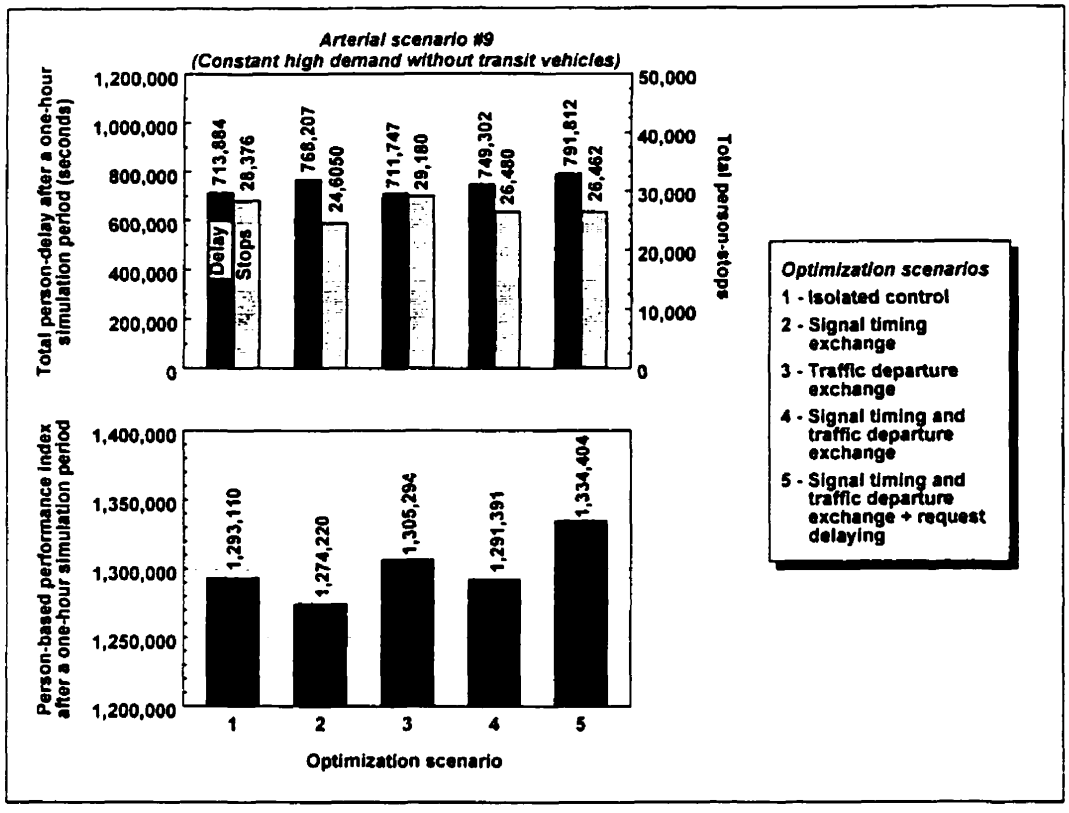


Figure 9.18 - Effect of Individual Elements of the Request Delaying Logic on Arterial Traffic Control Performance in Scenario without Transit Vehicles

Figure 9.19 compiles for each scenario the change in performance index obtained by adding the signal switches projected by adjacent controllers and the projected departures from upstream intersections to the information used by SPPORT to optimize the signal timings. From this figure, it appears that considering the projected switching decisions from adjacent intersections usually yields positive results. However, negative results are obtained in almost all the scenarios when SPPORT is considering projected departures from upstream intersections. When both projected timings and vehicle departures are considered, negative results are only obtained in a minority of cases.

A detailed analysis of the simulation outputs reveals that the negative effects of considering projected departures from upstream adjacent intersections are mainly due to the way SPPORT handles requests calling for future signal switches in relation to requests calling for immediate switches. Under non-coordinated control, only vehicles currently traveling on the immediate approaches to an intersection are considered in the signal optimization process of each intersection. When links are established with adjacent intersections, signal optimizers are then able to predict traffic stop line vehicle arrivals further

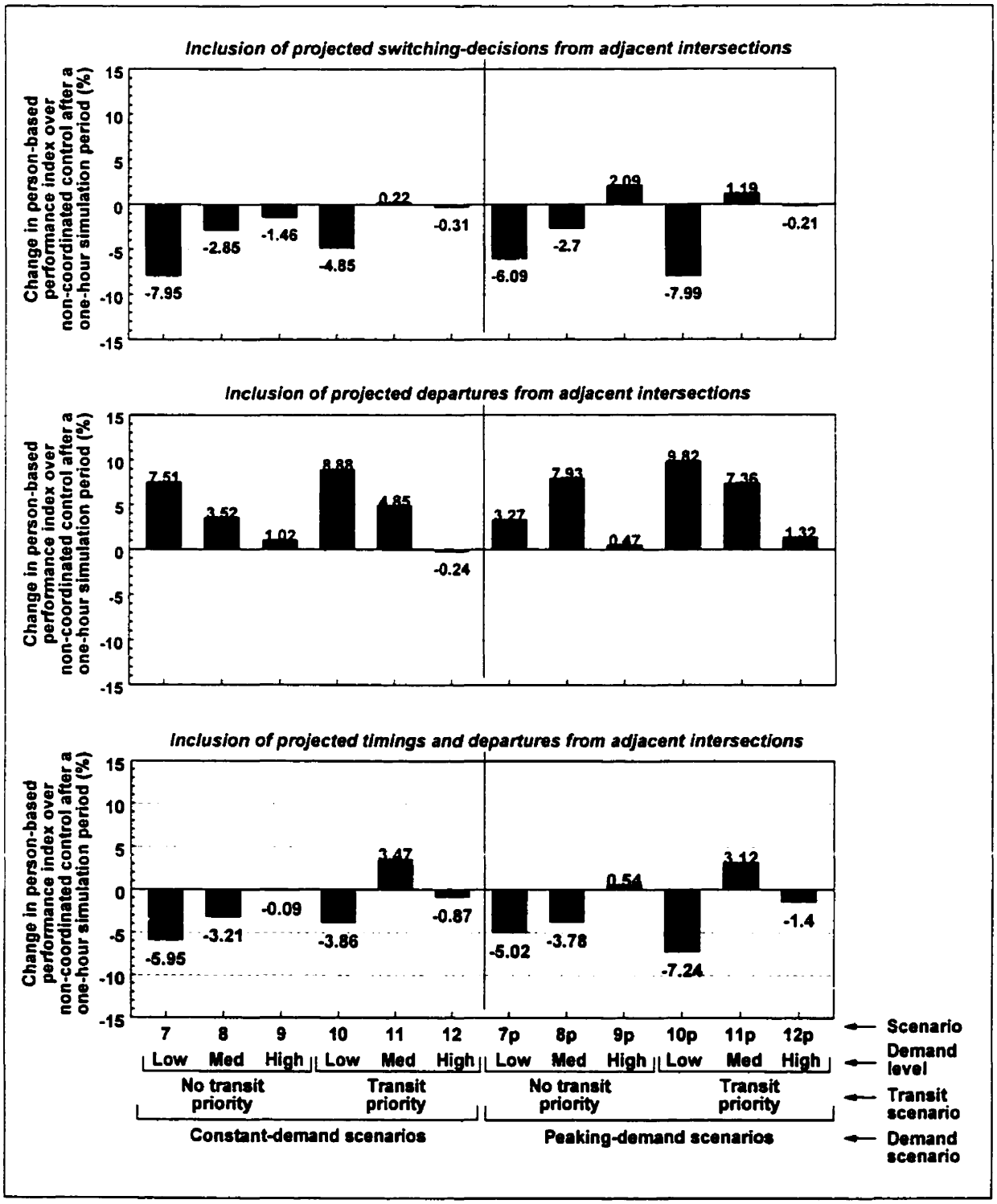


Figure 9.19 - Effect of Individual Elements of the Request Delaying Logic on Traffic Control Performance in all Arterial Scenarios

into the future, and thus better able to identify platoons scheduled to reach the intersection at some future time. This ability to predict the future arrival of incoming platoons causes a problem in the fact that SPPORT currently assigns priorities to traffic events without considering their time of occurrence.

For example, consider an arterial on which traffic detectors installed on each intersection approach allow the signal control system to detect approaching vehicles 20 seconds before their arrival time at the intersection stop line. Under non-coordinated control, this detection layout would only allow signal optimizers to project stop line arrivals at the intersection they each control over the next 20 seconds. As a result, if it were assumed that a minimum of 10 vehicles constitutes a platoon, a request to serve an incoming platoon would only be generated if the first 10 vehicles of a platoon were projected to reach the intersection within the next 20 seconds. If projected departures from upstream intersections over the next 20 seconds are also received, it then becomes possible for each optimizer to predict stop line arrivals over the next 40 seconds. Consequently, any platoon for which the first 10 vehicles arrive within the next 40 seconds could be identified and could result in the generation of a signal request. At this point, the same priority level would be assigned to a request calling for a green signal indication in 20 seconds to serve a platoon scheduled to arrive at that time than to a request attempting to serve a platoon that would reach the intersection within the next five seconds.

Ideally, a lower priority level should be assigned to the requests dealing with platoons scheduled to arrive later in time to reflect the greater importance of serving existing traffic over predicted demands that are still subject to changes. By assigning the same priorities to traffic events regardless of their time of occurrence, situations could arise in which the need to serve future traffic events may take more importance than to serve current traffic conditions. In such situation, the SPPORT model may be prevented from implementing signal switches that may not satisfy the need of the incoming platoons but that may be highly beneficial to the overall traffic performance. For instance, consider an intersection where a platoon is scheduled to reach the stop line of one of the four approaches in the next 20 seconds. In addition, consider that a green signal indication is currently displayed on that approach and that phasing requirements impose a 30-second minimum red interval following the end of any green interval. Also assume that the sum of the priority levels of all the requests calling for the start of a green interval on other approaches is lower than the priority level assigned to the request calling for a continuation of the current green signal indication to serve the incoming platoon. In this case, a decision could be taken by SPPORT to hold the current green signal indication until the platoon has been served, even though very few or no vehicles would be served in the next 20 seconds on the

approach on which the platoon is traveling. In this case, it might be better to switch the signal to the other approaches to immediately start serving the existing stop line queues and reduce the amount of delay incurred by vehicles on these approaches, even if such control treatment may cause the vehicles in the incoming platoon to be stopped for a short time at the intersection.

On the basis of the results presented in this section, it appears that the request delaying logic developed as part of this thesis performs its duties reasonably well. For most scenarios, improvements in traffic performance were obtained as a direct consequence of the application of the signal coordination logic developed as part of this thesis. More specifically, a detailed analysis of the simulation results indicates that both the use of projected signal-switches from adjacent intersections and the request delaying logic is generally beneficial. The only major identified problem is concerned with the use of projected departures from upstream intersections to predict future vehicle arrivals. As explained above, the problem is more particularly concerned with the current inability of the SPPORT model to reduce the priority level of future events to reflect the diminishing importance of serving events that are further in time.

9.6. Effectiveness of Transit Priority Rules

While the previous sections focussed attention on the ability of the SPPORT model to improve general traffic conditions around isolated intersections and along urban arterials, this section evaluates the ability of the model to provide effective transit priority treatments on a real-time basis in mixed-traffic environments. This is achieved by examining in more details the effects of implemented priority treatments on the stops and delays incurred by both transit and non-transit riders in the controlled networks.

A first indication of the effectiveness of the model in implementing efficient transit priority has already been given in the example of Figure 9.12. In this figure, six distinct spikes are observed along the line illustrating the duration of the main-street green interval throughout the one-hour simulated control period. As explained in Section 9.3, each one of these spikes corresponds to an extension of a green signal indication that has been implemented to accommodate a transit vehicle approaching the intersection under control. The diagram also illustrates the ability of the model to provide priority treatments tailored to the needs of existing traffic demands and to later compensate for the effects that these treatments have on other traffic. First, it is observed that different green signal indication

extensions were awarded in each situation in which transit priority was awarded. These variations are the effects of both different transit stop line arrival times with respect to the beginning of the main-street green interval and different surrounding traffic conditions. Second, it is also observed that SPPORT temporarily increases the duration of the cross-street green interval after each main-street green interval extension. These cross-street extensions are not granted to accommodate transit vehicles. Instead, they are granted to accommodate the longer queues that have formed on the cross-streets as a result of temporarily extending the red signal on these streets to accommodate transit vehicles traveling on the main-street.

Figures 9.20 to 9.23 illustrate the ability of the SPPORT model to provide effective transit priority treatments at controlled intersections. Figures 9.20 and 9.21 compare the operation of the SPPORT model at an isolated intersection with an optimal fixed-time operation based on Webster's delay minimization principles. Figures 9.22 and 9.23 provide comparisons for the five-intersection arterial configuration illustrated in Figure 9.3. In this case, timings generated by the TRANSYT-7F model are used as a reference for the optimal fixed-time operation. In both cases, the results are from the same simulations that were described in Sections 9.3 and 9.4 to evaluate the general ability of the SPPORT to improve traffic conditions at controlled intersections and along controlled arterials. More detailed results can be found in Appendices E and F.

In Figures 9.20 and 9.21, the ability of the SPPORT to provide effective transit priority is evaluated by comparing the delays incurred by transit riders and non-transit riders in both the fixed-time and SPPORT operations. In Figures 9.22 and 9.23, similar comparisons are made, but on the basis of a performance index considering the stops and delays incurred by transit riders and the passengers of other vehicles. In all cases, comparisons are made with stop and delay estimates that do not consider the stops made by public vehicles at transit stops, nor the time spent by these vehicles to board and discharge passengers. These stops and time spent at these locations are not considered, as they constitute an integral part of regular transit service and are not affected by the signal operation.

Together, these four figures clearly demonstrate the ability of the SPPORT model to favor transit operations without causing excessive traffic disruptions to other traffic. In Figure 9.20, it is first observed that the application of SPPORT results in a significant decrease in the amount of delay imposed to transit riders by the signal operation. In five of the six scenarios, delay reductions exceed 30 percent. In three of these scenarios, reduction exceeds 50 percent, with one example showing an

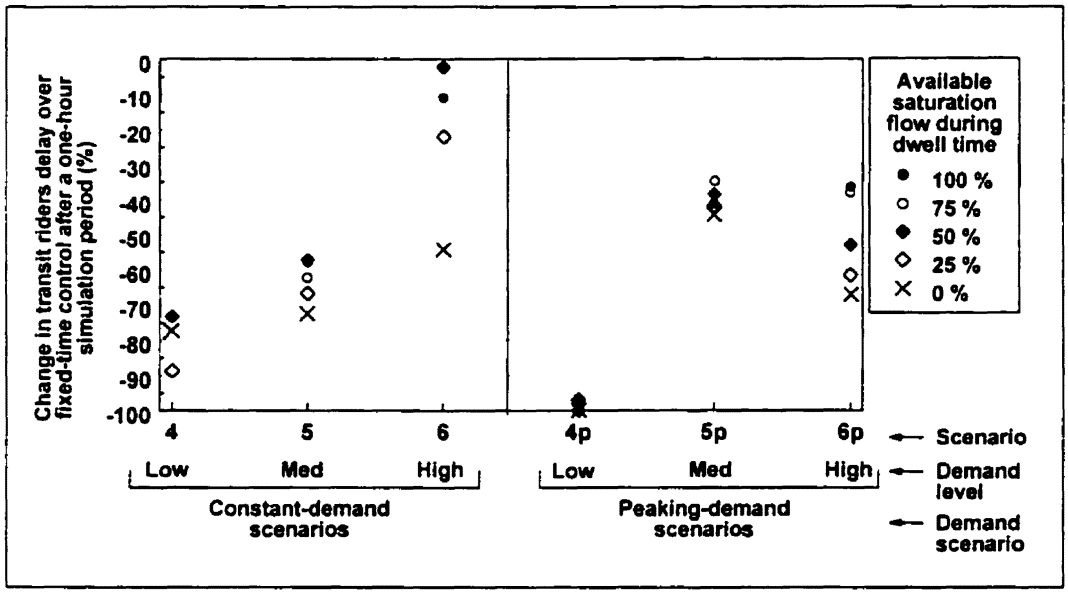


Figure 9.20 - Effect of Transit Priority Rules for Transit Riders in Isolated Intersection Scenarios

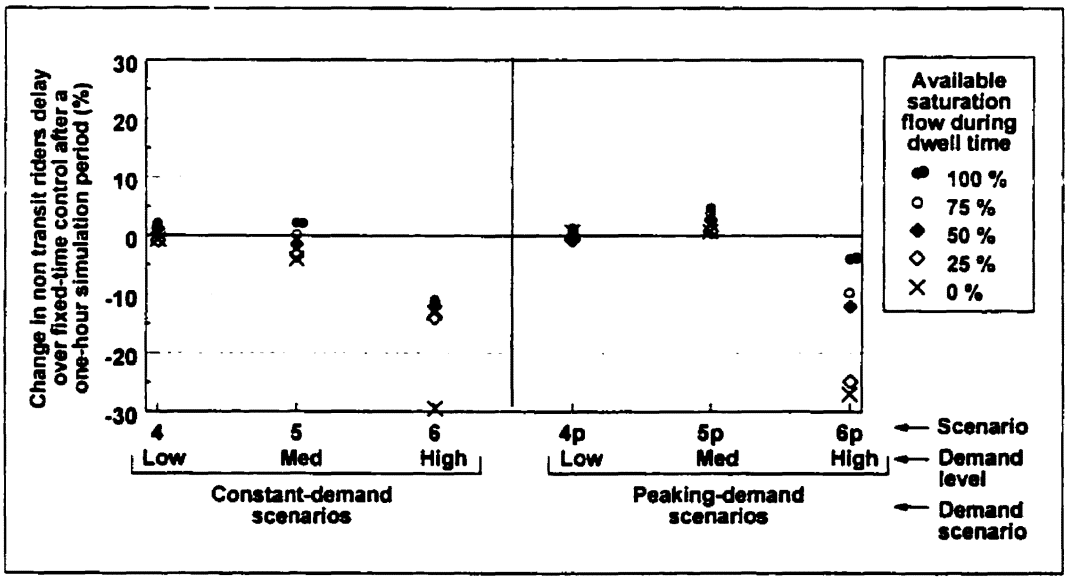


Figure 9.21 - Effect of Transit Priority Rules for Non-Transit Riders in Isolated Intersection Scenarios

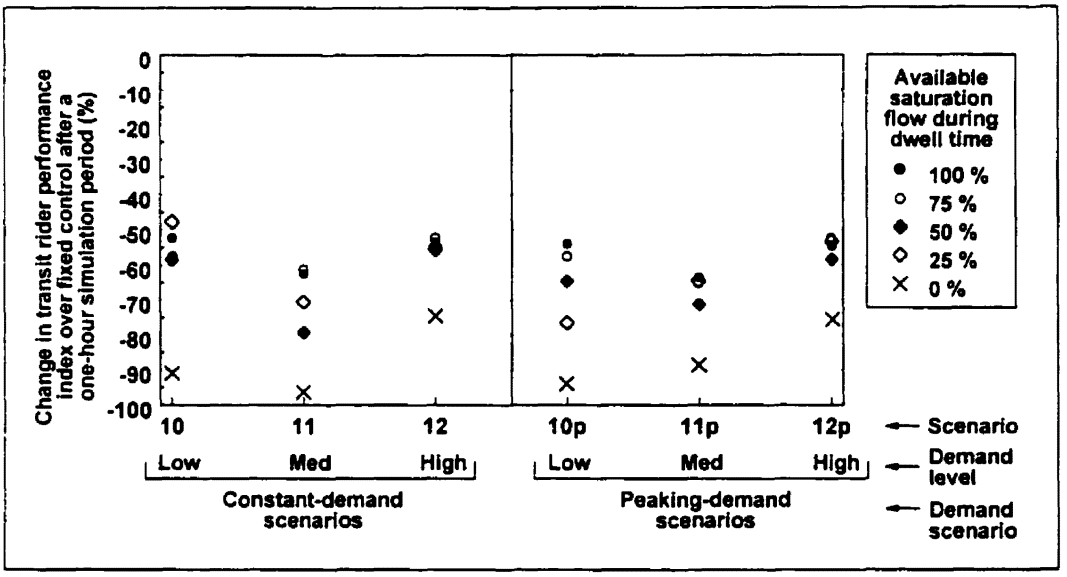


Figure 9.22 - Effect of Transit Priority Rules for Transit Riders in Arterial Scenarios

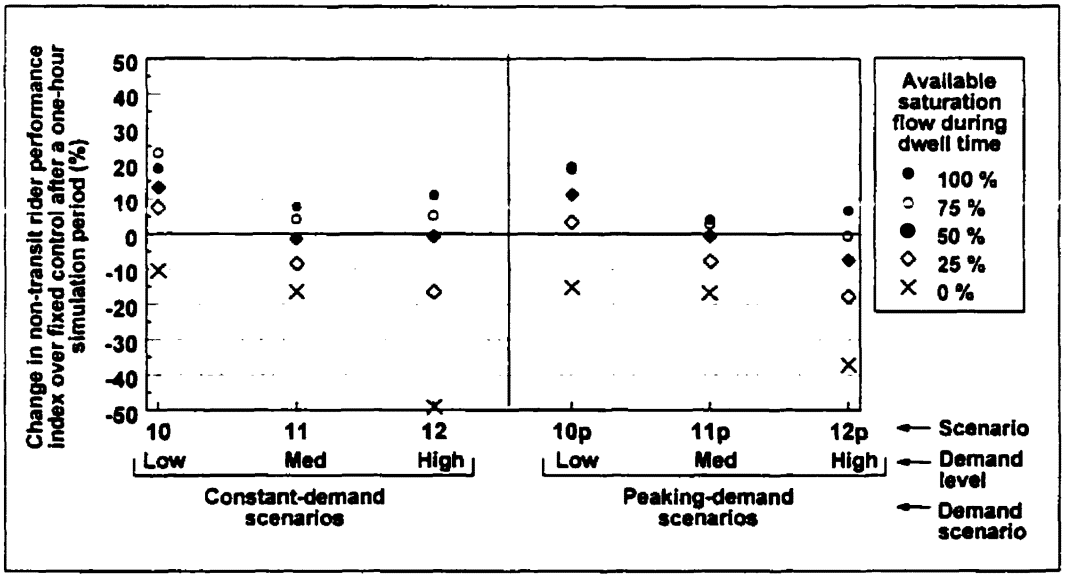


Figure 9.23 - Effect of Transit Priority Rules for Non-Transit Riders in Arterial Scenarios

almost complete elimination of transit delay. The exceptional performance of the SPPORT model in this case is attributed to the occurrence of very few conflicts between transit vehicles and the general traffic, as well as the facility with which the SPPORT model can move the green signal indication around when the general traffic demand is light. Similarly, the lower delay reductions obtained in the scenario with high demand and constant arrival patterns can be explained by the difficulties for the SPPORT model to switch the signal displays in response to transit priority requests at intersections operated near or at saturation.

Figure 9.21 complements the information of Figure 9.20 by indicating that delay reductions achieved by SPPORT for the transit riders were not made at the exaggerated expense of other traffic. While delay increases were observed for the non-transit riders in four of the six illustrated scenarios, these increases never exceeded five percent. Moreover, as indicated in Figure 9.8, none of these delay increases were large enough to negatively affect the overall performance of the controlled intersection in any one of the six scenarios.

For the two scenarios simulating high traffic demands, the application of the SPPORT model appears to be beneficial to both the transit and non-transit riders for all levels of transit interference. This result can be attributed to traffic conditions that were exceptionally favorable to transit priority treatments. For example, small signal alterations may have only been required to accommodate incoming transit vehicles each time priority of passage was requested. The switch in the signal control policies implemented by SPPORT that occur when the tail of a stop line queue threaten to spill across an upstream intersection may also be at the source of these benefits. In these cases, instead of attempting to minimize stops and/or delays, SPPORT attempts to dissipate as soon as possible the existing queue to avoid queue spillbacks across upstream intersections. While none of the isolated intersection examples featured intersections at the upstream end of intersection approaches across which queue could threaten to spill back, it was nevertheless assumed that "potential" spillback conditions would exist on these approaches if the tail of the stop line queue would cover more than 75 percent of their entire length. As a result, upstream queue management strategies may have been applied to contain the maximum size of the stop line queue on each approach while this would not have been the case in the fixed-time operation.

Figures 9.22 and 9.23 confirm the conclusions from Figures 9.20 and 9.21. Similar to the isolated intersection scenarios, these diagrams indicate that the application of the SPPORT model is highly

beneficial for transit riders in all arterial scenarios. For almost all scenarios, the performance index of transit riders was reduced by more than 50 percent under SPPORT control when compared to a fixed-time operation using TRANSYT-7F timings and passively favoring transit movements. In this case, however, the results are more stable as a result of the fact that each data point in the diagrams represents the combined results of five different intersections. In many scenarios, non-transit passengers also suffered from stop and delay increases. However, as shown in Figure 9.15, none of these increases negatively affected the overall performance of the arterial in any of the simulated scenarios, particularly in the cases in which on-lane transit activity severely disrupted other traffic.

9.7. Effectiveness of Multi-Objective Signal Control

In the previous sections, signal optimizations were carried out using the same combination of traffic cops for each test scenario. For the scenarios with no transit vehicles, three traffic cops were used. Each one of these cops held a different prioritized list of events in which the only variable element was the way incoming platoons were handled on each intersection approaches. Since there were no transit activities to consider, these lists also completely ignored the *Bring Transit Vehicle to Transit Stop* and *Bring Transit Vehicle to Stop line* rules designed to provide priority to incoming transit vehicles. For the scenarios with transit vehicles, three other traffic cops were added. These cops were an exact duplication of the three previous ones, except for the fact they each assigned high priorities to the two transit rules mentioned above.

In these tests, the main objective of using more than one prioritized list of events in the signal optimization was to allow SPPORT to determine the best control strategy at each decision point. For each list, a timing plan was generated and the plan resulting in the best performance index over the decision horizon was selected for implementation. This process allowed SPPORT to consider transit priority on a conditional basis. By combining lists favoring transit activities with lists that do not, a choice was given to SPPORT at each decision point and priority treatment would only be awarded to incoming transit vehicles if such treatment were found to be beneficial for the overall performance of the controlled intersection.

In order to evaluate the benefits of this process, additional optimizations were carried out for all peaking-demand scenarios involving transit vehicles using different combinations of traffic cops. The results of these tests are provided in Appendix H. Similar to previous tests, traffic performance was

evaluated in each case on the basis of a one-hour control period and using a performance function assuming that each stop made by the passengers of any vehicle had the same importance as 20 seconds of waiting time.

Figure 9.24 summarizes the results of the optimizations that were carried out for the medium demand arterial scenario. The first observation that can be made from the results in Figure 9.24 is the overall positive effect of using multiple traffic cops to perform signal optimization within SPPORT. This is evident from the several reductions in the performance index with increasing number of traffic cops. The second observation that can be made is the ability of the SPPORT model to provide priority to transit vehicles on a conditional basis.

In each section of the diagram, it is observed that the optimization in which the efforts of the largest number of traffic cops is pooled either produces the best signal control strategy or a strategy that is very close to the best one. It is also observed that the optimization using all six traffic cops defined in Table 9.2 produced the overall best solution. These results indicate the benefits of allowing SPPORT

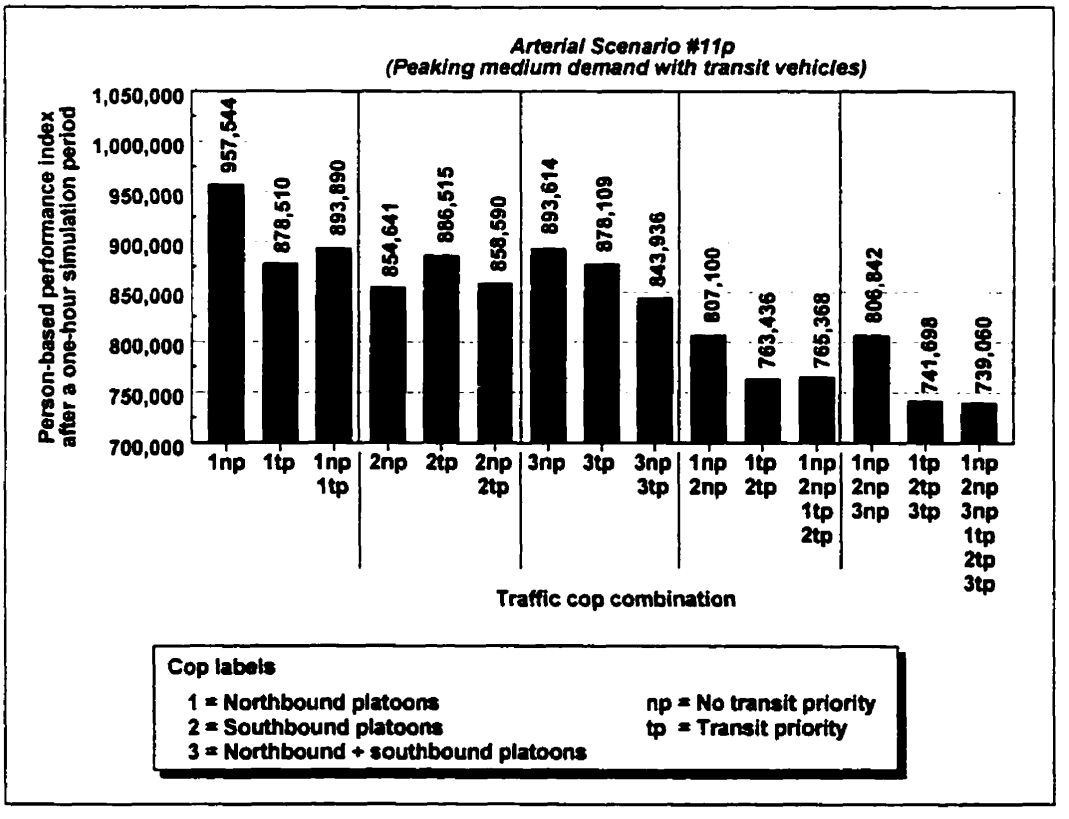


Figure 9.24 - Comparative Performance of Multi-Objective Traffic Signal Operations

to generate candidate timing plans using different prioritized lists of events and to choose among these plans at each decision point the one best suited for the existing traffic conditions. The results demonstrate the difficulty faced by traffic engineers in selecting a priori signal control strategies that can produce optimal signal-switching decisions at each decision point. In Figure 9.24, five of the six overall worst signal control strategies were obtained when only one traffic cop was used. At the same time, three of the four best overall solutions were obtained when the efforts of at least three traffic cops were evaluated before making any signal-switching decisions.

One of the main benefits of using multiple lists of events to generate signal control decisions within SPPORT is an improved coordination with adjacent intersections. In Figure 9.24, all traffic cops in each of the first three sections of the diagram attempt to allow the same platoons to cross the controlled intersections without being stopped. In the first section, the attention is given only to platoons approaching each intersection on the northbound approach. In the second section, the attention is given to platoons on the southbound approaches only, while platoons in both directions are considered in the third section. In the last two sections, different traffic cops look at different platoons. As it can be observed, results in these sections are much better. In this case, the use of multiple traffic cops allows SPPORT to evaluate the impacts of providing uninterrupted progression to incoming platoons in each direction and to determine, at each decision point, which direction should get the green signal indication.

The fact that the best or near best signal control strategies were obtained within each section of Figure 9.24 with the optimization using the largest number of traffic cops also demonstrates the ability of the SPPORT model to provide priority to transit vehicles on a conditional basis. Each section of the diagram presents the results of three distinct optimizations. In each section, the first optimization used only traffic cops that were not instructed to provide special treatments to incoming transit vehicles on any intersection approaches. The results of these optimizations show what happens when transit priority is completely ignored. The second optimization used only traffic cops instructed to provide priority of passage to all incoming transit vehicles. The results of these optimizations show what happens when priority is always granted. As it can be observed, while providing priority on an absolute basis may be beneficial in many circumstances, such benefits may not be obtained in every scenario. For instance, in the example of Figure 9.24, the decision to provide priority to all incoming transit vehicles causes a 1.7 to 8.3 percent decrease in the overall traffic performance index over non-priority scenarios in four of the five sections of the diagram. However, the same control strategy

causes a 3.7 percent increase in the overall traffic performance index in the scenarios of the third section.

The third optimization combined the two previous sets of traffic cops. In this case, SPPORT had the ability to evaluate signal control strategies that considered providing priority to transit vehicles and strategies that did not. As a result, SPPORT was able to determine at each decision point whether or not granting priority to the approaching transit vehicles would benefit the overall traffic performance. As shown in Figure 9.24, this strategy produced very good results in all the sections of the diagram, and the best strategy in two of them. In the other sections, the optimization implementing absolute transit priority produced the best results. However, a difference of less than 3.9 percent, and of only 0.3 percent separates the performance indexes of the two transit priority strategies in these cases. Such small differences confirm the ability of the SPPORT model to effectively consider transit priority on a conditional basis.

9.8. Sensitivity to Selected Signal Control Parameters

This section presents the results of analyses that were performed to determine the impact of selected control parameters on the signal timings produced by SPPORT. More specifically, the effects of the following parameters are reported:

- Duration of decision horizon (see Figure 5.7);
- Duration of look-ahead period at each decision point within the decision horizon (see Figure 7.4);
- Critical headway in the platoon identification process (see Section 7.4);
- Critical flow rate used in the request generation process to determine whether a given flow rate is high enough to warrant the generation of a request (see Section 8.1.1).

These parameters were chosen for examination on the basis of their assumed potential effects on the operation of the SPPORT model.

The results of the analyses conducted for one particular scenario are illustrated in Figures 9.25 to 9.28. The scenario chosen is the medium-demand arterial scenario with transit vehicles and peaking stop line arrival patterns. This scenario was chosen as it simulates mid-range varying traffic demands. Similar to the analyses conducted in the previous sections, the results reported in the figures evaluate the

performance of the SPPORT model on a person basis over a one-hour simulated operation of the model. For each point, a single replication was made. To perform the optimizations, all six traffic cops defined in Table 9.2 were used. It was also assumed that each stop incurred by the passengers of any vehicle would correspond to a 20-second delay.

In addition to the test results, each one of the above diagrams indicates the value that was assigned in Section 9.2 to the corresponding parameter to evaluate the effectiveness of the SPPORT model in the control of both isolated intersections and urban arterials. When these values are analyzed in relation to the illustrated curves, the first observation that can be made is that the parameter values selected for

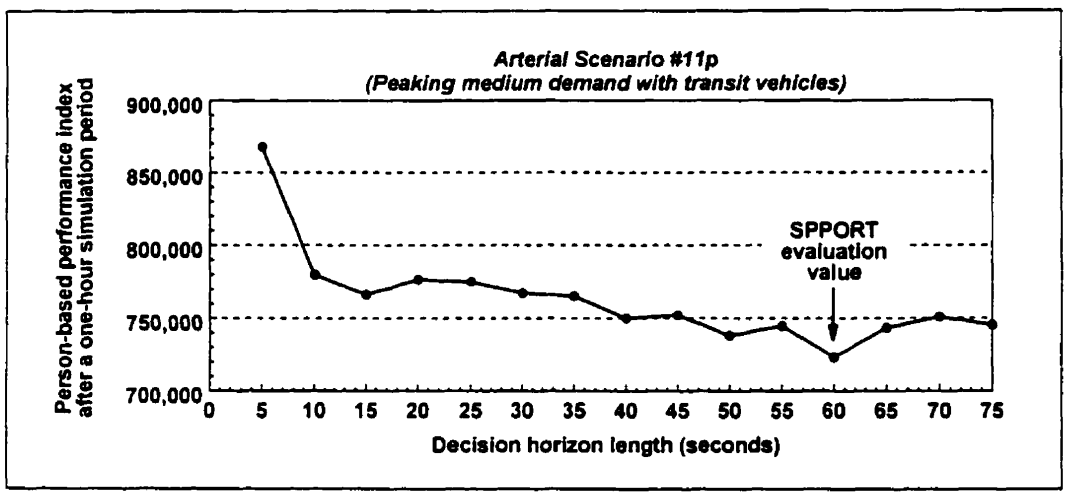


Figure 9.25 - Effect of Duration of Decision Horizon on Traffic Control Performance

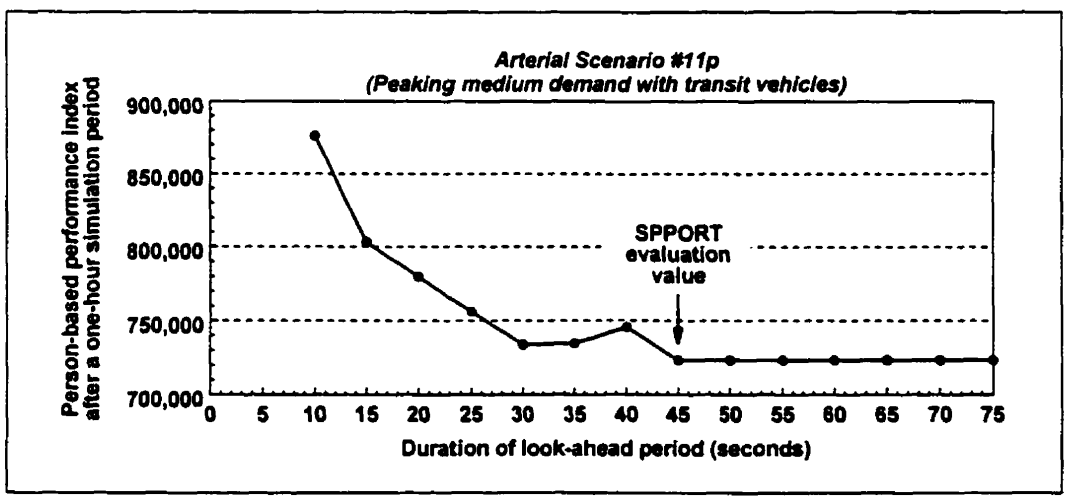


Figure 9.26 - Effect of Duration of Look-Ahead Period on Traffic Control Performance

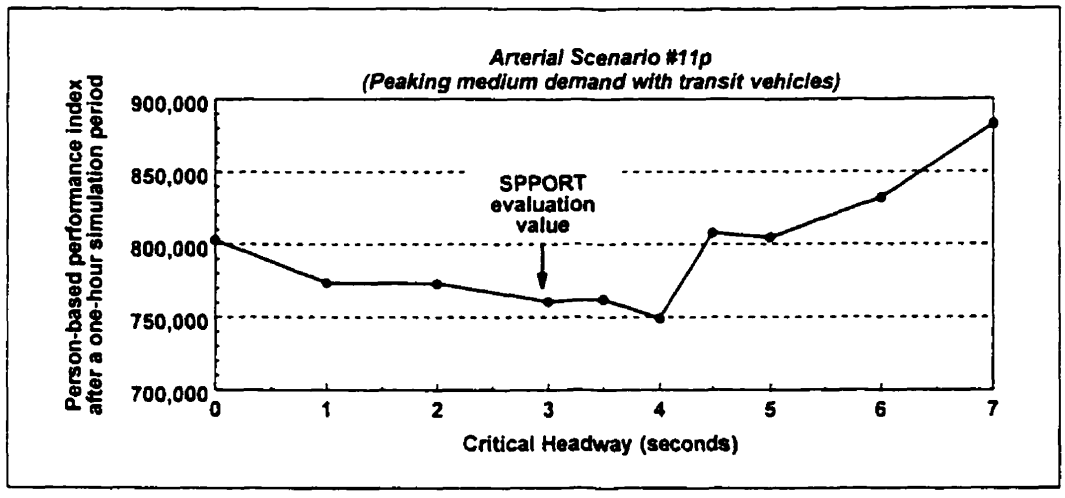


Figure 9.27 - Effect of Stop Coefficient in Performance Function on Traffic Control Performance

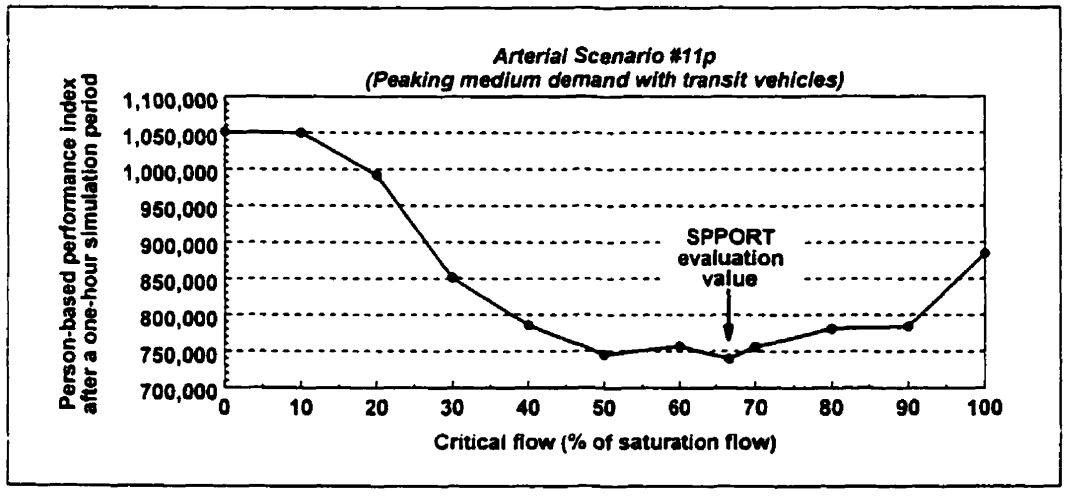


Figure 9.28 - Effect of Critical Flow Parameter on Traffic Control Performance

use within SPPORT are optimal or near optimal with respect to the minimization of the performance index. A second observation is related to the sensitivity of the simulation results to changes in the values assigned to each parameter. For all these parameters, the slope of the performance index relationship is very flat in the vicinity of the selected values. This indicates that model results are rather insensitive to small changes in the parameter values. However, the figures also indicate that large deviations from the selected values can have a more significant impact on the simulation results. While the above conclusions are made on the basis of a single simulated scenario, there is little reason to believe that similar behavior would not exist within the other scenarios used to evaluate the

SPPORT model. As a result, while true optimality cannot be guaranteed, a certain degree of confidence can be put in the assumption that the results reported in the previous sections are typical of what SPPORT can achieved.

Figure 9.25 also indicates that the overall performance of the SPPORT model tends to improve with the use of longer decision horizons. As it can be observed in the diagram, there is an almost continuous decrease in the estimated performance index up to the use of a 60-second decision horizon. Past this point, which corresponds to the duration of the decision horizon used in all the simulations conducted in this thesis, no further improvements are observed.

These results are consistent with expectations. As the duration of the decision horizon increases, signal optimizers are able to plan signal switches further ahead in time. This ability gives them the opportunity to better evaluate the future impacts of current signal-switching decisions by allowing them to project further in time the traffic conditions that would result from their proposed signal control strategy. Similarly, the availability of projected signal switches over longer periods at all controlled intersections allow the signal optimizer of each intersection to better plan its signal control strategy as a function of the projected signal control strategy of other optimizers at adjacent intersections. Past a certain point, a 60-second decision horizon in this case, no further improvements are observed as the benefits of allowing signal controllers to plan future signal switching decisions are completely offset by the increasing inaccuracy of traffic projections in the later part of the horizon.

Figure 9.26 indicates that the performance of the SPPORT model tends to improve with the use of longer look-ahead periods at each decision point within a decision horizon. As the interval over which stop line arrivals are projected is increased, the signal optimizers obtain a better picture of future projected arrivals and are consequently able to generate signal requests that better matches the true needs of the existing traffic demands. This effect is particularly significant in the generation of signal requests attempting to accommodate incoming platoons and transit vehicles. If the look-ahead period is short, approaching platoons and transit vehicles may only appear in the stop line arrival projections when they are very close to the intersection stop line. In many cases, such late detection may not allow the signal optimizers to switch the signal display in time to serve the incoming transit vehicle or platoon without causing unnecessary stops. This is especially true in situations in which recently implemented signal switches constrain the time at which the signal displays can be returned to green on the approach under consideration as a result of minimum green interval requirements and other phasing

constraints. Past a certain point, in this case a 45-second projection horizon, no further improvements are observed in the performance of the model as extending the interval over which stop line arrivals are predicted on each approach no longer affects the number of requests generated by SPPORT or the priority level assigned to these requests.

9.9. Real-time Control Applicability

As indicated in the research goals (see Chapter 4), any real-time, traffic responsive signal control system must be able to implement signal control methods that are suited to the objective of quickly responding to traffic demand variations. Such an objective can only be attained through the implementation of signal optimization methods that are capable of generating efficient signal timing plans in a few seconds.

For the development of the SPPORT model, a target optimization time of five to ten seconds has been established. Such a short target is not unreasonable given than some existing real-time traffic signal control systems are already designed to perform signal optimizations in less than 10 seconds (see Chapter 3 and Appendix A). For example, the UTOPIA system currently performs signal re-optimization at the intersection level every six seconds. In the case of PRODYN, signal optimizations are performed every five seconds. In addition, it is also expected that some of the RT-TRACS prototypes will feature such short intervals between signal optimizations once their development is completed.

In order to evaluate the real-time applicability of the SPPORT model, additional simulation runs were conducted with the objective of determining the time required by the model to perform signal optimizations at individual intersections. To perform these tests, the isolated intersection scenarios with low, medium and high demands, peaking arrival patterns and no transit vehicles were selected (Scenarios #1p, #2p and #3p of Table 9.1). Scenarios with peaking arrival patterns were selected to determine the variability of the optimization times under varying traffic conditions. Isolated intersection scenarios were used rather than arterial scenarios to reflect the fact that the SPPORT model has been designed as a fully distributed signal control system in which signalized intersections are individually optimized using parallel processing. While the optimization times could be recorded for each intersection in the arterial scenarios, these scenarios would not provide a correct estimate of

the time required by the signal controllers to perform data exchanges between signal optimizations, since the simulation is not executed in a distributed fashion.

Figure 9.29 presents the results of the simulation runs that were performed on a 233 MHz Pentium II personal computer with 64 Mb of memory and using Windows 95 as an operating system. For each scenario, the diagram reports the minimum and maximum optimization times that were observed over the course of one-hour simulations using different number of traffic cops. Similar to the evaluation tests performed in the previous sections, each simulation was allowed to run for five minutes before starting to record optimization times. In each case, it was also assumed that signal optimizations took place every five seconds. This allows the recording of a total of 720 optimization times for each simulation scenario.

In the figure, it is observed that both the minimum and maximum optimization times for both the low and medium demand scenarios remain below the five-second target for all combinations of traffic cops. While the minimum optimization times recorded for the high demand scenarios show a similar behavior, very long maximum optimization times are observed for the cases in which five and six

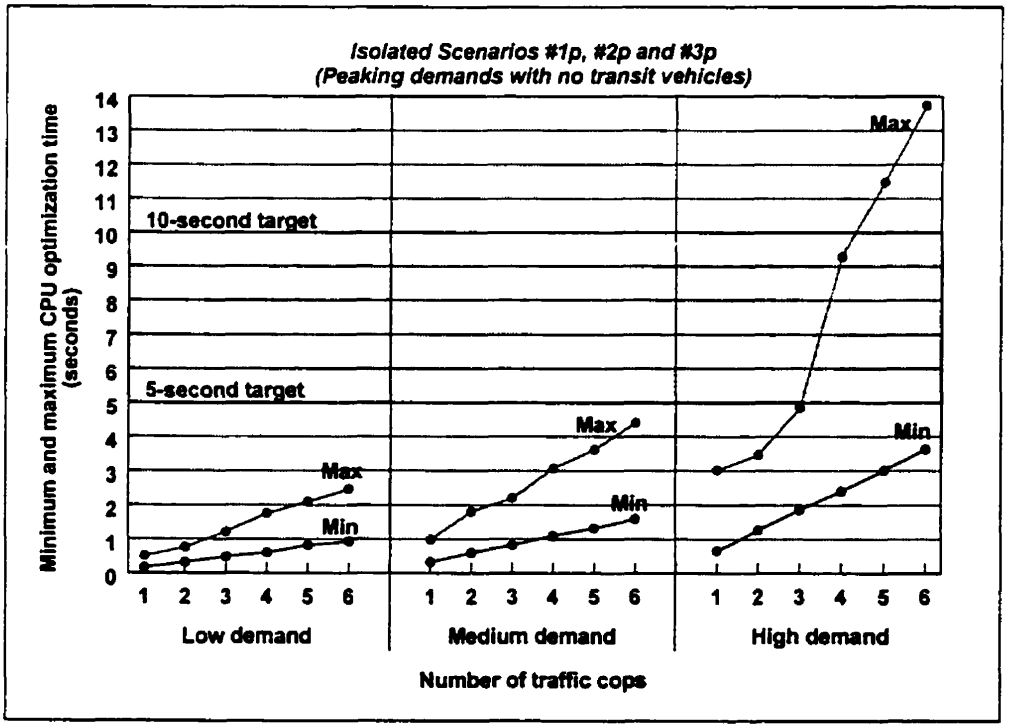


Figure 9.29 - SPPORT Traffic Signal Optimization Times on a 233 MHz Pentium II Computer

traffic cops are used to generate candidate signal switching decisions. These long optimization times are not caused by problems associated with the signal optimization routines, but by the discrete-event simulation model used to project stop line arrivals and evaluate the potential effects of candidate-switching decisions. Like all microscopic simulation models, the time required to perform traffic simulations within the discrete-event simulation used by the SPPORT model is a function of the number of vehicles within the simulated network. The greater the number of vehicles, the longer it takes to perform a simulation. In addition, the relationship between simulation time and number of vehicles in the system is not linear.

Despite the above problem, it can be concluded that the SPPORT model that has been developed has part of this thesis, meets, at least in simulation, the requirements for its application in a real-world real-time traffic signal control system. For both the low and medium demand scenarios, the observed optimization times are well below 10 seconds, which would leave enough time for the signal controllers to communicate with their neighbors between the optimizations and exchange pertinent traffic and signal information. For the high-demand scenarios, the ten-second target time is not attained in all the cases; however, it can be assumed that continued improvements in computing performance capability will reduce the optimization times. It is also estimated that improvements resulting in a reduction of simulation time could be made in the way the simulation model used by SPPORT operates. For example, simplifications could be made in the way vehicles queued at a red signal are handled so that the simulation model would not have to determine whether or not these vehicle can move at the occurrence of every event when the signal display has not changed.

10. Summary, Conclusions and Recommendations

The main objective of the research described in this thesis was to develop efficient real-time, traffic responsive signal control strategies for urban signalized networks that would solve the current limitations of existing real-time traffic signal control systems with respect to their ability to consider transit-related issues in mixed-traffic environments. The first limitation is concerned with the modeling of transit interference on streets on which passenger cars and transit vehicles share the right of way. Existing systems currently assume that transit vehicles do not interfere with the progression of other traffic. In reality, transit vehicles stopping in the right of way often cause non-negligible disruptions to traffic flows, which may then result in a significant loss of traffic control performance at signalized intersections. The second limitation is associated with the way priority is granted to transit vehicles approaching the intersection. Existing systems do not evaluate the consequences of altering the signal timings to accommodate approaching transit vehicles. Green interval extensions and phase recalls are often implemented with the sole objective of avoiding delays to the approaching transit vehicles, without due considerations for the additional stops and delays that might be imposed to other traffic.

The purpose of this chapter is threefold. First, to briefly describe the solution that has been adopted to address the two major limitations of existing real-time signal control systems. Second, to highlight the main findings of the research regarding the application of the procedures developed to provide real-time traffic signal control in an urban environment with mixed-traffic conditions. Third, to present some directions for future research in the field of real-time traffic signal control for urban networks.

10.1. Summary of Modifications to the SPPORT Model

To achieve the stated objective, it has been proposed to expand the applicability of the SPPORT model (*Signal Priority Procedure for Optimization in Real-Time*) to actively consider networks of coordinated signalized intersections. This model, which considered signalized intersections

individually, solved the two main limitations of existing real-time traffic signal control systems through the two following unique features:

- The use of a discrete-event microscopic traffic simulator that explicitly models transit interference on other traffic during dwell times when predicting traffic behavior around controlled intersection.
- The use of a heuristic rule-based signal optimization process that allows the generation and evaluation of several candidate signal timings strategies before choosing for implementation the one yielding the best performance.

The research described in this thesis consisted of expanding and substantially enhancing the existing SPPORT model to enable it to be applied to a network of signalized intersections. To expand the model to the network level, signal optimization procedures and data communication processes had to be developed to allow the model to coordinate the operation of adjacent intersections. At the same time, it was required to maintain sufficient flexibility of operation at individual intersections to allow the model to react quickly to changes in traffic demands and transit priority requests. To guide the development of the proposed model, the four following general prescriptions were stated:

- The system must be truly demand-responsive in that it must adapt to actual traffic conditions and not to historical or predicted values that may be far from reality.
- The system should not be arbitrarily restricted to control periods of a specified duration but should be capable of updating plans at any time, at any location.
- The system must implement signal control methods suited to the objective of quickly responding to traffic demand variations.
- The system must be designed to provide better performance than off-line signal optimization methods.

The above objectives were achieved by developing a fully distributed signal control system in which signalized intersections are individually optimized on the basis of traffic and signal information describing traffic conditions on each approach to the controlled intersection and at adjacent intersections.

In order to provide truly demand-responsive signal control, traffic control is primarily based on traffic information provided by traffic detectors located some distance upstream from the stop line on each intersection approach. These detectors provide SPPORT with advance information regarding future

stop line arrivals. In order to provide priority to transit vehicles, they must be able to selectively detect transit vehicles from other traffic on approaches on which passenger cars and transit vehicles share the right of way. In order to extend the period over which future arrivals are known, while maintaining reasonable accuracy of predictions of future arrival patterns, each signal controller also receives a projection of vehicle departures from adjacent intersections under the proposed local signal control strategy. It is only when traffic information is not available that SPPORT uses artificial traffic generators to predict future traffic behavior.

To provide quick responses to detected changes in traffic demands, signal control is performed using a rolling horizon process in which projected signal timings over the next minute or so are typically updated every five seconds. To provide increased flexibility in the operation of individual intersections, signal control is also achieved on a phase-by-phase basis, with no explicit reference to the traditional concepts of cycle time, green allocation and signal offset. At each decision point, the main control decisions are whether or not to end the current phase, and if the current phase is to be ended which phase to go to next. These decisions are taken on the basis of minimum and maximum green interval constraints, and on the need to coordinate the signal operation with adjacent intersections. There is also no common cycle time imposed on groups of coordinated intersections, as is commonly done in some real-time signal control systems.

A unique feature of the SPPORT model is the use of a heuristic rule-based signal optimization procedure based on the recognition that efficient signal-switching decisions usually occur after the realization of specific discrete events. Examples of such events are queues of vehicles reaching a certain size, dissipation of queues or transit vehicles arriving or completing their dwelling process. By ignoring all events that have no importance for the signal operation, the model is able to significantly reduce the number of potential switching combinations that need to be considered to find an optimum solution to the traffic control process. This process, which replicates the strategy that many traffic control officers adjust, make the SPPORT model more amenable to real-time control than exhaustive optimization methods such as dynamic programming.

At each intersection, the signal optimization process starts with the analysis of projected stop line arrivals and the evaluation of traffic events defined as having an influence on the signal operation. Each event results in the generation of a request calling for either a green or a red signal display on a specific approach at a specific time. To account for the fact that different events do not carry the same

importance, a priority level is assigned to each request. This level is determined on the basis of user-defined values, and is also dependent on the size of existing queues and platoons for requests dealing with approaching platoons and existing stop line queues.

Coordination with downstream intersections is considered by adjusting the times at which green and red signal displays are requested on each approach. Adjustments are made on the basis of the projected signal timings at the downstream intersection, the queue dissipation times along the approach to that intersection, and the level of interference caused by transit vehicles stopping in the right of way to board and discharge passengers on exit links. Coordination with upstream intersections is enhanced by examining the potential for queue spillback across the upstream intersection, the recently implemented and projected timings at that intersection, and queuing conditions on the link joining the intersection with the one being optimized.

Once all requests have been generated, a decision to switch the signal or not is made with the objective of accommodating as well as possible the various requests for green and red signal displays that have been generated in the demand evaluation process. A signal timing plan is then produced by repeating the decision-making at fixed intervals (typically five-second long) within the decision horizon. If more than one prioritized list of events is provided for consideration by SPPORT, a candidate timing plan is generated by the model for each list. The one yielding the best projected performance over the decision horizon is then selected for implementation. However, plans are implemented with a commitment to the first few seconds only (typically five seconds) and are reevaluated at the next decision-making time (typically five seconds from the current decision time).

10.2. Conclusions

On the basis of a series of simulations evaluating the operation of the SPPORT model under a range of traffic conditions at an isolated intersection and along a five-intersection urban arterial, the following conclusions can be drawn regarding the extensions to the heuristic rule-based signal optimization process developed in this thesis to consider transit related issues in mixed-traffic environments:

1. Effective transit priority treatments can be developed by applying the two transit-related rules defined within the SPPORT model. Unlike existing real-time traffic signal control systems, which typically only attempt to provide transit vehicles with uninterrupted progression across

controlled intersections, the rule-based signal optimization process developed in this thesis also attempts to provide these vehicles with a clear passage up to their loading point. In a majority of scenarios, assigning a high priority level to these two transit rules resulted in a significant decrease in the amount of stops and delays incurred by transit riders when compared to a signal operation in which the transit rules were not considered.

2. The multi-objective signal optimization process used by SPPORT allows transit priority requests to be granted only if such preferential treatment will not negatively affect the overall performance of the intersection under control. In some of the scenarios that were simulated significant stop and delay reductions were obtained, on an overall person basis, by unconditionally awarding priority to all approaching transit vehicles when compared to scenarios in which priority is not provided or in which transit movements are only passively favored. However, for each scenario the best or near best performance was obtained when the signal optimization process was allowed to compare strategies providing transit priority with strategies that did not. These results clearly indicate the difficulties faced by traffic engineers in determining a priori which signal control strategy should be the best. They also demonstrate the benefits of allowing a signal control system to continuously evaluate alternative control strategies and change their active strategy when new traffic conditions are detected.

3. By combining requests considering general traffic and transit needs in a single decision-making process, the rule-based signal optimization process followed by SPPORT has the ability to consider all the effects of transit priority treatments and to minimize their potential negative effects on general traffic. For example, extending a green signal indication to accommodate an incoming transit vehicle often causes larger than usual queues to be formed on the cross-streets. In such situation, the SPPORT model is able to temporarily increase the cross-street green duration after each main-street extension by assigning higher priority levels to the signal requests generated by the cross-street queues. This response allows the queue to be quickly dissipated and minimizes the negative impacts of the implemented transit priority treatments. In the traffic networks scenarios examined in this thesis, the implementation of transit priority treatments generally increased stops and delays for non-transit riders. However, the same treatments also resulted in reductions in the total amount of stops and delays incurred by all network users.

4. By considering the effects of transit vehicles stopping in the right of way to load and unload passengers, the SPPORT model is able to make more accurate predictions of vehicle arrival patterns at controlled intersections, and consequently, better able to plan efficiently signal control operations. For the various scenarios that were examined, the benefits of using the SPPORT model over an optimal fixed-time operation generally increased with the severity level of transit interference. These benefits were mostly attributable to the increasing gains of switching the green signal to other approaches each time a transit vehicle stopping in the right of way impeded traffic at a transit stop. This indicates a clear effect of transit interference on the performance of signalized intersections that should not be neglected in any signal optimization process.

On the basis of the scenarios that were simulated in this thesis research, the following general conclusions can also be made regarding the ability of the basic SPPORT-based signal control processes and the enhancements developed in this thesis to provide efficient and suitable real-time, traffic responsive signal control at isolated intersections and in coordinated urban networks with mixed-traffic:

1. The application results clearly demonstrate the ability of the rule-based signal optimization process to provide efficient responses to traffic demand variations. For example, the application of the SPPORT model reduced total vehicle passenger delays at an isolated intersection by as much as 35 percent in scenarios involving transit priority, transit interference and peaking arrival patterns when compared to an optimal fixed-time operation. Benefits were also obtained in scenarios evaluating the operation of the model along a five-intersection arterial. In these cases, reductions of up to 50 percent were observed in the value of a performance index considering the total amount of stops and delays incurred by all vehicle passengers.

2. The application results show a certain inability for the rule-based optimization process to improve traffic conditions over an optimal fixed-time operation in scenarios involving no transit vehicles, and more particularly in scenarios with constant vehicle arrivals. This inability is not surprising given that fixed-time signal optimizations are specifically designed to produce optimum timings for these conditions. However, other factors may also have contributed to these negative results. One potential factor is the possible non-optimality of the

priority levels assigned to the various requests for green and red signal displays that are generated during the demand evaluation process. Another factor is concerned with how priorities are combined before determining which phase would best meet the needs of current demands. Currently, the ability of a given phase to serve current traffic demands without delaying service to future high priority requests is determined by simply adding and subtracting request priorities. This process assumes a linear relationship between the number of request served and their overall priority level while the ability to serve more than one request at once may command an overall priority level greater than the sum of the individual priorities.

3. When attempting to coordinate the operation of an intersection with its immediate neighbors, benefits were generally obtained by instructing each signal optimizer to consider the projected signal switches at both upstream and downstream intersections, as well as the projected departures from upstream intersections. In almost all the scenarios that were considered, additional benefits were also obtained by allowing the signal optimization process to adjust the times at which green signal displays are requested as a function of both upstream and downstream traffic conditions. When considered altogether, these results point to the effectiveness of signal coordination process developed in this thesis. They also demonstrate the ability of the ruled-based signal optimization process to successfully coordinate signalized intersections in a control environment in which signalized intersections are individually optimized.

4. While benefits were generally obtained when projected signal switches at adjacent intersections were only considered, the consideration of projected departures from upstream intersections only generally resulted in poorer traffic control performance. These negative results are not attributed to inaccuracies the prediction of stop line vehicle arrival patterns but rather to the fact that the rule-based signal optimization process currently assigns a priority level to traffic events without considering their time of occurrence. When projected upstream departures are not considered, almost all the signal requests generated during the demand evaluation process are associated with current or imminent events. When stop line arrival patterns are estimated using departures from upstream intersections, additional requests dealing with events scheduled to occur at some time in the future are added to the list of requests. If the priorities assigned to these requests are sufficiently high, situations can then arise in which a green signal is held on an approach while serving no or very little traffic for

the only purpose of avoiding delay to an incoming platoon or transit vehicle that might still be 10 or 20 seconds away from the intersection. While such control decision result in a poor use of green time, it can often be avoided by instructing the signal optimizer to evaluate a series of candidate controls strategies before making any decision. However, there is no guarantee that all defined control strategies will not lead to the same decision. One way to solve the problem is to assign lower priorities to future traffic events. This lead to the conclusion that truly general optimal control can only be achieved with the proposed rule-based signal control process by considering the temporal relationship of traffic events in the signal design process.

5. In the various tests that were conducted, the best signal control strategy was generally obtained with the optimization evaluating the largest number of signal control strategies before making any signal-switching decision. These results are a strong indication of the difficulties faced by traffic engineers in determining a priori which signal control strategy would produce optimal, or near-optimal, signal switches at each decision point. While the use of a single strategy often appears to be sufficient when adequate traffic conditions are considered, better responses to detected changes in traffic demands are provided by not constraining the signal operation to one single strategy. As a result, the ability to consider various control strategies at each decision point will often open the doors for significant improvements in traffic signal control performance.

6. By combining requests attempting to prevent queue spillbacks with requests attempting to provide service to existing queues, incoming platoons and incoming transit vehicles, the ability is given to the rule-based signal optimization process possesses automatically adjust its signal control strategy to prevailing traffic conditions. For example, a switch from the minimization of stops and delays to the maximization of traffic flow across controlled intersections would automatically occur when stop line queues would threaten to spill across upstream intersections if the highest priorities are assigned to the queue spillback rules. Similarly, a return to the initial control strategy would also occur when the congestion would have dissipated. This is an advantage for the rule-based signal optimization process described in this thesis as its ability to switch signal controls strategies provide the necessary flexibility to allow true real-time, traffic responsive signal control to take place.

7. Simulation results indicate that rule-based signal optimization process developed in this thesis is currently applicable to real-time traffic control. Similar to existing real-time traffic signal control systems for coordinated urban networks, the SPPORT model was able to generate efficient signal timing plans within a 10-second target optimization period in a majority of cases. These results are based on the simultaneous consideration of up to six control strategies. While the target optimization time was exceeded when high demand levels were considered, it is estimated that improvements in computer technology could further increase the applicability of the model.

8. A sensitivity analysis revealed that the rule-based signal optimization process is rather insensitive to changes in the decision horizon, the traffic projection period at each decision point within the decision horizon, the critical headway used to determine whether two vehicles are traveling closely, and the critical flow used to determine if a given stop line flow rate is high enough to warrant the generation of a given request, at least around the values assigned to these various parameters in the simulation studies. The same analysis also revealed that the use of longer decision horizons is generally preferable over shorter horizons. Another conclusion is that increasing the duration of the traffic projection period only affects the performance of the model up to the point where the availability of any additional traffic information does not result in the generation of new signal requests or the modification of existing ones.

10.3. Recommendations for Further Research

1. In completing this research project, numerous studies were performed to demonstrate the efficiency of the proposed SPPORT model. However, none of the tests that were conducted fully considered the variability of real-world traffic demands. While traffic demands are normally subjected to numerous sources of randomness, the only source of randomness in the simulations that were conducted was how individual drivers selected their path where traffic split into through and turning movements. In all scenarios, vehicle speeds and transit dwell times were assumed to remain constant over time and not to vary from one vehicle to the other. In addition, while the average arrival rate was varied at specific points in time in the peaking demand scenarios, vehicle inter-arrival times were assumed constant between each one of these points. In order to determine the behavior of the SPPORT model under traffic demands that would be more representative of the

reality, it is therefore suggested that additional evaluations with scenarios considering randomly variable vehicle inter-arrival times, vehicle speeds and transit dwell times be performed.

2. In addition to the need to evaluate the behavior of the SPPORT model under more realistic traffic demands, additional studies should also be conducted for determining the effects on the operation of the model of changes in the values assigned to each one of various signal optimization configuration parameters. Such studies could notably help devising guidelines for the selection of appropriate control parameters in given control situations. A first step in that direction was taken in the thesis when the sensitivity of the signal operation to the values assigned to selected control parameters was examined. However, the evaluation work only considered the parameters that were thought to have the greatest impact on the signal operation. One particular element requiring further investigation is the selection of appropriate priority levels for each rule of the signal optimization process. In the current research projects, priority levels were selected on the basis of engineering judgement and test simulations. However, given the complex interdependencies between traffic events (one event often leads to another one) it is unclear whether the selected values were truly optimal for the test scenarios considered. It is also unclear whether these values could be successfully applied to other test scenarios.

3. As a corollary to the need to determine optimal priority levels for each, there is also a need to determine the optimal number of control strategies that should be considered by the SPPORT model to produce optimal signal switching decisions. In the current research project, a maximum of six different control strategies were used, with the only elements differentiating the lists being whether or not priority is provided to transit vehicles and on which approach incoming platoons were considered. In some scenarios, improved traffic performance might have been obtained by using additional lists of events considering only potential queue spillbacks, or completely ignoring rules dealing with incoming platoons.

4. Additional research is still required to enhance the signal optimization process. For instance, in the evaluation of the SPPORT model, problems were attributed to the fact that the signal optimization process currently assigns a priority level to traffic events regardless of their time of occurrence. Ideally, requests dealing with events projected in the future should receive a lower priority than similar events projected to occur imminently to reflect the fact that these future events are subject

to more uncertainty. It is also still unclear whether linearly adding the priorities of active and inactive requests in the phase selection process leads to optimal signal control decisions.

5. It is anticipated that allowing the SPPORT model to determine the order in which the phases should appear would lead to further reductions in stops and delays. Such results are anticipated as imposing a fixed sequence of phases on the operation of the SPPORT model, as what was done in this thesis, often prevented SPPORT to switch to a particular phase when the phase was first requested. This recommends that the performance impacts of permitting SPPORT to select the phase sequence be quantified for a range of traffic network scenarios.

6. In order to improve the ability of the SPPORT model to realistically predict stop line arrival patterns and accurately evaluate candidate timing plans, further research efforts should also be directed at improving the discrete-event simulation model used by SPPORT. As an example, the model does not currently simulate the dispersion of platoons between successive intersections. This dispersion, which occurs as a result of the differing speeds at which individual vehicles travel within a platoon can significantly affect traffic arrival patterns at signalized intersections, especially along arterials where intersections are relatively far apart. Another limitation is associated with the inability of the model to consider traffic movements on a lane-by-lane basis. Currently, traffic movements are simulated using single-file, unidirectional segments. Consequently, the model may not be able to correctly simulate traffic behavior on approaches on which through and turning movements share a traffic lane. For example, while a left turn gap acceptance logic was added to the simulation model as part of this thesis, the logic can currently only be applied to segments used to represent single-lane approaches or exclusive turning lanes. Since it is assumed that all vehicles travel in a single file, the use of the logic on a segment carrying both through and turning movements would cause a complete blockage of the segment each time a left-turning vehicle would have to stop to wait for an appropriate gap in the opposing stream, even if the left-turning vehicle only blocks in reality one of the two or more available traffic lanes.

7. In the thesis, the problem regarding the accurate simulation of left-turning movements from a traffic lane shared by both through and turning movements was not considered as the main focus of the research was to develop signal control procedures providing efficient responses to transit interference and transit priority requests in coordinated urban networks. In each scenario, left turns

from the main-street were all assumed to be made from an exclusive left turn lane. This allowed a correct usage of the gap acceptance logic to simulate left-turning behavior. Since left turns from the minor streets were assumed to be made from a shared lane, the gap acceptance logic was not applied to these movements. This modeling choice introduced some inaccuracies in the simulation of cross-street traffic, but such inaccuracies were judged reasonable on the basis of the main objective of the research. In order to determine the impact of this modeling choice on the performance of the SPPORT model, it is therefore recommended that additional simulations be performed assuming the availability of an exclusive left turn lane on all cross-streets and the use of the gap acceptance logic by all left-turning vehicles. Since many sources of traffic variability were neglected in the scenarios considered in this thesis, it is also recommended that signal control performance evaluations be made with scenarios considering mid-block entry and exit flows, as well as variable inter-arrival times, speeds and transit dwell times.

8. Other improvements can also be made to reduce the time required by the discrete-event simulation model to simulate traffic behavior. Such reduction could greatly reduce the time required by SPPORT to generate signal timing plans since the time required by the discrete-event simulation model to project stop line arrival patterns and evaluate candidate timing plans accounts for most of the total optimization time. In this case, the time required to perform the necessary simulations is a function of the magnitude of the demand to simulate. The larger the number of vehicles being simulated is, the longer it takes to perform the simulations. This relationship was observed in the simulation tests, were significantly longer optimization times were required for the high demand scenarios than for the low demand scenarios. While the development of faster computers might preclude the need for a faster simulation model, any modification to the existing model that might significantly improve its efficiency should not be neglected. One particular point that merits investigation is the way stopped vehicles are processed within a given simulation. For example, all stopped vehicles are transferred to the current event chain from the future event chain each time the simulation clock is advanced. As a result, a scan is made by the simulation model to determine whether each one of these vehicles can make a move at the current simulation time even if it obvious that some of these vehicles cannot move. Simulation times could then be reduced by not performing such scans for vehicles waiting at a red signal until the signal display has changed. At this point, it is anticipated that such modification would provide increasing benefits with increasing magnitudes of traffic demands.

APPENDIX A

Existing Real-Time, Traffic-Responsive Signal Control Systems

A.1. SCATS - Sydney Coordinated Adaptive Traffic System

Description based on the following references: Sims (1978), Sims and Dobinson (1979), Fehon and Moore (1982), Lowrie (1982), Sims and Finlay (1984), Luk (1984), Cornwell et al. (1986), Luk and Cahill (1986), Luk and Doughty (1988), Lowrie (1990), and Wood (1993).

A.1.1. Control Architecture

SCATS is a two-level centralized hierarchical signal control system. Its upper control level, hosted by the central computer and known as the strategic control, is in charge of producing the basic signal control strategy. The lower control level, known as the tactical control, is composed of intelligent signal controllers.

The central computer optimizes the signal settings on an area basis based on average traffic conditions. Its role is to react to long-term fluctuations in traffic demand, as well as to changes in the capacity of each intersection that may result from traffic accidents, changing weather, pedestrian friction, etc. Local controllers, on the other hand, are designed to provide more flexibility at individual intersections. They can adjust the signal timings they receive from the central computer to adapt them to the current of level traffic demand at each intersection, but only within specified limits imposed by the central computer.

SCATS controls a network of signalized intersections by breaking it up into a series of smaller sub-networks containing between one and ten intersections. All signals in a given sub-network are coordinated and share the same cycle length. SCATS' strategic control logic can also create larger coordinated area by "marrying" adjacent sub-networks. Two adjacent sub-networks are automatically married if imposing the same cycle length in both and if coordinating the signals located along their boundary result in lower estimates of stops and delays for all vehicles traveling inside these two sub-networks. Married sub-networks are also automatically "divorced" when their association provides no longer an advantage.

A.1.2. Traffic Prediction Model

Detectors located immediately upstream of the stopline monitor traffic demand on each approach to an intersection. The detectors are installed where the first vehicle in a queue would normally stop. They enable SCATS to monitor in real-time the flow of vehicles going through an intersection during each green phase, as well as the detection of phases for which there is no demand, i.e., no vehicle queued at the stopline waiting for the green signal to come on.

The information collected by the detectors is used to estimate the degree of saturation of each approach, which is defined in SCATS to be the ratio of the effectively used green time to the total available green time. It is the main control variable used by the optimization algorithms to determine the suitable signal timings for each intersection.

SCATS does not possess a traffic prediction model. Because traffic detectors are located at the stopline, they cannot provide advance information about future arrivals at the intersection. The detectors only record the passage of a vehicle when it is about to enter the intersection. Therefore, there is no need to use a traffic prediction model to predict near-future arrivals at a given intersection. This forces SCATS to use past demand measurements to calculate its optimum signal timing plans.

The absence of a traffic simulation model and the use of stopline detectors prevent SCATS to simulate the queuing of vehicles at the stopline and to estimate directly the stops and delays generated by the queuing process. The only queue prediction capability that SCATS possesses is its ability to detect the presence of an excessively long queue on an approach through the detection of a decrease in the outgoing flow rate at the upstream intersection. If a queue gets very long and reaches almost the upstream end of a link, fewer vehicles will be able to enter the link from the upstream intersection and, consequently, to leave that intersection.

Luk et al. (1986, 1988) developed a traffic prediction model using traffic data provided by SCATS' stopline detectors that could predict arrivals and simulate the queuing process at the following downstream intersection. However, recent literature about SCATS (Lowrie, 1990; Wood, 1993) does not indicate that this model was finally introduced in SCATS.

The main difficulty associated with the use of a prediction model based on information provided by stopline detectors lies in the estimation of turning percentages at the intersection were traffic in monitored. Because all vehicles in the lanes shared by turning and through movements do not follow the same path, there is a relative uncertainty regarding the actual proportion of vehicles that turn and go straight ahead. The greater this uncertainty, the less accurate is the traffic demand prediction. This may lead the optimization algorithm to develop timing strategies that are more or less appropriate for the existing traffic demand. Estimates could be used, but these will not consider the possible variations of turning percentages over time.

SCATS developers chose to use stopline detectors rather than upstream ones to be able to estimate on-line the saturation flow of each traffic movement. The saturation flow defines the maximum number of vehicles that can cross an intersection during a given period of green time. This maximum flow is function of many factors, such as weather, time-of-day, parking activities, pedestrian friction, downstream conditions, and type of vehicles. Because stopline detectors enable SCATS to directly monitor the traffic when a queue is discharging, the saturation flow can thus be directly estimated. When upstream detectors are used, as in all other models, queue discharge conditions cannot be directly monitored. Therefore, estimated saturation flow rates must be used to simulate the dissipation of queues of vehicles.

A.1.3. Timing Principles

SCATS is designed to minimize stops and delays in each sub-network, but with a great emphasis on the minimization of stops incurred by vehicles on main arterials. SCATS attempts to provide good progression along major streets, rather than to find a global optimum timing policy that would minimize stops and delays at all intersections. When the network is operating near saturation, the control objective may also be biased to maximize throughput along major arterials, i.e., to maximize the green time available at each intersection on these arterials so that more vehicles can travel along them. The following paragraphs describe in more details the timing principles used by the central computer (network, or strategic control) and by the signal controllers (local, or tactical control) to achieve this objective.

A.1.3.1. Network Timing Principles

SCATS strategic control responds to detected changes in traffic demand and in intersection capacity (change in the measured saturation flows) by slowly varying the timing parameters at each intersection. Basic timing parameters are the common cycle length under which all signals are operated in a given sub-network, the green splits at each intersection (proportion of green time dedicated to each phase in a cycle) and the offset between adjacent intersections (interval of time between the apparition of a given phase at the two intersections). The phase sequence is fixed and is then not a control parameter.

The length of the common cycle in a given sub-network is adjusted once every cycle. This common cycle length is equal to the cycle required at a critical intersection defined by the traffic engineer. The cycle can be varied by up to six seconds each time (nine seconds if a trend is recognized) to maintain the degree of saturation of each approach to the critical intersection within user defined limits (usually 90 percent). All other intersections in the sub-network share the same cycle length and are consequently forced to operate with a longer cycle than their optimum.

Green splits can be varied at the critical intersection by up to four percent of the cycle length each cycle to try to maintain an equal degree of saturation on all approaches. It is believed that maintaining an equal degree of saturation contributes to the minimization of stops and delays at a given intersection. For stability purposes, however, calculations are based on the average degree of saturation calculated over the past three cycles to reduce the random fluctuations between two or three different green splits. At non-critical intersections, splits are either non-variable or selected in a plan library by a matching process that selects splits compatible with the ones currently in use at the critical intersection. Because critical intersections typically represent only 10 to 20 percent of all intersections in a SCATS network, this means that between 80 to 90 percent of all controlled intersections do not have their green splits explicitly optimized.

Optimum offsets between signalized intersections in a given sub-network are calculated each cycle for the direction having the highest flow rate, to ensure that the most significant direction of travel receives adequate progression through the system. For stability purposes, offsets are only changed when a modification has been requested three times in the past five cycles. Offsets are also optimized between intersections at the boundary of two sub-networks when these sub-networks are married to establish larger coordinated areas. Otherwise, signals in different sub-networks are not coordinated.

A.1.3.2. Intersection Timing Principles

Local controllers can shorten or entirely skip a phase according to the level of demand measured by the stopline detectors on the corresponding approaches.

The adjustment of signal timings provided by the network control level is made by following simple vehicle actuation logic. The duration of a green phase is for instance extended over a certain minimum duration until the gap between two detected vehicles at the stopline exceeds a threshold value or until the maximum allowable extension has been reached. A green phase can be completely skipped if no vehicle has been detected to reach the stopline on the links served by that phase during the preceding red phase.

All these adjustments are subject to restrictions imposed by the network control level. Moreover, one phase cannot be omitted of shortened. This phase is defined by the user of the control system, but is usually associated with the phase giving the green indication to the main road.

A.1.4. Priority to Transit Vehicles

Passive priority treatments can be granted to transit vehicles by allocating a favorable bias on the link traveled by these vehicles. The bias can be permanent, modified by time-of-day or by operator command.

Active bus and tramway priority treatments have also been implemented in some SCATS systems (Wood, 1993). In these systems, when a transit vehicle is detected, the facilities exist to switch to a particular phase or to retain the current green phase, if appropriate, to avoid unduly delaying an incoming transit vehicle detected on one of the approaches to an intersection. An automatic adjustment of green splits may also follow the change made to minimize the impact on other approaches of suddenly reducing their green time, as in the system implemented in Melbourne, Australia (Cornwell et al., 1986).

Because SCATS' stopline detectors cannot provide advance information about the arrival of transit vehicles, separate detectors must be used. For instance, in the tramway detection system implemented

in Melbourne, additional detectors were installed 200 metres upstream of the stopline to give an appropriate advance warning about the arrival of the tramways having priority of passage.

A.1.5. Implementation Results

SCATS has been installed so far in over 30 cities around the world. This fact alone is a good sign of its ability to control in real-time networks of signalized intersections.

Very few results regarding SCATS' performances are available in the literature. The most recent one is a survey that was carried out in 1981 by the Australian Road Research Board in Paramatta, Australia (Luk, 1984). In that survey, SCATS' performance was compared with a fixed-time plan optimized with TRANSYT. It was found that SCATS and TRANSYT performances did not differ significantly in the highly constrained downtown network of Paramatta. SCATS increased travel times by six percent and decreased stops by nine percent during the afternoon control period. These poor results were explained by the impossibility, either for TRANSYT or SCATS, to establish a good progression scheme on most of the major routes of the downtown network. A second test on a less constrained arterial showed that SCATS consistently performed better than TRANSYT. It did not significantly improve the travel times (4 percent reduction, not significant at the 95 percent confidence level), but reduced stops by about 25 percent. Luk also indicates that other studies have demonstrated improvements in travel times, but that they derived their results by comparing SCATS with fixed-time control systems that were not necessarily operating optimum timings.

The Paramatta evaluation work did not look at priority to transit vehicles. According to Wood (1993), there is no performance results reported in the literature for networks in which signal timings ca be changed to offer priority of passage to transit vehicles. The only performance results reported are for isolated intersections only.

A.2. SCOOT - Cycle, Split and Offset Optimization Technique

Description based on the following references: Hunt et al. (1981), Robertson (1986), Wood (1993), Bowen et al. (1994), Martin et al. (1995), Hounsell et al. (1996), Bretherton (1996), Bowen and Bretherton (1996).

A.2.1. Control Architecture

SCOOT is a fully centralized control system, in the sense that all timing calculations are done by the central computer. Local controllers have no adaptive or optimization capability of their own. They only implement the timings they receive from the central computer.

This control architecture enables SCOOT to gather at a single location all data required for signal optimization. This allows the central computer to attempt to perform a system-wide optimization off all timing parameters. As such, it readily allows for pro-active control, where the central system can attempt to simultaneously optimize all the signals under its control, providing for progression between pairs of intersections, while at the same time pre-conditioning at each intersection the arrivals from upstream intersections to actively promote progression.

This pro-active control ability is amplified by the fact that SCOOT can coordinate large number of signalized intersections. Like SCATS, SCOOT controls a signalized network by breaking it up into a series of smaller sub-networks. However, SCOOT's sub-networks are much larger than those used in SCATS. The only limit put on the maximum size of a given sub-network is the available computer power. This increases SCOOT's ability to produce global optimum timing plans for signalized networks.

A.2.2. Traffic Prediction Model

Contrary to SCATS, traffic demand is monitored by detectors located at the upstream end of each approach to an intersection. These detectors enable SCOOT to obtain the most direct advance information about future arrivals at the downstream intersection. They also enable SCOOT to know if

a queue of vehicles fills entirely and approach and, consequently, to know if vehicles can or cannot leave the upstream intersection.

The traffic flows observed by the detectors on each approach are stored in SCOOT's central computer in the form of cyclic flow profiles (see figure A.1). A cyclic flow profile can be though of as a histogram giving the average number of vehicles that were observed to pass at a given location during each of the intervals dividing the current cycle length. However, SCOOT's cyclic flow profiles do not truly represent future arrival. Although the upstream detectors provide advance arrival information,

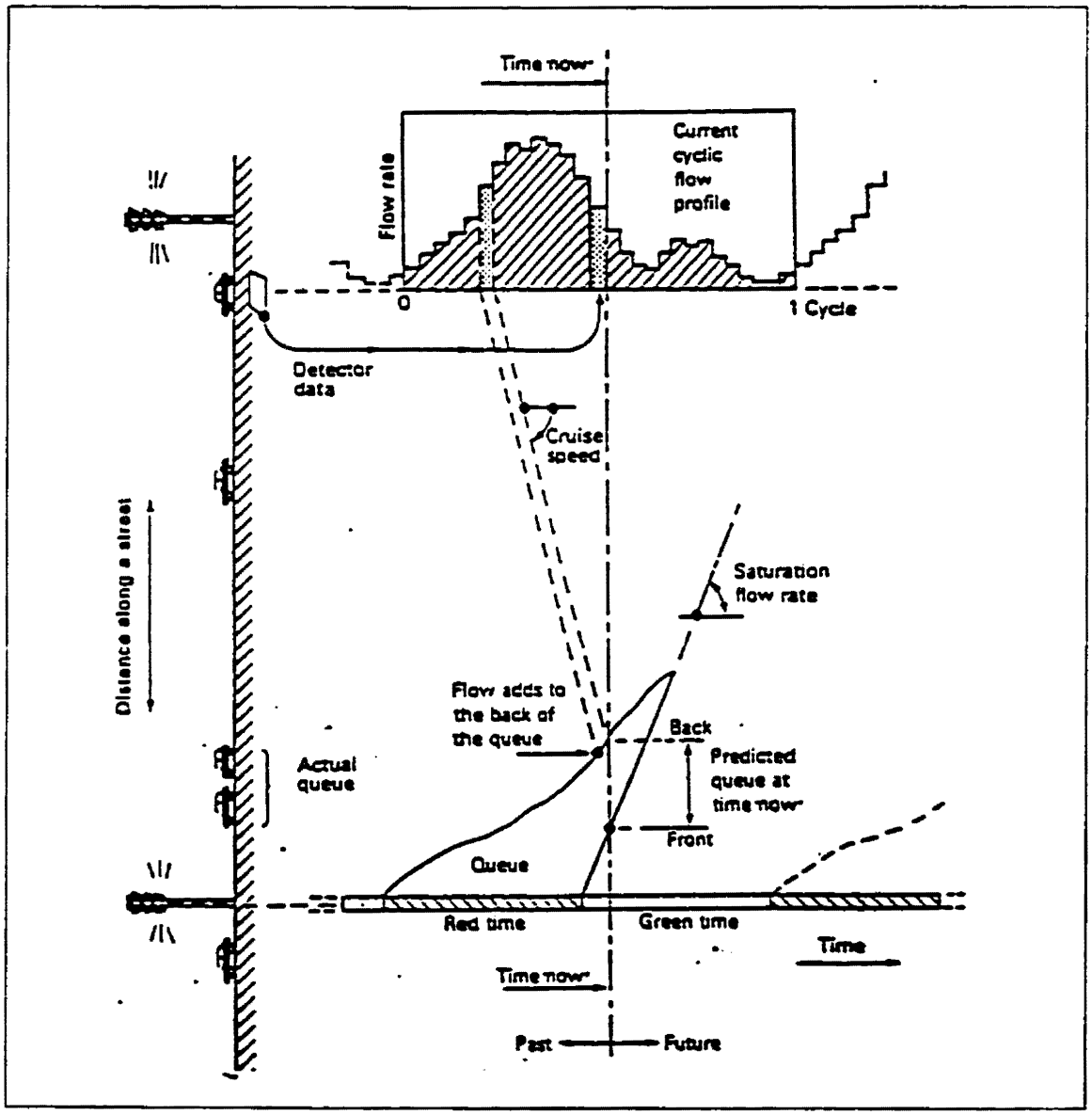


Figure A.1 - Principles of the SCOOT traffic model (source: Robertson, 1986)

SCOOT's flow profiles only represent the average number of vehicles that were observed during each four-second interval over the past few cycles. SCOOT is effectively designed to react to long-term slow variations in traffic demand, not to short-term random fluctuations. For stability purposes, the most recent data on traffic flows for the current interval are combined with existing values in the current flow profiles using a moving average process so that unduly large fluctuations in these profiles are avoided.

Traffic demand at the stopline is estimated by projecting to the stopline the shape of the upstream cyclic flow profile. The profile is translated into time, to take account of the average travel speed between the upstream detector and the stopline, and passed through a platoon dispersion model that spreads vehicles in time. Once this arrival profile is known, a simple queuing model is used to simulate the evolution of the queue of stopped vehicles over time under a specified signal control strategy and to estimate the stops and delays generated by this queuing process.

For simplicity, the queue is modeled vertically. Therefore, vehicles are assumed to travel at free speed until they reach the stopline. This affects the accuracy of the prediction, as vehicles will usually reach the back of the queue before the stopline. The queue estimated by SCOOT may thus be shorter than the actual one. This gives more importance to the upstream detectors, since they can detect a queue reaching the upstream end of a link and warn the central computer about this situation. Prediction accuracy will also depend on the accuracy of the assumed average travel time and saturation flow rate on each approach.

A.2.3. Timing Principles

SCOOT is designed to find a global system optimum by minimizing stops and delays incurred by all vehicles or persons traveling inside the controlled network. There is no a priori preferences given to the optimization of signal timings along major routes as in SCATS, except when required by the traffic engineer. All approaches to an intersection and all intersections are equally important.

SCOOT optimizes signal timings by acting on the three traditional timing parameters of a coordinated system: cycle length, green splits, and offsets. Its basic control philosophy is to react to variations in traffic conditions by implementing frequent, but small, changes to these three timing parameters. The phase sequence, however, is fixed.

As explained, changes are not made to respond to all detected variations in traffic demand. SCOOT only reacts to permanent changes in the average traffic conditions that reflect themselves in the average cyclic flow profiles, i.e., that are only apparent after a few updates. SCOOT was designed to control networks in which traffic demand patterns typically repeat from one cycle to the next.

Changes in the common cycle length of each sub-network are tested every 2.5 minutes or so. The cycle is allowed to change by up to four seconds each time. The objective of this change is to keep the degree of saturation at the critical intersection in each sub-network at or near a user-defined degree of saturation (usually 90 percent). However, the critical intersection is not pre-defined. SCOOT algorithms determine the critical intersection based on the estimated degree of saturation of each intersection. The one with the highest degree of saturation is selected as the critical intersection. The cycle length retained is the one that results in the lowest performance function, a weighted sum of stops and delays.

The benefits of extending or shortening the current phase are tested at each intersection each time the phase comes near to its end. The main objective of this change is to maintain an equal degree of saturation on each approach to the controlled intersection. It is assumed that maintaining an equal degree of saturation contributes to the minimization of stops and delays. Initially, the signal optimizer was only allowed to make a temporary change of four seconds and a permanent change of one second. This has been changed in 1995 to make the optimizer more flexible in its response to changes in traffic demand. It is now possible to specify for each node and each cycle time a set of alternative permanent and temporary changes. At each optimization, the signal optimizer considers all the permitted authorities and implement the most beneficial.

Finally, the benefits of altering the offset between an intersection and its four surrounding ones are evaluated once per cycle. The offset selected is the one that minimizes a performance function that is the sum of the performance functions of each of the five intersections involved in the optimization. A network-wide simultaneous optimization of offsets is not achieved as it is considered that operating all mini-networks at their optimum timings should normally result in a network operating near its global optimum. As for the split optimization, a set of different authorities can be specified, resulting in the most beneficial one being implemented.

The optimization criteria can be modified by applying movement-specific weights to the offsets and green splits, typically to give priority to certain routes in the network. For instance, the split weighting facility can be applied to give priority to the main roads at low degrees of saturation. SCOOT may also be instructed to automatically reduce the split weighting as the degree of saturation increases, leading to a normal minimization of stops and delays. When traffic becomes congested, a congestion logic dominates the optimization procedure. Minimization of stops and delays is then replaced by queue management objectives, gating strategies, etc.

A.2.4. Priority to Transit Vehicles

SCOOT can provide both passive and active priority treatments for transit vehicles. On the passive side, links on which transit vehicles are traveling can be systematically favored by using the split and offset weighting features. These features allow the user to provide different weights to different links in the split and offset optimization processes. Fixed offset based on average transit travel times can also be provided. In such case, there would be no offset optimization carried out by SCOOT for the links for which fixed offsets have been provided.

Active priority features have been introduced in SCOOT in 1995. Contrary to passive priority treatments, these features allow the model to favor transit operations only when transit vehicles have been detected to approach signalized intersections. The model currently allows transit vehicles to be detected either by selective vehicle detectors or by an automatic vehicle location system. In the first case, each transit vehicle is fitted with a transponder unit and special inductive loops installed in the pavement are used to transmit a transit identifier code to the corresponding signal controller each time a transit vehicle crosses them. In the second case, transit vehicles are fitted with equipment that allows their location to be determined by dead reckoning. Each transit vehicle transmits its identifier and location by radio when it is polled from the central computer. While these detection systems allow SCOOT to know the location of transit vehicles, there is no attempt to measure or predict the time that individual transit vehicles wait at a service stop. While it is generally best to detect transit vehicles as far as possible from the intersection, this means that detection on a given link must take place downstream of any transit stop.

Once an approaching transit vehicle had been identified, signal timings are optimized to benefit the vehicle by either extending the current green phase or causing successive phases to occur early.

Extensions can be awarded by the central computer or by programming the signal controllers to implement them locally. Extensions directed awarded by the signal controllers are preferred, as they eliminate three to four seconds of communication time and allow controllers to grant extensions to vehicles arriving the last seconds of a green signal. In all cases, SCOOT's central computer remains in control of the optimization by deciding each second if local controllers could grant extensions or not. To avoid oversaturating an intersection, local extensions and phase early recalls are only allowed when the degree of saturation of the intersection at which priority is request is sufficiently low. This recognizes the fact that transit priority would be most effective at intersections that have spare capacity. Preset values based on simulation studies put the priority threshold for green extension at a degree of saturation of 110 percent and for phase early recall at 95 percent.

Once the transit vehicle that had requested priority of passage has passed through the intersection, a period of resynchronization occurs to bring the local timings back into line with the normal SCOOT optimization. Four methods of recovery are provided for operation after green extensions and early recalls. Of these methods, two are recommended for normal use and operate by default.

A.2.5. Implementation Results

SCOOT has been shown to give significant benefits over the best fixed-time operation and is now in use in more than 130 cities around the world.

The effectiveness of the SCOOT control strategy has been assessed by major trials in five cities in the United Kingdom. Comparisons with fixed-time plans optimized with TRANSYT and with simple isolated vehicle-actuated signals showed that SCOOT can significantly reduce travel times and delays in signalized networks. Table A.1 gives the results of these tests. The relative effectiveness of SCOOT varies by area and time of day, but it was concluded that SCOOT can achieve savings in delays of about 12 percent when compared with good fixed-time plans. As shown in Table A.2, significant benefits also seem to emerge from the demonstration project carried out in Toronto comparing SCOOT control with the city's current fixed-time system.

Delay savings due to transit priority were estimated through both simulation studies and field trials in the Candem area of London (Hounsell et al., 1996). The simulation studies indicated average delay savings to buses in that area of typically 20 to 30 percent per bus per intersection, depending on the

priority strategy implemented, the traffic and bus flows, and signaling characteristics at the intersection. These results were then generally confirmed through field trials in the same area. In these trials, average delay savings of five seconds per bus per intersection were measured. This corresponded to a 22 percent reduction. The highest savings (ten seconds per bus per intersection) were achieved at intersections with light traffic using both green extensions and early phase recall facilities. At intersections with higher traffic volumes, the bet results were obtained by using local extensions only. In this case, there was no significant increase in the delay to general traffic. Higher overall delay savings were obtained for transit vehicles by using both green extension and phase early recall, but only at the expense of a statistically significant increase in delay for the general traffic of up to five seconds per intersection.

Table A.1 - Changes in Delay from the use of SCOOT in United Kingdom (source: Wood, 1993)

Location	Previous control	A.M. Peak (%)	Off Peak (%)	P.M. Peak (%)
Glasgow	Fixed-time	+2	-14 *	-10 *
Coventry - Foleshill	Fixed-time	-23	-33 *	-22 *
- Spond End	Fixed-time	-8	-0	-4
Worcester	Fixed-time	-11	-7 *	-20 *
	Vehicle actuation	-32 *	-15 *	-23 *
Southampton	Vehicle actuation	-39 *	-1	-48 *
London	Fixed-time	8 % in average		

*significant at the 95 % confidence level

Table A.2 - Changes in Travel Time, Delays, and Stops from the use of SCOOT in Toronto (source: Kelman et al., 1993)

Control area	Number of intersections	Travel time (%)	Delay (%)	Stops (%)
Lake Shore Blvd.	20	-6.8	-25.6	-25.2
Yonge St. North	13	-11.4	-18.7	-31.4
Downtown	42	-5.7	-6.1	-9.7

A.3. UTOPIA - Urban Traffic OPtimization by Integrated Automation

Description based on the following references: Donati et al. (1984), Mauro and Di Taranto (1989), Mauro (1991), Davidsson and Di Taranto (1992), and Wood (1993).

A.3.1. Control Architecture

UTOPIA places strong emphasis on the decentralization of signal timing optimization. It is a two-level hierarchical control system in which signal controllers play a dominant role in the development of signal timing strategies. They determine the signal settings to be implemented at each intersection according to local traffic demand, priority for incoming transit vehicles, and coordination with adjacent controllers. The central computer, which hosts the network control level, adds robustness and stability to the control strategies developed by the signal controllers by imposing timing constraints based on network-wide coordination objectives, the performance function to be used at each intersection, and a network-based reference timing plan that each controller can partly or completely modify.

Adjacent signal controllers coordinate their operation by exchanging pertinent information. For instance, each controller sends its latest optimal timing plan to its upstream neighbors so that each one of them can evaluate the stops and delays incurred by the vehicles leaving their intersection once they reach one of the four downstream intersections. In return, these controllers send to their downstream neighbors forecast of near future traffic demand (vehicles expected to leave their intersection). This communication link is also used by a controller to transmit to its upstream and downstream neighbors actions to take (increase or decrease green time) when demand becomes higher than the capacity or when queue spillback conditions are detected on one of the downstream links.

A.3.2. Traffic Prediction Models

UTOPIA uses two different traffic prediction models. One is used at the intersection level by the signal controllers and the other one is used at the network level by the central computer. The intersection model predicts future arrivals for the next 120 seconds, while the network model makes prediction for the next 30 minutes of signal control. These two models are described in turn in the following paragraphs.

A.3.2.1. Intersection Traffic Prediction Model

At the intersection level, traffic prediction is made by combining information about traffic demand coming from two sources. The first source is the detectors located at the upstream end of each approach to an intersection. These detectors provide 20 to 30 seconds of advance information about vehicle arrivals at the downstream intersection, depending on their location. The neighboring upstream controllers are the second source of information. They provide data that enables a given controller to estimate (with less accuracy than the detectors) the future arrivals for an additional 90- to 100-second period. This additional information enables each controller to obtain advance information about traffic that has not yet been detected by the upstream detectors. It results in a better coordination between adjacent intersections, since controllers are then able to better plan signal timings for future arrivals.

The information provided by the upstream detectors and neighboring controllers is used to produce a stopline flow profile on each approach to an intersection. This profile gives the number of vehicles expected to reach the stopline (when assuming there is no queue at the stopline) during each of the three-second intervals dividing the next 120 seconds of signal operation. Contrary to SCOOT, these profiles truly represent near-future traffic demand, since the information they contain is not smoothed with past data. Contrary to SCATS and SCOOT, UTOPIA is designed to react to near-future traffic demand, and not only to respond to changes detected in past demand.

Traffic progression is simulated every three seconds. Simulation takes the form of a flow profile update using a modified Kalman filtering technique. Queue length estimation is done by considering the predicted arrivals during each three-second interval and the expected status of the traffic signals controlling the flow discharge at the stopline. For simplicity, queues are modeled vertically and vehicles are assumed to travel all along the link before they join a vertical queue. There is no information in the literature regarding the use of a platoon dispersion model.

For robustness in the prediction logic, the estimated queue length is verified, and corrected if necessary, after each phase change. The accuracy of the prediction is estimated by comparing the number of vehicles predicted to have cross the intersection with the number of vehicles that were counted by the detectors located at the upstream end of the three downstream links (detectors located immediately downstream of the intersection).

Local controllers are also in charge of estimating slow-varying parameters such as travel time on each approach, turning percentages, and saturation flow rates. Estimation of these parameters is made through on-line filtering techniques based on the signal status and on the traffic counts made by the detectors located at the upstream of each approach and immediately downstream of the intersection. An update is made every cycle and corrective measures are again based on differences between predicted and measured traffic counts.

A.3.2.2. Network Traffic Prediction Model

The central computer analyzes traffic conditions over the whole network and predicts, with a traffic assignment model, the most important routes that will be taken in the next 30 minutes by the private vehicles. Traffic demand on each one of these routes is defined for each three-minute time interval.

Traffic prediction is made by merging the most recent traffic counts provided by the traffic detectors with filtered statistical traffic characteristics stored in a continuously updated database. A macroscopic network model of major routes is used to predict traffic movements. This network is represented by a collection of oriented one-way storage units, upon which is superimposed a series of fixed routes.

The network traffic prediction model is composed of two parts. The first one operates on a day-to-day basis and makes the optimal updates of the origin-destination attributes and assumed day-by-day similarities of all defined routes. The second runs in real-time. It counts the vehicles detected at each expected origin and makes flow predictions on major routes for the next 30 minutes of signal control based on these counts and on the attributes associated with the specified origin.

A.3.3. Timing Principles

UTOPIA aims at providing good traffic signal coordination, improving traffic conditions for private vehicles, giving priority to transit vehicles on selected routes by forecasting and monitoring the progress of individual transit vehicles, and to permit high flexibility in signal settings at each intersection from a network reference plan.

To achieve these objectives, UTOPIA divides the network control problem into a series of intersection optimization problems and a global network coordination problem. The intersection problems are

solved separately by each controller, while the central computer solves the network coordination problem. The two control levels, as described below, use different timing principles.

A.3.3.1. Intersection Timing Principles

UTOPIA's intersection control is based on acyclic timing principles. Concepts such as cycle length, green splits and offsets have no explicit meaning here. Each signal controller determines the best time to switch from one phase to the other over the next 120 seconds. However, the sequence of phase is fixed.

Real-time control is achieved through a rolling-horizon framework. Signal settings are re-optimized every six seconds based on newly detected traffic conditions. Between each optimization, only the first six seconds of the latest optimal plan is committed to implementation. This ensures that the signal timing strategy in operation is always based on the most recent traffic demand information.

UTOPIA gives to each signal controller a high degree of freedom to adjust the signal timings at each intersection to the specific requirements of the local traffic demand. Each controller also only attempts to coordinate its operation with its immediate neighbors, thereby producing a very loose coordination scheme over the whole controlled network. To compensate for this weak coordination, timing constraints based on network-wide signal control objectives are imposed by the network control level to ensure the existence of a minimum level of coordination.

As explained earlier, detectors located at the upstream end of each approach typically provide 20 to 30 seconds of advance arrival information. Since such a short horizon is not long enough to guarantee smooth signal control, because of the hard constraints of minimum and maximum phase duration (which may exceeds this horizon) and of the expressed goal of providing priority to transit vehicles, a 120-second decision horizon is used. Demand for the additional period is estimated from data provided by adjacent controllers. Then, because there is no correct prediction of vehicle arrivals for the period for which no detector data are directly available, signal settings over the whole decision horizon are optimized following two different procedures:

- In the first part of the decision horizon (typically, the first 20 to 30 seconds), a branch-and-bound algorithm is used to calculate the optimal signal timings. This algorithm takes advantage of the fact that advance arrival information in that part of the decision horizon is

directly provided by the detectors located at the upstream end of each approach. It also attempts to coordinate the timings of the current intersection with those of the surrounding ones by considering the delays incurred at those intersections by vehicles leaving the current intersection. The various cost elements being considered by the branch-and-bound algorithm in its signal optimization process are:

- The delays incurred by private vehicles on the incoming links;
- The number of stops made by private vehicles on the incoming links;
- The excess queueing on the incoming links;
- The delays incurred on the outgoing links by vehicles leaving the intersection;
- The delays incurred at the current intersection by the transit vehicles requesting priority of passage through the intersection;
- The cost of the modifications needed in the timing strategy already defined in the second part of the decision horizon to comply will the signal operation constraints.

- For the second part of the decision horizon (e.g., from 30 to 120 seconds), a sub-optimal heuristic selection procedure is used. Signal settings are optimized by taking into account the probability of not providing priority to incoming transit vehicles, the deviation from the signal settings established six seconds earlier in the previous optimization stage (to ensure smooth behavior over time) and the deviation from the fixed reference plan generated by the central computer based on network coordination criteria (to ensure good coordination and to find a stable and robust solution). No direct coordination with adjacent controllers seems to be attempted in this second part of the decision horizon.

The overall optimization technique consists of the following steps.

1. Adaptation of the intersection signal control strategy over the whole 120-second decision horizon to the reference timing plan provided by the network control level.
2. Adaptation of the control strategy to transit vehicle requirements. The control strategy obtained in step 1 is modified to ensure that transit vehicles expected to reach the intersection within the decision horizon obtain a green indication.
3. Analytical optimization of the signal timings in the first part of the decision horizon (first 20 to 30 seconds) by using the branch-and-bound algorithm mentioned above.

A.3.3.2. Network Timing Principles

The overall network optimization objective is to minimize the total travel time of private vehicles, subject to any necessary delay to give priority to transit vehicles on selected routes at signalized intersections.

UTOPIA achieves this objective by suggesting average travel speeds and saturation flows for all links in the network that will minimize travel times within the controlled area and by setting signal timings accordingly. During this optimization process, suitable constraints reflecting the actual ones are imposed on the average speeds and saturation flow rates. The timings that result from this optimization constitute the network reference plan that each controller receives from the central computer, together with the specific timing constraints and performance function they should use to optimize the signals they control. These elements, considered all together, may contribute to the establishment of preferred routes in the network and may impose traffic-dependent criteria on the intersection signal control level.

Real-time signal control is achieved at the network level through another rolling-horizon framework. The reference timing plan, the timings constraints and the performance function each controller should use are optimized every 30 minutes for each 3-minute interval dividing the decision horizon. However, a new optimization is performed only when a significant change in traffic conditions is detected. A reference plan may thus be in use for a full 30 minutes.

A.3.4. Priority to Transit Vehicles

In UTOPIA, transit priority is considered at the intersection level. Priority is granted on a simple call basis by local controllers, with no direct consideration for the disruptions to that may be caused to other traffic. As a general rule, incoming transit vehicles get an absolute priority of passage, except when there are conflicts with other transit vehicles at the same intersection.

In Turin, UTOPIA's only implementation site, transit vehicles for which priority is providied do not share the right-of-way with other traffic. Tramways run on tracks that are mainly separated from other traffic, while buses are traveling on either reserved or protected lanes. In the first installation of the system, tramways and buses were selectively detected from other traffic by using special traffic detectors installed at the upstream and downstream ends of every transit stops, and at the entrance and

exit of signalized intersections for which priority must be provided. These detectors allowed the system to know when a transit vehicle started end ended dwelling, and when a vehicle would have entered and intersection and would no longer require special priority treatment. Each time a transit vehicle would be detected, a message was sent to the central computer asking it to predict the arrival times of the vehicle at successive downstream intersections for the next eight minutes. Once the predictions were completed, the projected arrival times would be sent to every interested signal controller for use in their signal optimization process. These times, and the resulting timings, would then be updated each time the vehicle would pass over another detector. Improvements to this detection scheme were later made when the City of Turin put its transit vehicle location system into operation. Instead of using a series of special traffic detectors, UTOPIA now relies on the prediction capabilities of the city's transit vehicle location system to obtain estimates of vehicles' arrival times at successive intersections.

A.3.5. Implementation Results

An evaluation of UTOPIA's effectiveness made in 1986 showed that the implementation of this model on a network of about 40 intersections resulted in a significant increase in average travel speeds (Mauro and Di Taranto, 1989). For instance, it was found that the average travel speed of private vehicles increased globally by 15.9 percent. Specific increases were between 11.4 and 20.9 percent depending on the test routes. A deeper analysis of the system performances showed that UTOPIA was able to gain more in heavy traffic conditions. During peak hours, measured travel speeds increased for instance by about 35 percent. Tramway's commercial speed also increased significantly as a result of UTOPIA's implementation. The global increase was 19.9 percent. It was also found that tramways obtained absolute priority at almost all the intersections. Tramways had to stop more often at only two intersections having a very critical design.

A.4. PRODYN - "PROgrammation DYNamique" (DYNamic PROgramming)

Description based on the following references: Henry et al. (1983), Henry and Farges (1984), Barrière et al. (1986), Farges and Henry (1988), Kessaci (1988), Henry and Farge (1989), Kessaci et al. (1989), and Wood (1993).

A.4.1. Control Architecture

PRODYN adopts a fully distributed approach to control a network of signalized intersections. The signal controllers do all timing calculations, whereas the role of the central computer is limited to monitor signal operation and perform general data management tasks.

Coordination of signal operation between adjacent intersections is achieved by enabling controllers to communicate with their immediate neighbors. However, contrary to UTOPIA, communication is done in a single direction only. Information always travels from an upstream controller to a downstream one. Each controller sends to its downstream neighbors a forecast of the expected departures from the intersection it controls under its current optimum timing plan. The downstream intersections will convert these forecasted departures in stopline arrivals and consider them in their signal timing optimization. There is no information exchanged, and thus no coordination, between intersections more than 200 metres away.

A.4.2. Traffic Prediction Model

PRODYN predicts vehicle arrivals for a period extending 75 seconds into the future. Traffic demand in that period is modeled by a flow profile giving the number of expected arrivals at the stopline in each of the 16 five-second intervals contained in the decision horizon. These profiles represent near-future traffic demand. They are built upon traffic counts that are not smoothed with past data, except when detectors and adjacent controllers cannot provide advance arrival information for the full length of the 75-second decision horizon.

Traffic prediction is done with the help of two detectors installed in every traffic lane: one near the upstream intersection and one near the stopline. The upstream detector provides every five seconds the

number of vehicles that passed over it. The second detector indicates every second if a vehicle had passed over it or not. The information gathered by this detector is used to adjust the forecast of vehicle arrivals near the stopline, to check if the measured queue is consistent with the predicted one, and to calculate the parameters of the Normal probability distributions associated with the queue predictor (see below).

As explained above, a controller receives additional near-future arrival information from its upstream neighbors if they are located less than 200 metres away. At the beginning of each optimization step, each controller simulates the arrival and departure processes at the intersection under its control according to its current optimum timing plan, turning percentages calculated off-line, and fixed saturation flow rates. The expected departures are then sent to the corresponding downstream controllers, which will convert them into arrivals at their own intersections. This additional information enables each PRODYN controller to obtain advance information about traffic that has not yet been detected by the upstream detectors. It results in a better coordination between adjacent intersections, since controllers are able to better plan signal timings for future arrivals.

Where there is no communication between adjacent controllers, a moving average estimation technique is used to obtain a forecast of vehicle arrivals over the part of the 75-second decision horizon for which installed detectors can provide no advance arrival information. Additional detectors can be installed at more than 200 metres of the stopline, but they may not provide advance arrival information for the full length of the decision horizon if the travel time between them and the stopline is less than 75 seconds. Kessaci (1988) indicates that the optimal size of the window for the moving average is about 20 observations.

Queue prediction at the stopline is done using a Bayesian state filtering procedure. PRODYN recognizes that traffic behavior is stochastic in nature and that a precise estimate of the length of the queue can hardly be obtained by simply adding and subtracting vehicles arriving at the stopline and departing from it. PRODYN's developers felt that a stochastic prediction model was more appropriate than a deterministic queue predictor. The procedure used makes use of a number of probability distributions to estimate the size of the queue. It defines three possible classes of queue, small, average and large, and uses normal probability distributions to define the possible length of a queue belonging to each category. The parameters defining the probability distributions are refreshed every five seconds by comparing the predicted and measured arrivals at the detectors located near the

stopline on each approach. Bayes rule is then applied to define a general probability distribution based on an estimate of the current size of the queue. From this general probability distribution and the expected number of vehicle arrivals within each five-second interval, another probability distribution is obtained. This distribution describes the possible length of the queue at the specified time into the future (5 seconds into the future, 10 seconds, etc.) The average predicted queue length used in the signal optimization procedure is the mean value of this last distribution.

Among other important characteristics of the traffic prediction model are the use of a vertical queue modeling and the assumption that all vehicles travel at a constant speed (i.e., no platoon dispersion) between signalized intersections. A model with a more realistic horizontal queue model was first envisioned (Farges and Henry, 1988), but the complexity of the resulting dynamic programming optimization problem dictated the use of a simpler model. Memory and computation problems associated with the use of Forward Dynamic Programming techniques to control a network of signalized intersections at the time the model was initially developed led to many simplifications in the signal control problem formulation. Using Forward, and not Backward Dynamic Programming complicates the optimization problem because non-linear state equations must be inverted, which is not always easy or possible. The use of a vertical queuing model is however consistent with most existing traffic simulation models.

A.4.3. Timing Principles

PRODYN, like UTOPIA, is an acyclic real-time signal control model. Signal timings are optimized on a phase by phase basis without any common cycle length requirement between adjacent intersections. Concepts such as cycle length, green splits and offsets are not explicitly used. This use of acyclic timing principles enables PRODYN to operate each signal in a network at its individually optimum timings. The weak network coordination scheme (each controller only coordinates its operation with its immediate neighbors) may however reduce the global optimality of the network timing strategy, especially since there is no network control level.

PRODYN determines the best phase sequence and switching times (with a five-second resolution) for the next 75-seconds of signal control by using Forward Dynamic Programming. Real-time control is achieved through a rolling-horizon framework similar to the one used by UTOPIA's intersection

control. Signals are re-optimized every five-seconds, and only the first five seconds of the new optimum timing plan is committed to implementation at the end of each optimization stage.

As explained above, each controller attempts to coordinate its operation with its immediate upstream neighbors if they are less than 200 metres away. No area-wide coordination is attempted. Simulation tests reported by Barrière et al. (1986) and discussed in section A.4.5 below have indicated that a distributed control approach in which controllers only attempt to coordinate their operation with their neighbors can produce, with a lot less calculations, signal timings almost as good as a hierarchical control structure with an upper control level in charge of coordinating all controllers. A fully distributed approach with no coordination attempted (signal control as if all intersections were isolated) produced the worst results.

The optimization criterion followed by each PRODYN's controllers is the minimization of the sum of delays incurred by vehicles on all approaches to the intersection under its control over the 75-second projection horizon. There is no network-wide optimization criterion. Delays incurred by vehicles at adjacent intersections are not included in the performance evaluation of a given intersection. However, a terminal cost function is included to penalize timing strategies leaving long residual queues on some approaches at the end of the decision horizon.

A.4.4. Priority to Transit Vehicles

Transit vehicles are not modeled. Consequently, there is no priority offered to these vehicles at signalized intersections. Wood (1993), however, states that such facilities are planned.

A.4.5. Test and Implementation Results

The PRODYN model was tested by simulation with various control architectures (Barrière et al., 1988): as a hierarchical system with an upper control level coordinating the operation of local controllers, as a fully distributed system in which no coordination is attempted, and as distributed system with coordination attempted between adjacent intersections (current model). It is the only model for which such extensive testing is reported in the literature. Very important conclusions can thus be drawn from this experience.

As indicated in Table A.2, the distributed approach with attempted coordination compared favorably with the hierarchical model despite the lack of network wide coordination objectives. The distributed approach with no coordination produced the overall worst results, thus demonstrating the need of coordinating signalized intersections in a network. From these results, it was decided to pursue PRODYN's development with the distributed approach featuring coordination between adjacent controllers. Until 1986, PRODYN was mainly developed as a two-level hierarchical control system (Henry et al., 1983, 1984). Local controllers were in charge of predicting traffic demand and computing optimum local signal timing strategies over the decision horizon using Forward Dynamic Programming. The upper control level was attempting to find the best coordination strategy between all controllers in the network. The optimum timing strategies was obtained by iterating between the two control levels.

Table A.3 - Efficiency of various PRODYN control architectures

Control Model	Delay reduction when compared to TRANSYT optimized plans	
	4-intersection arterial	4-intersection grid network
Hierarchical model	34.4 %	13.5 %
Distributed model with no coordination	-5.0 %	12.0 %
Distributed model with coordination with adjacent signals	30.7 %	11.9 %

The distributed control approach was preferred although it did not produce the best overall timings because of the major limitations associated with the hierarchical model. These limitations resulted from the necessity to make iterations to find an optimum signal control strategy for the whole network and from the increasing complexity of the control problem each time new intersections were added. These elements created the need for high capacity communication links between local controllers and the central computer, which hosted the upper control level. The use of a powerful central computer in regards of computation speed and memory was also required in order to be able to compute within a five-second time limit an optimum timing plan. With the 8086-8MHz microprocessor technology available then to implement to control logic in the signal controllers, it was estimated that a network of no more than 10 intersections could be controlled in real-time with the PRODYN model. The distributed approach had the advantage of not limiting the size of the controlled network, since each

controller only cares about its immediate neighbors. Moreover, adding new intersections in the controlled area does not create the need to modify the control algorithm.

Later tests made on a real seven-intersection network (Henry and Farges, 1989) with the distributed model showed that PRODYN was able to reduce delays in the test network of Toulouse, France, by an average of 10 percent with a 99.99 percent statistical significance when compared to a fixed-time control plan optimized by TRANSYT. Measured gains varied between 6 and 12 percent according to the time of the day.

Kessaci (1988) and Kessaci et al. (1989) also report that improvements could be made to the PRODYN algorithm when queue spillback condition occurs by introducing an upper control level in the model formulation, and thus, by going back to a hierarchical model. They indicate that queue spillback could theoretically be treated in the current PRODYN distributed modeling by introducing a control logic that would take into account the output limitation of an intersection due to the spillback of downstream queues across the intersection. However, this solution was not seen as practically possible. At the time they analyzed the congestion problem, the introduction of such model would have increased drastically the computation time required to perform the signal optimization with PRODYN's forward dynamic programming problem formulation. The increased communication needs between adjacent controllers would also have increased significantly the costs of the physical control system. They thus proposed a second approach, which consisted in the development of a network control level constraining the operation of PRODYN's local controllers when queue spillbacks are detected. This upper control level would have determined the minimum and maximum duration of phases that each controller could implement in such way that congestion would be shared among controlled links proportionally to their storage capacity. These constraints were to be determined using an off-line optimization procedure dealing with more average traffic conditions than local controllers (sampling time of several minutes).

This work, however, does not indicate that PRODYN is an inadequate model, but simply that PRODYN's ability to efficiently model and control a signalized network is highly dependent on computer technology. The better the computers are, the better PRODYN can be. There is notably no indication in the recent literature that the control hierarchy mentioned in the previous paragraph was finally introduced in PRODYN. The availability of faster computers and of better communication technologies may have solved the problem described above.

APPENDIX B

Isolated Intersection Test Scenarios

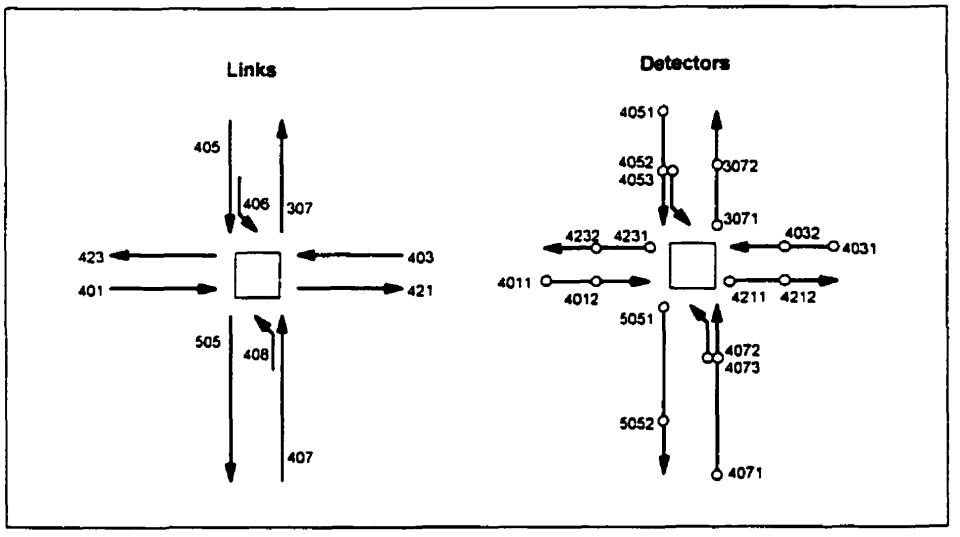


Figure B.1 - Simulation Model of Isolated Intersection Test Configuration

Table B.1 - Isolated Intersection Low Volume Scenario

Link	Traffic volume (veh/hr)	Transit volume (veh/hr)	Saturation flow		Private automobile turning volumes		
			Through (pcu/hr)	Left (pcu/hr)	Left (veh/hr)	Through (veh/hr)	Right (veh/hr)
407	1013	6	3400	1385	51	914	48
405	575	6	3400	1385	29	519	27
401	501	0	2706	---	0	464	37
403	581	0	2872	---	0	454	127

Table B.2 - Isolated Intersection Medium Volume Scenario

Link	Traffic volume (veh/hr)	Transit volume (veh/hr)	Saturation flow		Private automobile turning volumes		
			Through (pcu/hr)	Left (pcu/hr)	Left (veh/hr)	Through (veh/hr)	Right (veh/hr)
407	1350	6	3400	1385	68	1218	64
405	767	6	3400	1385	38	692	36
401	668	0	2706	---	0	619	49
403	774	0	2872	---	0	605	169

Table B.3 - Isolated Intersection High Volume Scenario

Link	Traffic volume (veh/hr)	Transit Volume (veh/hr)	Saturation flow		Private automobile turning volumes		
			Through (pcu/hr)	Left (pcu/hr)	Left (veh/hr)	Through (veh/hr)	Right (veh/hr)
407	1688	6	3400	1385	84	1523	80
405	959	6	3400	1385	48	865	46
401	835	0	2706	---	0	774	61
403	968	0	2872	---	0	757	211

Table B.4 – Isolated Intersection Transit Dwell Times

Approaches							
Northbound		Southbound		Eastbound		Westbound	
Link	Average Dwell (sec)	Link	Average Dwell (sec)	Link	Average Dwell (sec)	Link	Average Dwell (sec)
407	16	405	14	No transit service		No transit service	
307	16	505	13				

Table B.5 – Isolated Intersection Phasing Parameters

Phase	Link served	Minimum Green (sec)	Maximum Green (sec)	Amber Period (sec)	Start Lag (sec)	Stop Lag (sec)
Low / Medium Demand Levels						
1 (Left-turn arrow)	406, 408	5	10	2	3	2
2 (Arterial green)	405, 406, 407, 408	10	60	5	3	2
3 (Cross-street green)	401, 403	10	60	5	3	2
High Demand Level						
1 (Left-turn arrow)	406, 408	5	10	2	3	2
2 (Arterial green)	405, 406, 407, 408	10	90	5	3	2
3 (Cross-street green)	401, 403	10	60	5	3	2

APPENDIX C

Arterial Test Scenarios

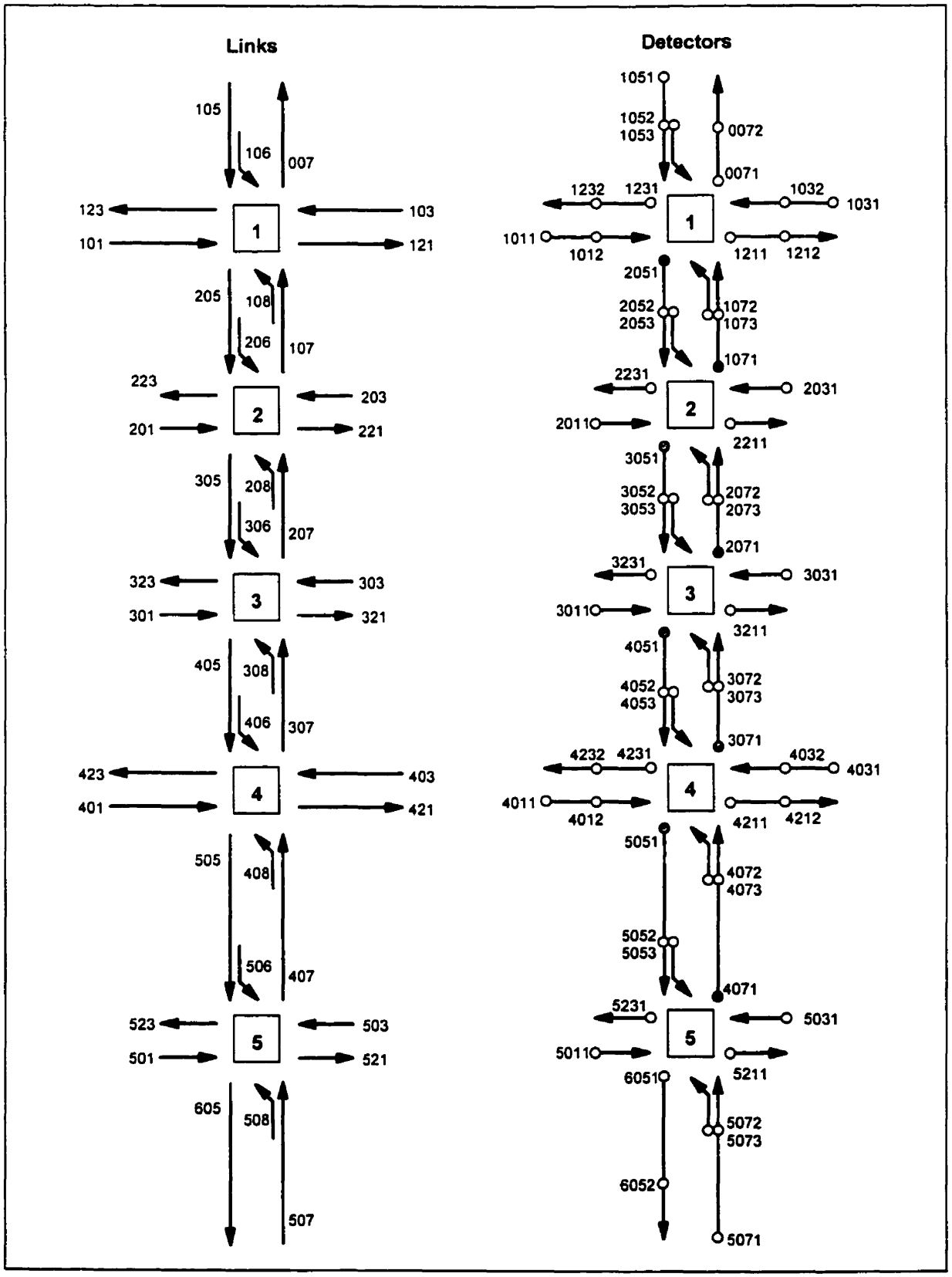


Figure C.1 - Simulation Model of Arterial Test Configuration

Table C.1 - Arterial Low Volume Scenario

Link	Traffic	Transit	Saturation Flow		Traffic Turning Volumes		
	Volume	Volume	Through	Left	Left	Through	Right
Northbound Arterial							
507	1013	10	3400	1385	51	914	48
407	961	10	3400	1385	48	867	46
307	994	10	3400	1385	50	897	47
207	983	10	3400	1385	49	887	47
107	981	10	3400	1385	49	885	47
007	885	10	3400	---	---	799	---
Southbound Arterial							
105	575	10	3400	1385	29	519	27
205	519	10	3400	1385	26	468	25
305	555	10	3400	1385	28	501	26
405	565	10	3400	1385	28	510	27
505	547	10	3400	1385	27	493	26
605	493	10	3400	---	---	493	---
Eastbound Cross-Street Approaches							
101	450	0	2820	---	0	450	0
201	225	0	2218	---	14	158	54
301	374	0	2201	---	29	314	31
401	501	0	2706	---	0	464	37
501	342	0	2529	---	16	326	0
Westbound Cross-Street Approaches							
103	450	0	2774	---	0	450	0
203	254	0	2238	---	33	141	80
303	284	0	2276	---	33	195	56
403	581	0	2872	---	0	454	127
503	342	0	2406	---	0	312	30

Table C.2 - Arterial Medium Volume Scenario

Link	Traffic	Transit	Saturation Flow		Traffic Turning Volumes		
	Volume	Volume	Through	Left	Left	Through	Right
Northbound Arterial							
507	1350	10	3400	1385	68	1218	64
407	1280	10	3400	1385	64	1156	61
307	1324	10	3400	1385	66	1195	63
207	1309	10	3400	1385	65	1182	62
107	1307	10	3400	1385	65	1179	62
007	1179	10	3400	---	---	1064	---
Southbound Arterial							
105	767	10	3400	1385	38	692	36
205	692	10	3400	1385	35	625	33
305	741	10	3400	1385	37	669	35
405	753	10	3400	1385	38	680	36
505	729	10	3400	1385	36	658	35
605	658	10	3400	---	---	658	---
Eastbound Cross-Street Approaches							
101	600	0	2820	---	0	600	0
201	338	0	2218	---	18	210	72
301	378	0	2201	---	39	418	41
401	774	0	2706	---	0	619	49
501	456	0	2529	---	22	434	0
Westbound Cross-Street Approaches							
103	600	0	2774	---	0	600	0
203	338	0	2238	---	44	187	107
303	378	0	2276	---	44	259	75
403	774	0	2872	---	0	605	169
503	456	0	2406	---	0	416	40

Table C.3 - Arterial High Volume Scenario

Link	Traffic	Transit	Saturation Flow		Traffic Turning Volumes		
	Volume	Volume	Through	Left	Left	Through	Right
Northbound Arterial							
507	1688	10	3400	1385	84	1523	80
407	1601	10	3400	1385	80	1445	76
307	1656	10	3400	1385	83	1495	79
207	1637	10	3400	1385	82	1478	78
107	1634	10	3400	1385	82	1475	78
007	1475	10	3400	---	---	1475	---
Southbound Arterial							
105	959	10	3400	1385	48	865	46
205	865	10	3400	1385	41	781	43
305	926	10	3400	1385	44	836	44
405	942	10	3400	1385	45	850	47
505	912	10	3400	1385	43	823	46
605	823	10	3400	---	---	823	---
Eastbound Cross-Street Approaches							
101	750	0	2820	---	0	750	0
201	375	0	2218	---	23	263	90
301	623	0	2201	---	49	523	51
401	835	0	2706	---	0	774	61
501	570	0	2529	---	27	543	0
Westbound Cross-Street Approaches							
103	750	0	2774	---	0	750	0
203	423	0	2238	---	55	234	134
303	473	0	2276	---	55	324	94
403	968	0	2872	---	0	757	211
503	570	0	2406	---	0	520	50

Table C.4 – Arterial Transit Dwell Times

Approaches							
Northbound		Southbound		Eastbound		Westbound	
Link	Average Dwell (sec)	Link	Average Dwell (sec)	Link	Average Dwell (sec)	Link	Average Dwell (sec)
507	17	105	36	No transit service		No transit service	
407	16	205	12				
307	16	305	16				
207	19	405	14				
107	33	505	13				
007	27	605	13				

Table C.5 – Arterial Phasing Parameters

Phase	Link served	Minimum Green (sec)	Maximum Green (sec)	Amber Period (sec)	Start Lag (sec)	Stop Lag (sec)
Low / Medium Demand Levels						
1 (Left-turn arrow)	x06, x08	5	10	2	3	2
2 (Arterial green)	x05, x06, x07, x08	10	60	5	3	2
3 (Cross-street green)	x01, x03	10	60	5	3	2
High Demand Level						
1 (Left-turn arrow)	x06, x08	5	10	2	3	2
2 (Arterial green)	x05, x06, x07, x08	10	90	5	3	2
3 (Cross-street green)	x01, x03	10	60	5	3	2

APPENDIX D

Webster-Cobbe Signal Optimization for Isolated Intersections Operated in Fixed-Time

The Webster-Cobbe delay minimization principle states that the delay incurred by drivers at a signalized intersection is minimized when the available green time is allocated to each phase in proportion to the demand for each phase.

In this optimization principle, the optimal cycle is given by Equation D.1:

$$C_{opt} = \frac{1.5L + 5}{1 - Y} \qquad [D.1]$$

where L = Total lost time over the cycle (seconds).

Y = Load of the intersection.

The total lost time, L, represents the time during which the intersection is not effectively used by any movement. During a typical signal operation, lost time first occur at the beginning of a green phase as a result of the need for the drivers to react to the initiation of the phase and to accelerate to ambient speed. Lost time also occur at the end of a green phase, during the last few seconds of the yellow interval, when driver decide to stop instead of crossing the intersection. The lost time also includes any all-red period during which the signal displays are red on every approach, as no vehicle effectively enters the intersection during such period.

To calculate the load Y of a given intersection, the maximum volume to saturation flow ratio, or load, of each phase is first determined using Equation D.2. Once all the maximum ratios have been determined, the load Y is obtained by adding together the maximum ratios of all the phases in the signal cycle according to Equation D.3.

$$y_i = max\left(\frac{v_{m,i}}{s_{m,i}}\right) \qquad [D.3]$$

$$Y = \sum_{i=1}^{n} y_i \qquad [D.2]$$

where: Y = Intersection load.

y_i = Load of phase i.

n = Number of phases in the signal cycle.

$v_{m,i}$ = Traffic volume in movement m during phase i (veh/h).

$s_{m,i}$ = Saturation flow of movement m during phase i (veh/h).

In Equation D.3, the volume represents the number of vehicles observed to cross the intersection, while the saturation flow represents the maximum rate at which vehicles could traverse the intersection assuming that the green signal is available at all time and that there is no lost time at the beginning or end of every phase. For each phase, the maximum volume to saturation flow ratio is determined by calculating the ratio for each particular traffic movement and retaining the highest one. Typically, the load Y of a given intersection should not exceed 0.8, as a higher ratio indicates that the intersection is operating near capacity. The capacity of an intersection, or maximum flow that can be process in a given amount of time is attained when the load of the intersection reaches 1.0. Congestion occurs when the load exceeds 1.0, as the intersection there is then not enough capacity to serve all vehicles arriving in a given amount of time.

To apportion the cycle length between the various phases, the maximum load factor for each phase is used. Typically, the available usable green time, that is, the total cycle time minus the total lost time, is divided among all phases in proportion to their respective maximum volume to saturation flow ratio according to Equation D.4:

$$g_{e,i} = \frac{y_i (C - L)}{Y} \qquad [D.4]$$

where $g_{e,i}$ = Effective green duration of phase i (seconds).

The effective green represents the portion of green time that is effectively used by vehicles. This interval corresponds to the actual green and yellow intervals, less any time loss at the beginning and end of the phase. Equation D.5 mathematically formulates this calculation:

$$G_{e,i} = g_{e,i} + t_{l,i} - A_i \qquad [D.5]$$

where: $G_{e,i}$ = Actual green duration of phase i (seconds).

$t_{l,i}$ = Lost time during phase i (seconds).

A_i = Amber time at end of phase i (seconds).

Tables D.1 to D.4 presents the results of a sample calculation of Webster's optimal timings using Equations D.1 to D.5. The calculations shown are for the medium-demand isolated intersection scenario of Table 9.1 with no transit vehicles and constant arrival patterns.

Table D.1 –Timing Parameters

Phase	Description	Links served	Fixed Intergreen (sec)	Minimum green (sec)	Start-up lost time (sec)	Clearance lost time (sec)	Total lost time $t_{l,i}$ (sec)
1	Main street flashing green	406 408	2.0	5.0	2.0	0.0	2.0
2	Main-street green	405, 406 407, 408	5.0	10.0	2.0	2.0	4.0
3	Cross-street green	401 403	5.0	10.0	2.0	2.0	4.0
						L =	10.0

Table D.2 – Calculation of Volume to Saturation Ratios

Link	Volume (veh/sec)	Saturation flow (veh/sec)	v/s ratio
401	668	2706	0.247
403	774	2872	0.269
405	729	3400	0.214
407	1282	3400	0.377
406	38	1385	0.027
408	68	1385	0.049

Table D.3 – Determination of Critical Volume to Saturation Ratios

Phase	Description	Links served	Critical *v/s* ratio, y_i
1	Main street flashing green	406 408	0.049
2	Main-street green	405, 406 407, 408	0.377
3	Cross-street green	401 403	0.269
		Y =	0.695
		C =	65.6 sec

Table D.4 – Determination of Webster-Cobbe's Optimal Signal Timings

Phase	Description	Links served	Critical *v/s* ratio, y_i	Green ratio y_i / Y	Effective green, $g_{e,i}$ (sec)	Actual green, $G_{e,i}$ (sec)
1	Main street flashing green	406 408	0.049	0.071	4.0	4.0
2	Main-street green	405, 406 407, 408	0.377	0.542	30.1	29.1
3	Cross-street green	401 403	0.269	0.387	21.5	20.5
		Y =	0.695			

Table D.5 – Adjustment for Minimum Green Durations

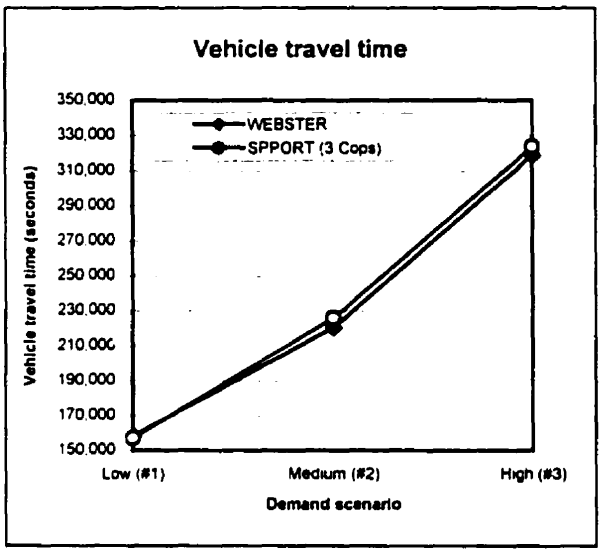
Phase	Description	Links served	Actual green, $G_{e,i}$ (sec)	Minimum green (sec)	Critical v/s ratio, y_i	Green ratio, y_i / Y	Effective geen, $g_{e,i}$ (sec)	Actual green, $G_{e,i}$ (sec)
1	Main street flashing green	406 408	4.0	5.0 (1)	--	---	---	5.0
2	Main-street green	405, 406 407, 408	29.1	10.0	0.377	0.584	29.6	28.6
3	Cross-street green	401 403	20.5	10.0	0.269	0.416	21.0	20.0
					Y =	0.646		

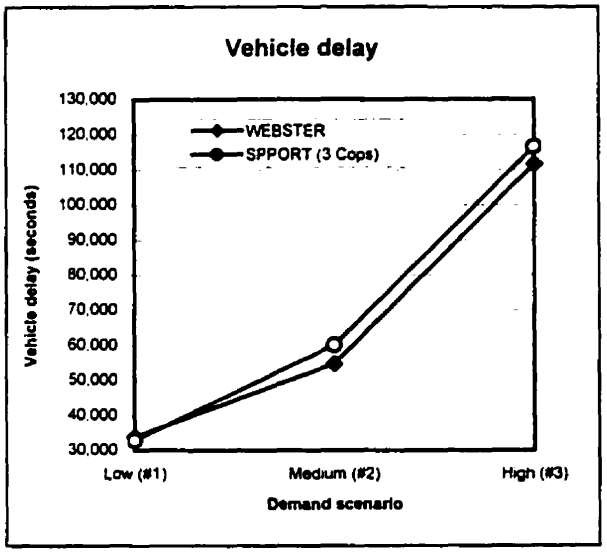
(1) Since calculated optimal green time is less than the required minimum, new effective green times have to be calculated for the main-street and cross-street phases. These times are calculated assuming that the 5 second assigned to the flashing green is now part of the lost time and cannot be assigned to the other phases (that is, C-L = 65.6-15 instead of C-L = 65.6-10 in equation D.4).

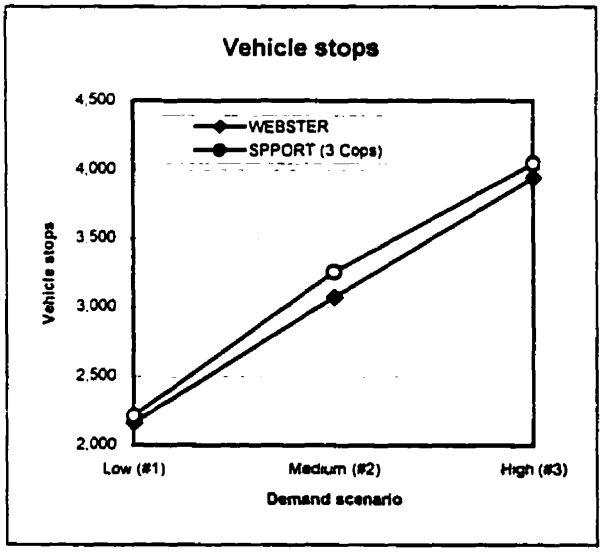
APPENDIX E

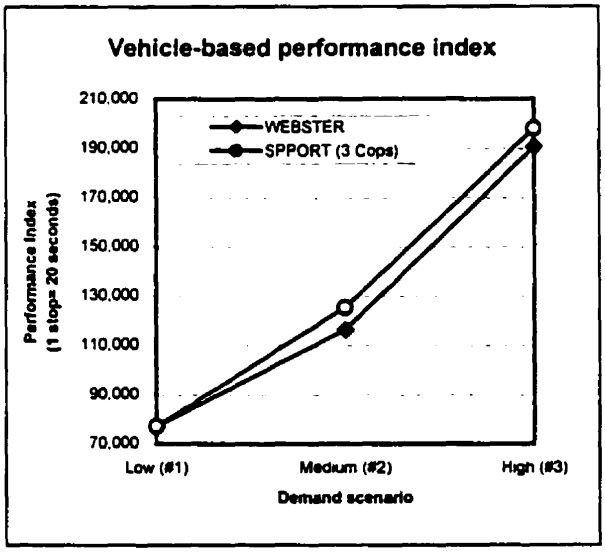
Comparison Results Between SPPORT and Fixed-Time Control for Isolated Intersection Scenarios

Isolated intersection scenarios #1, #2, #3
(Constant demand without transit vehicles)

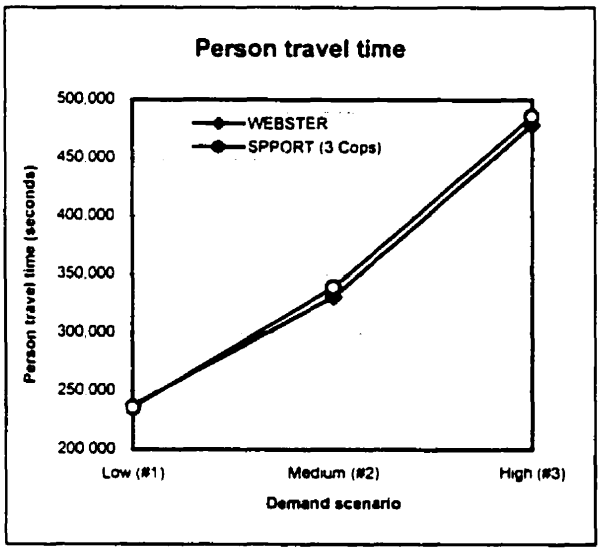


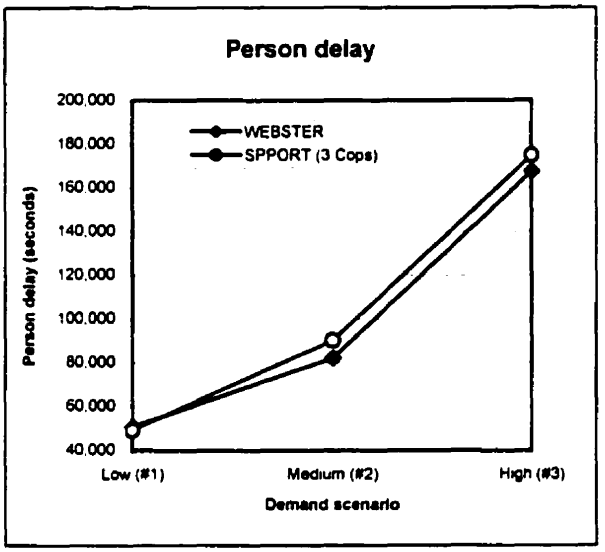


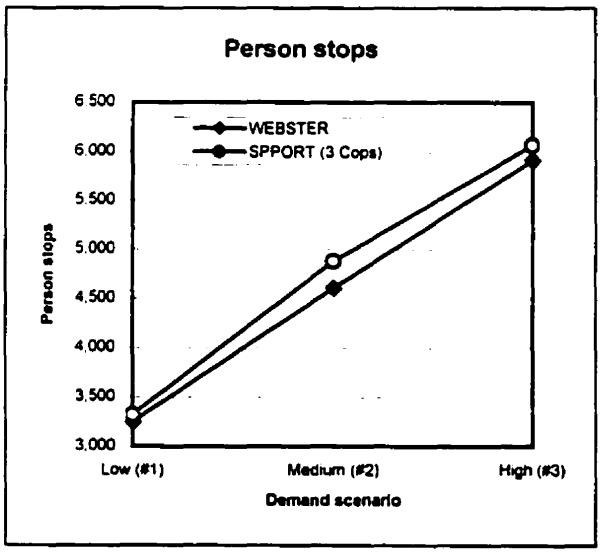


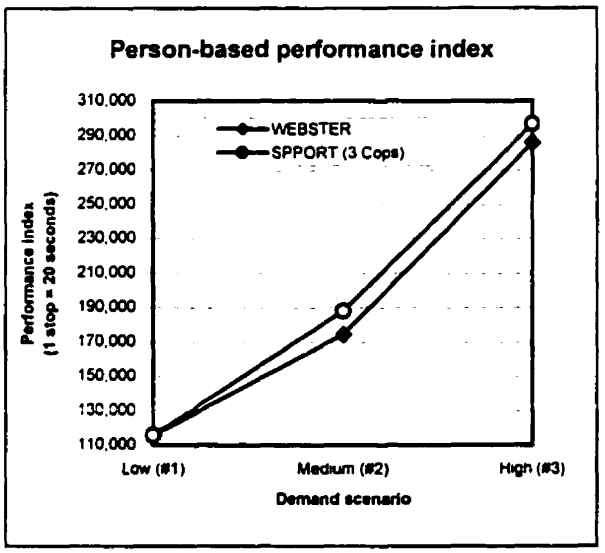


Isolated intersection scenarios #1, #2, #3
(Constant demand without transit vehicles)

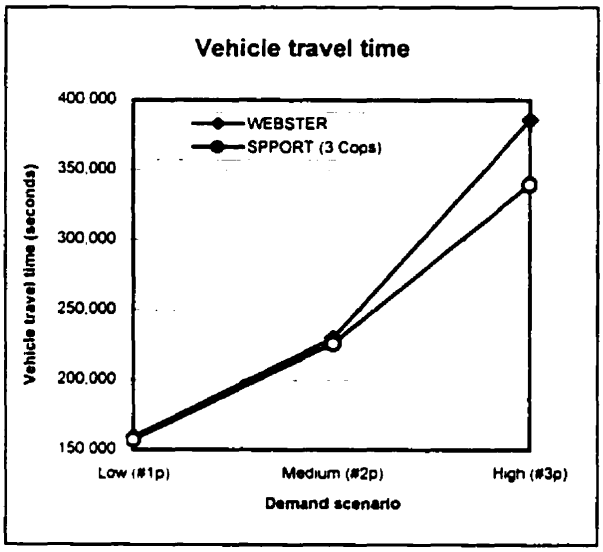


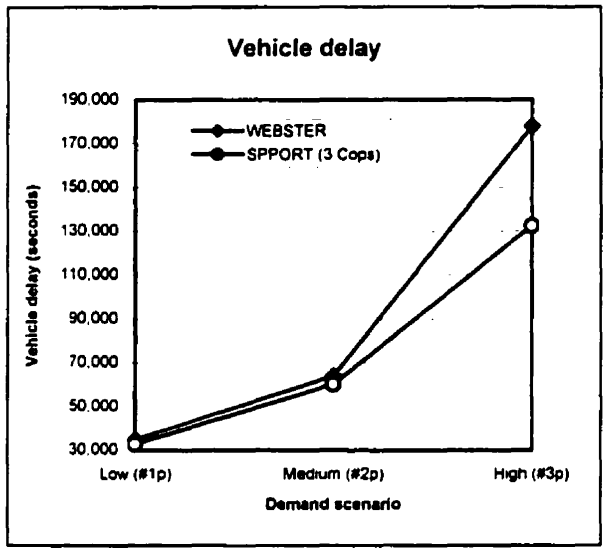


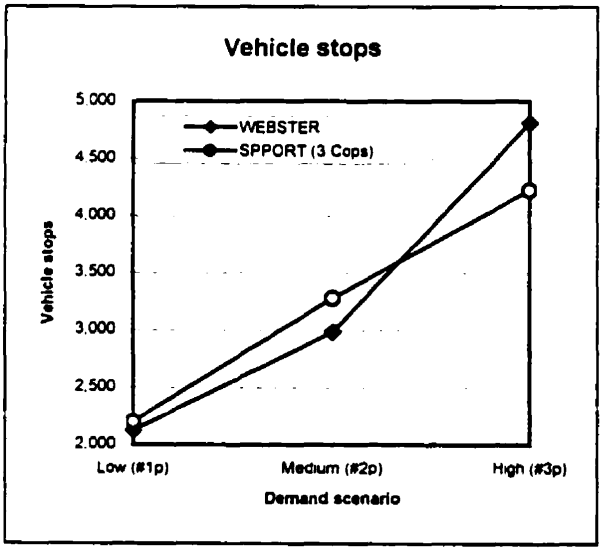


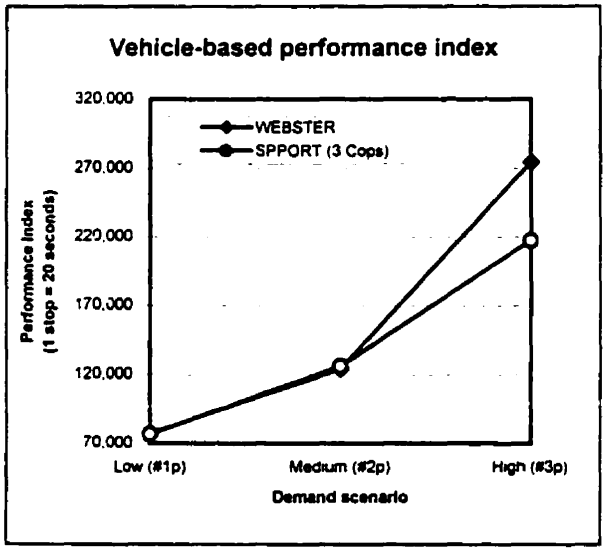


Isolated intersection scenarios #1p, #2p, #3p
(Constant demand without transit vehicles)

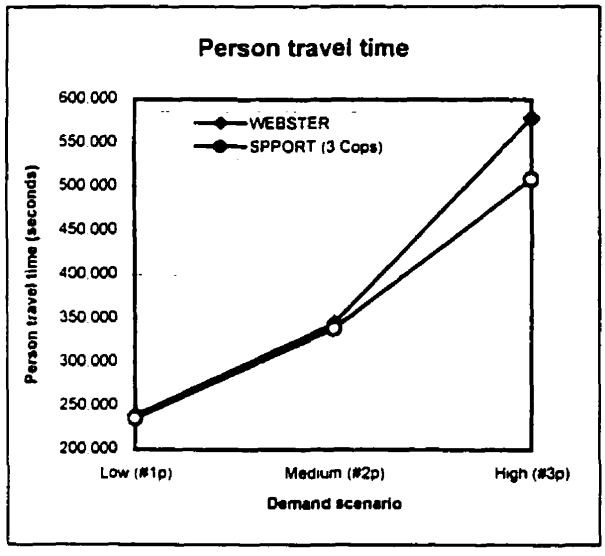


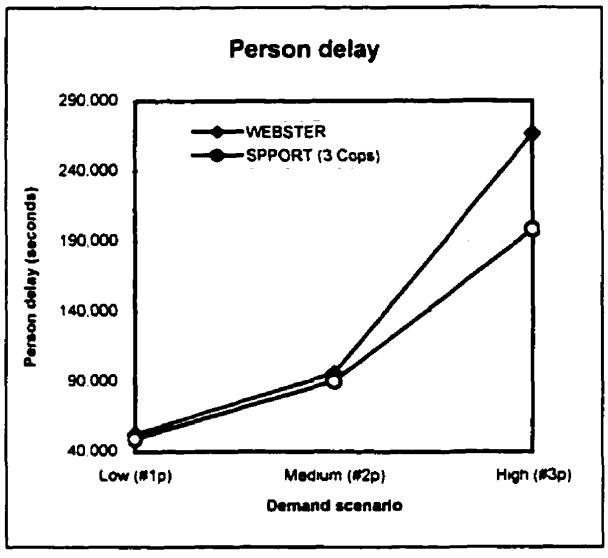


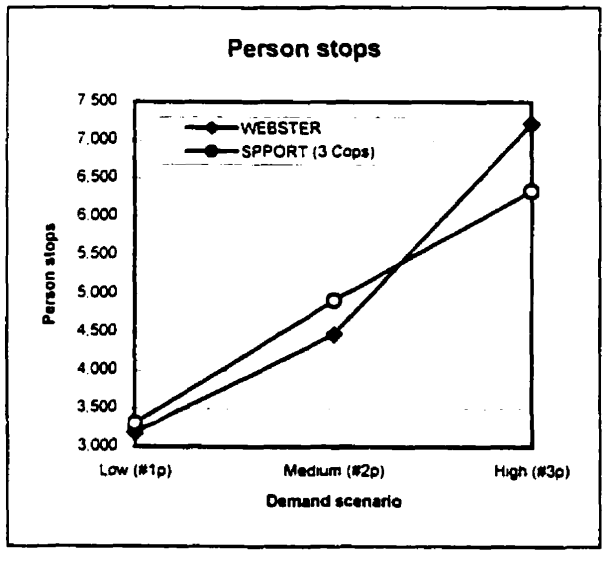


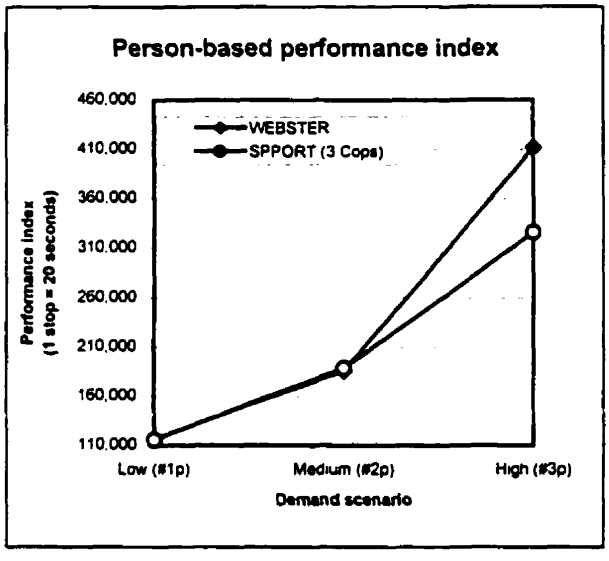


Isolated intersection scenarios #1p, #2p, #3p
(Constant demand without transit vehicles)

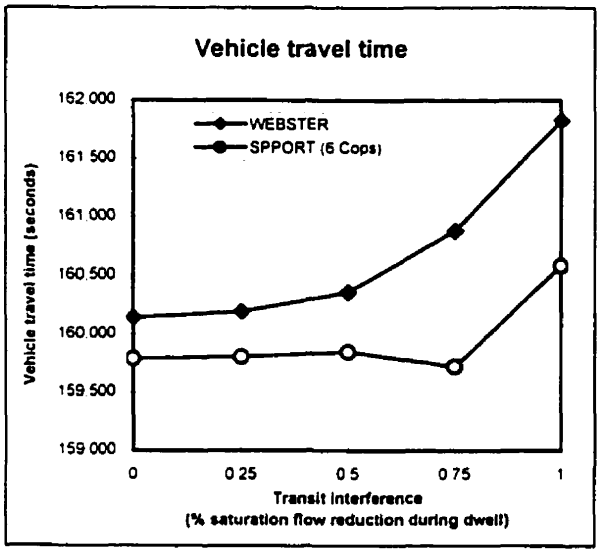


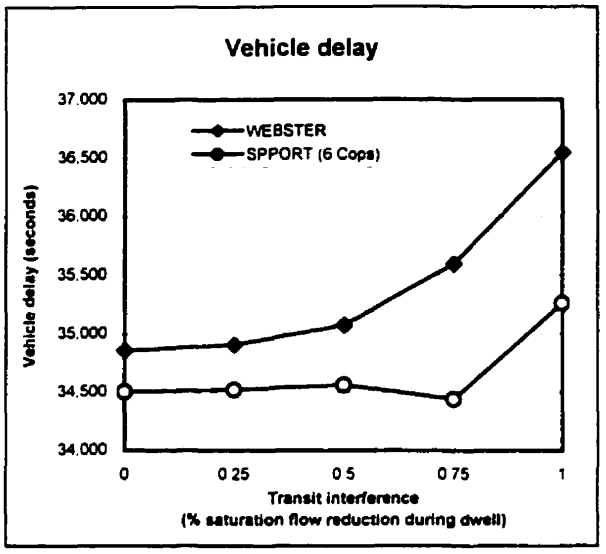


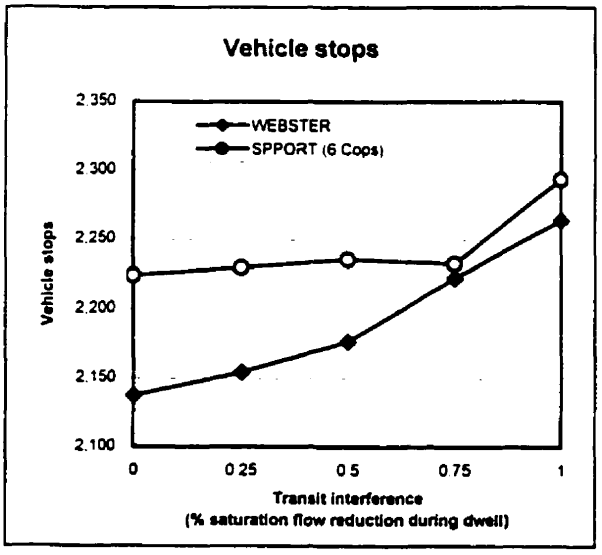


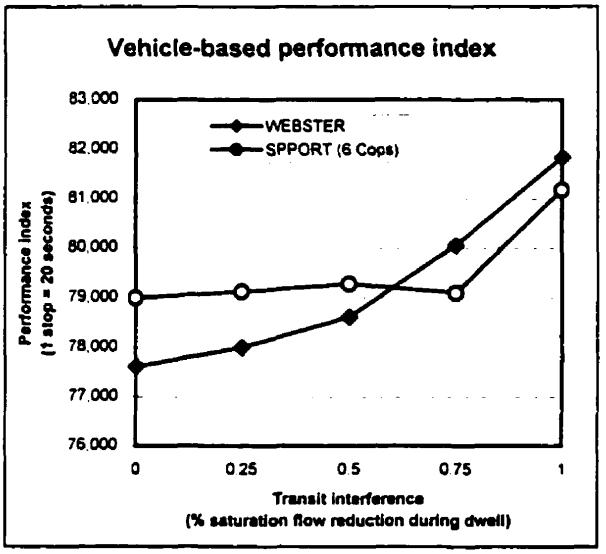


Isolated intersection scenario #4
(Low constant demand with transit vehicles)

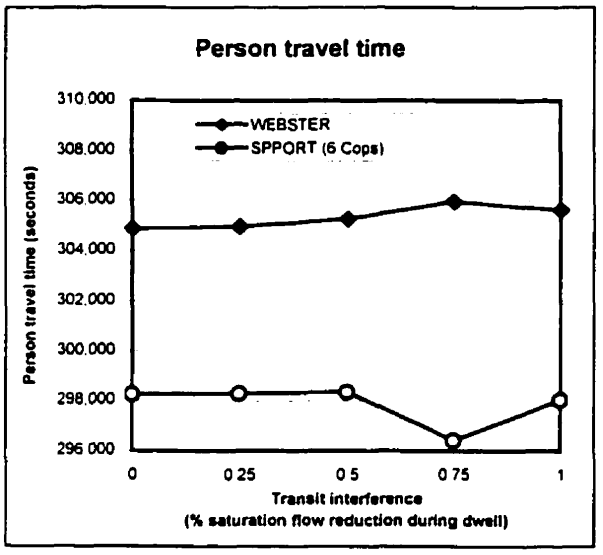


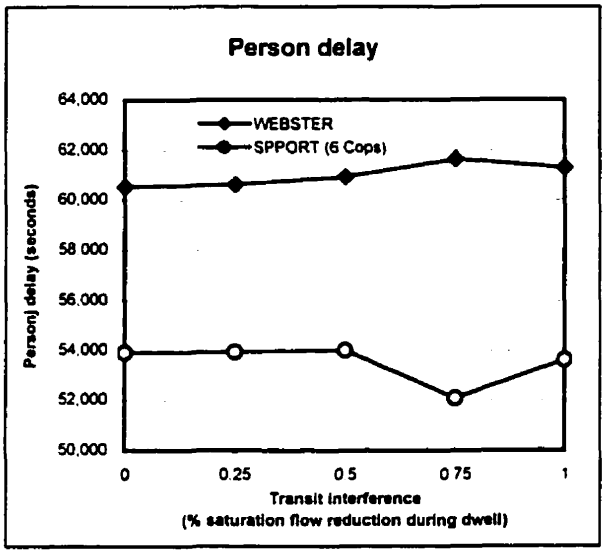


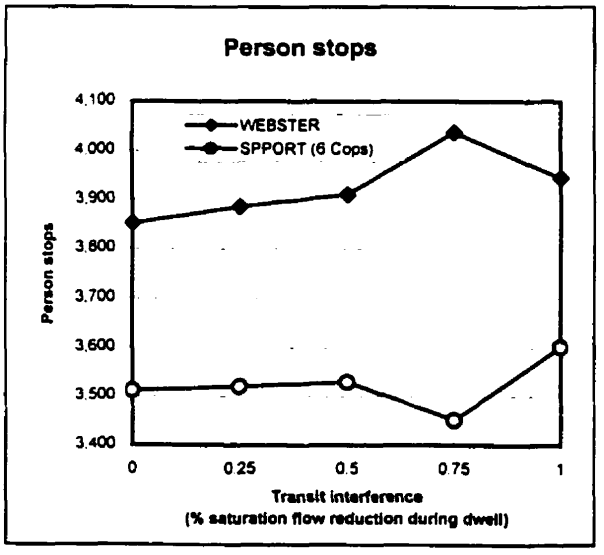


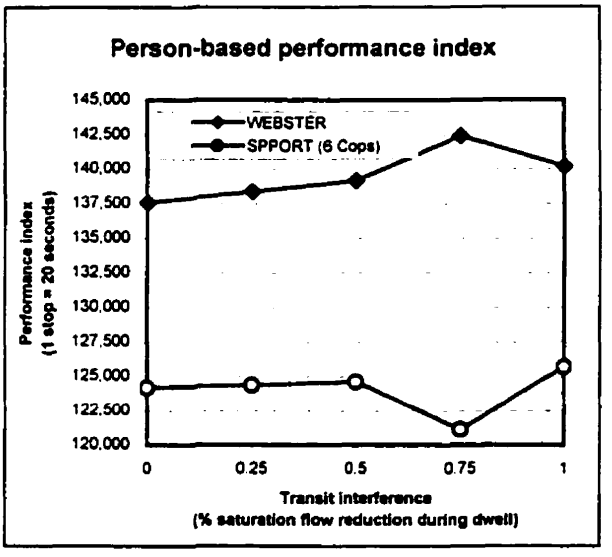


Isolated intersection scenario #4
(Low constant demand with transit vehicles)

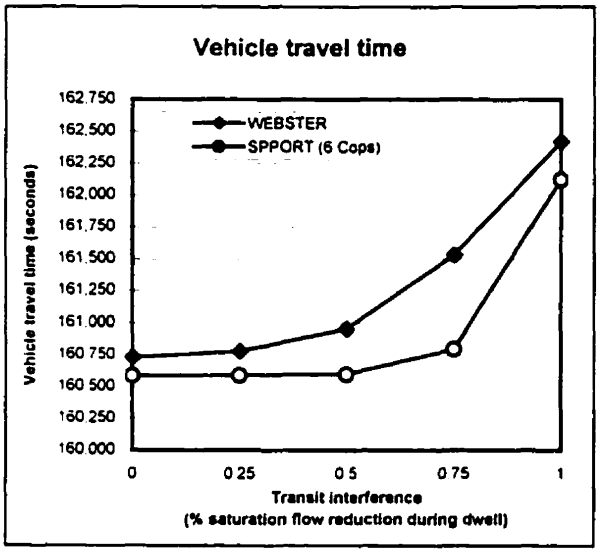


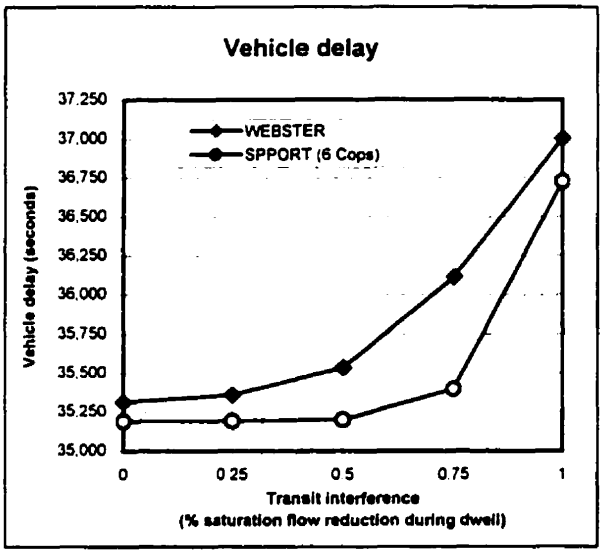


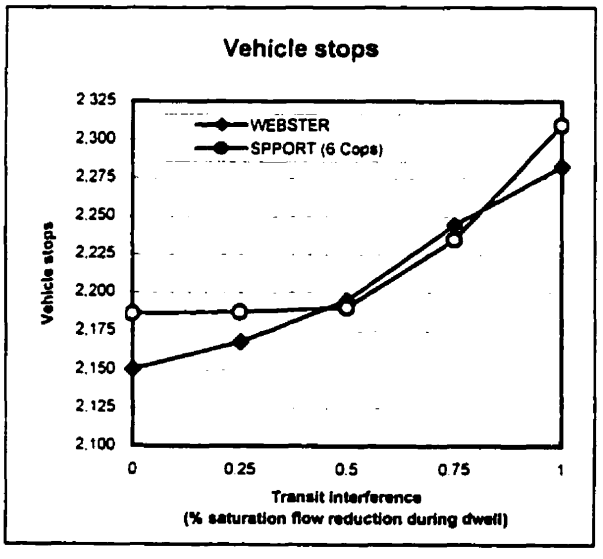


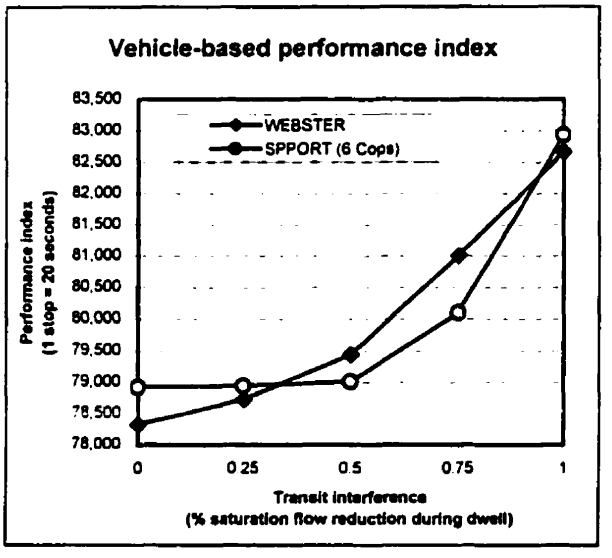


Isolated intersection scenario #4p
(Low peaking demand with transit vehicles)

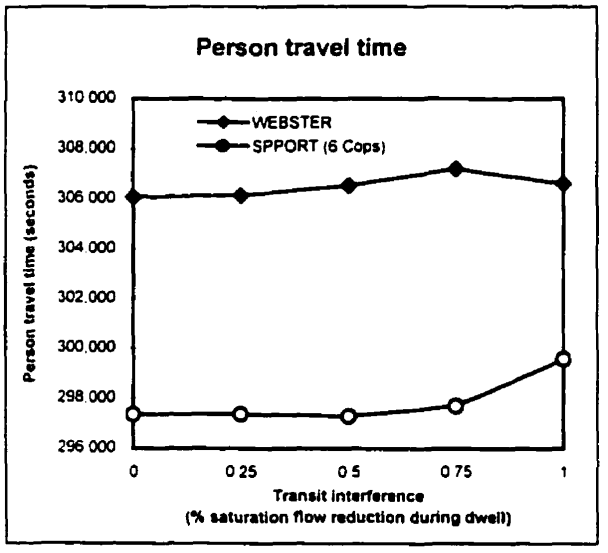


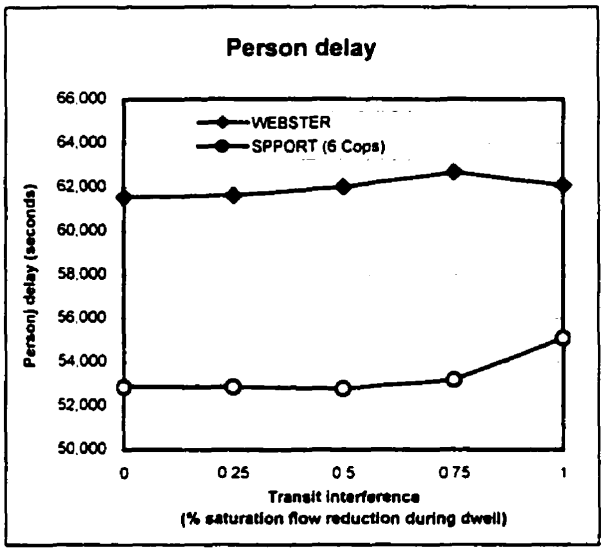


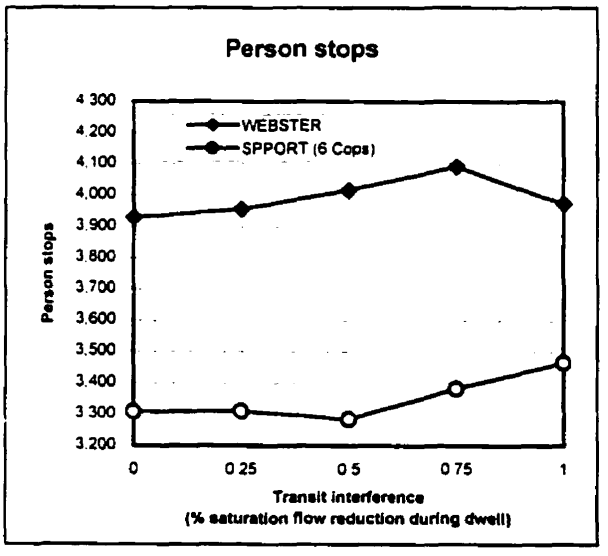


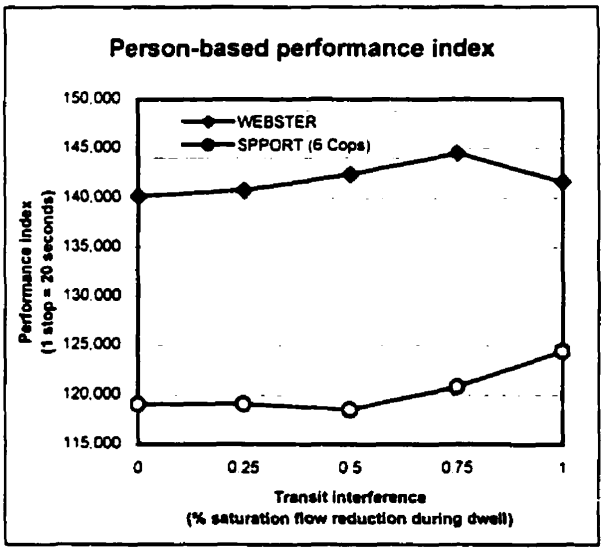


Isolated intersection scenario #4p
(Low peaking demand with transit vehicles)

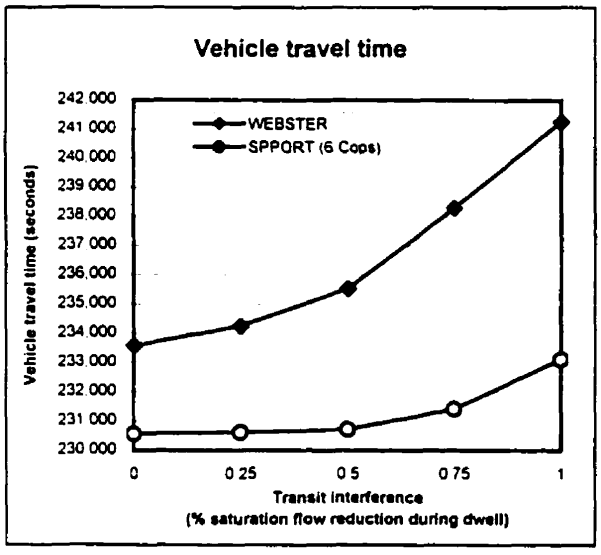


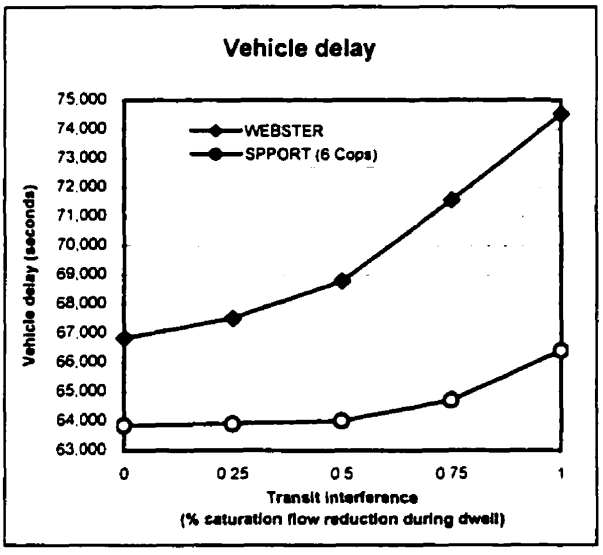


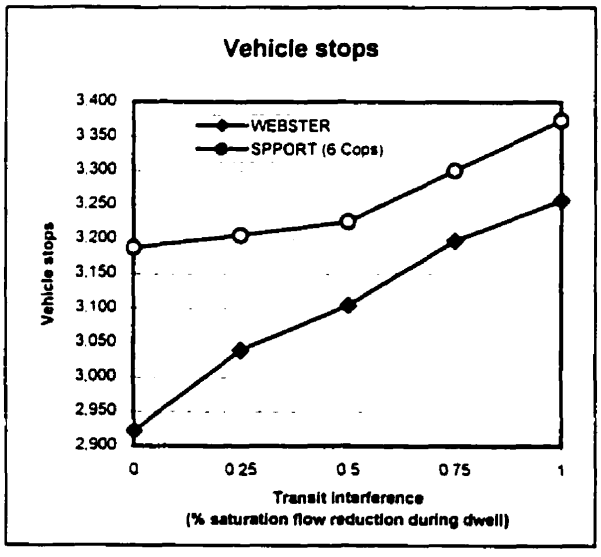


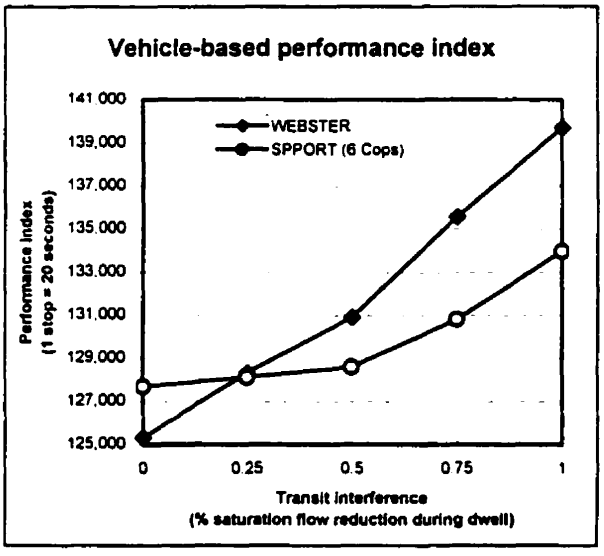


Isolated intersection scenario #5
(Medium constant demand with transit vehicles)

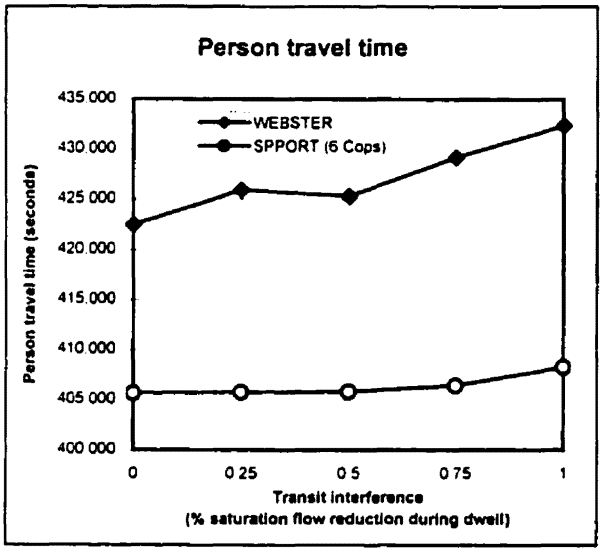


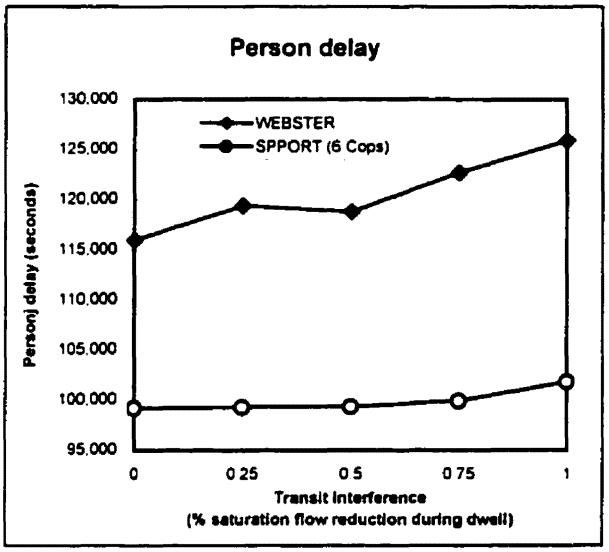


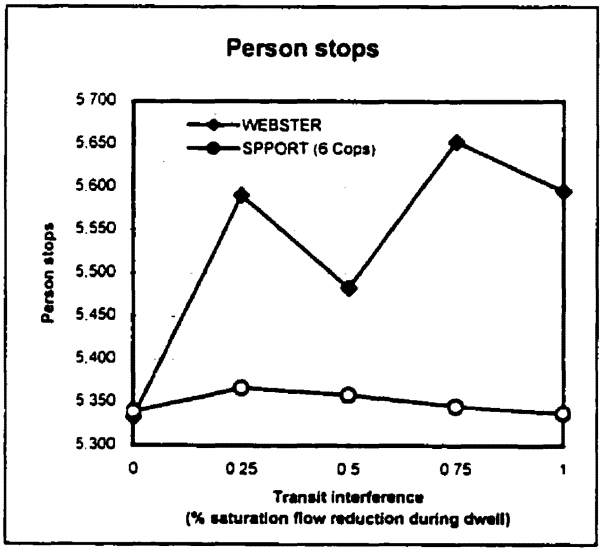


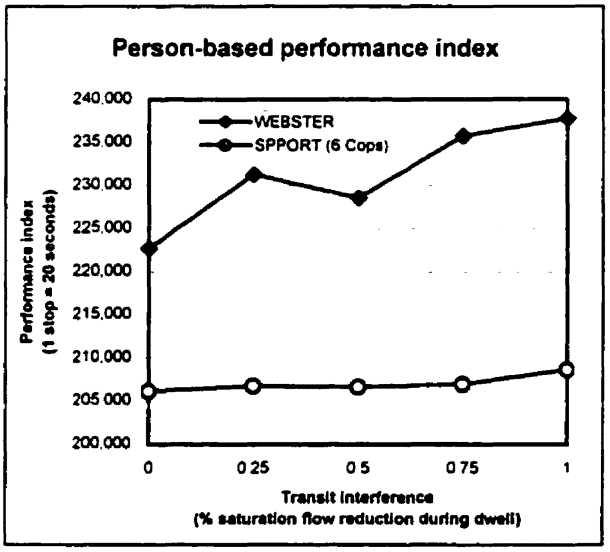


Isolated intersection scenario #5
(Medium constant demand with transit vehicles)

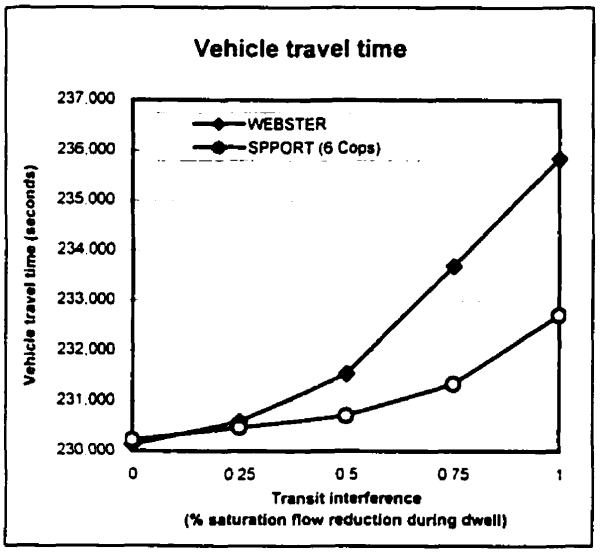


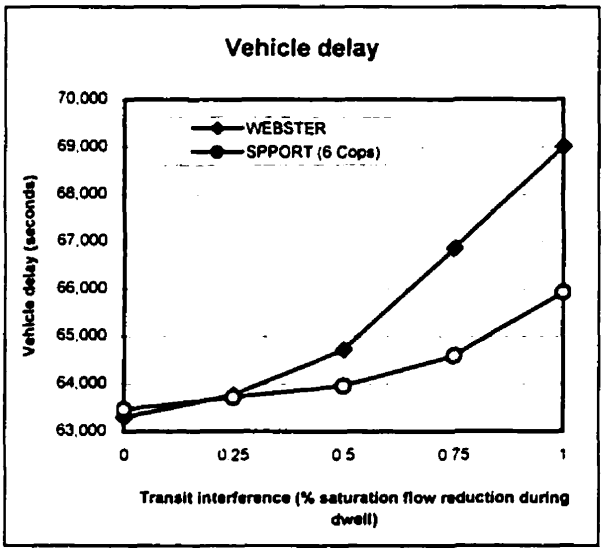


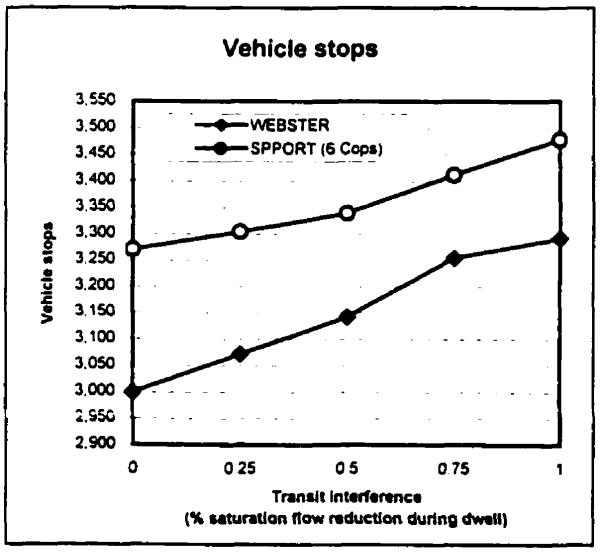


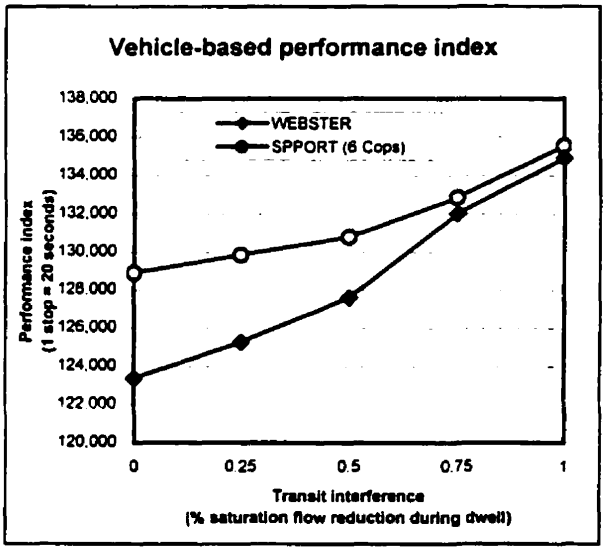


Isolated intersection scenario #5p
(Medium peaking demand with transit vehicles)

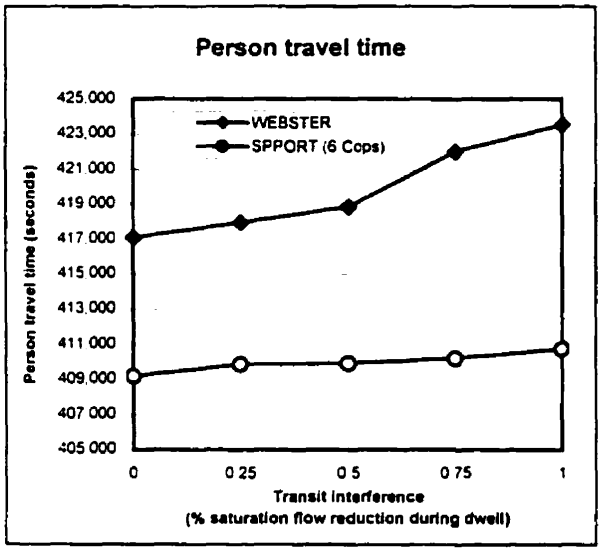


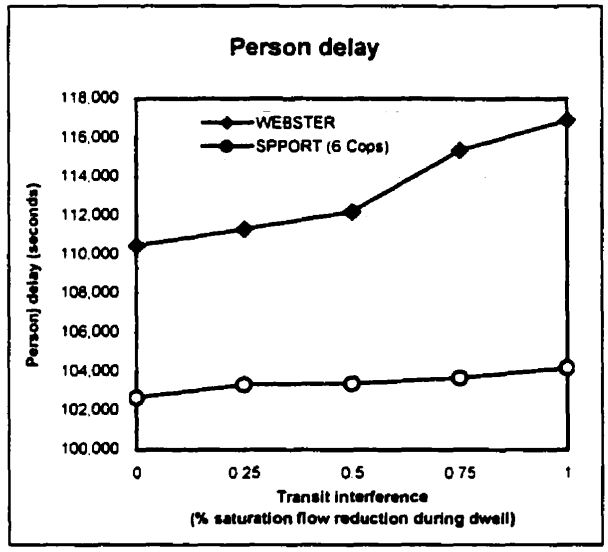


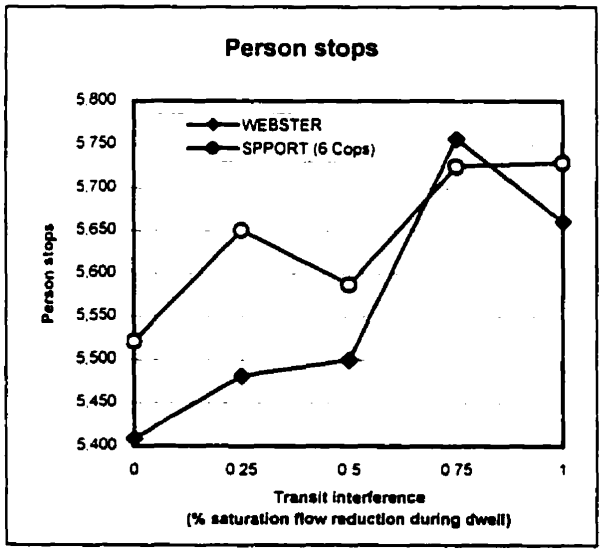


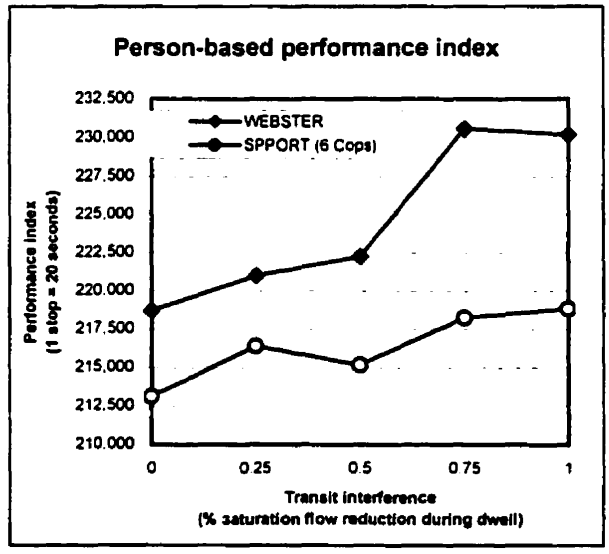


Isolated intersection scenario #5p
(Medium peaking demand with transit vehicles)

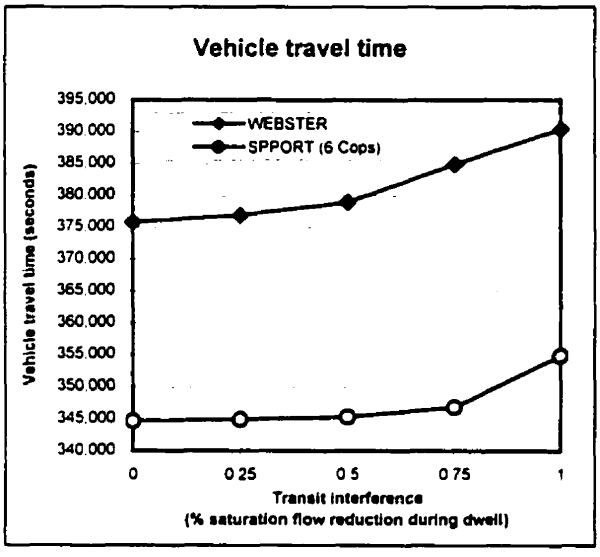


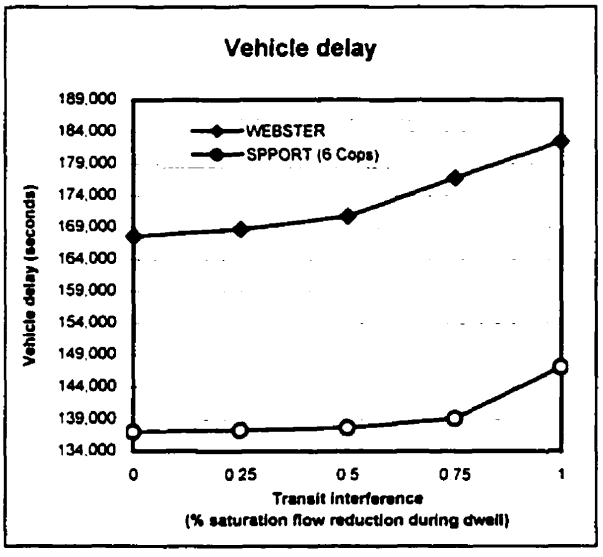


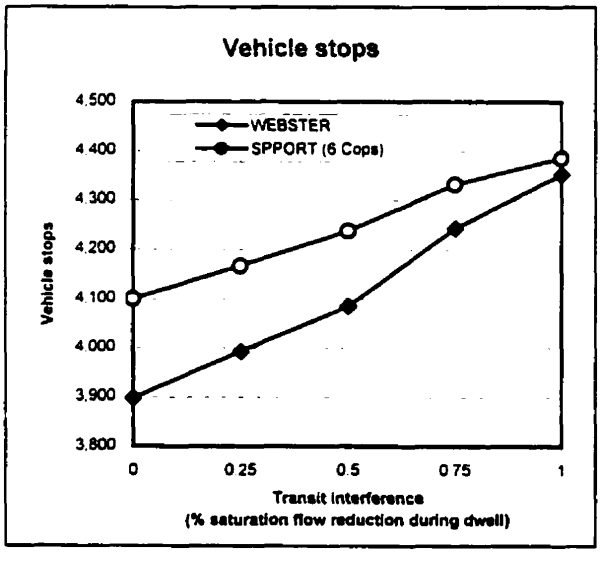


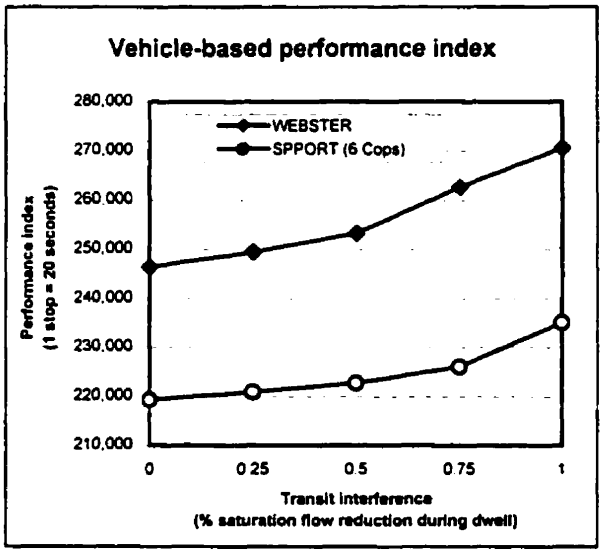


Isolated intersection scenario #6
(High constant demand with transit vehicles)

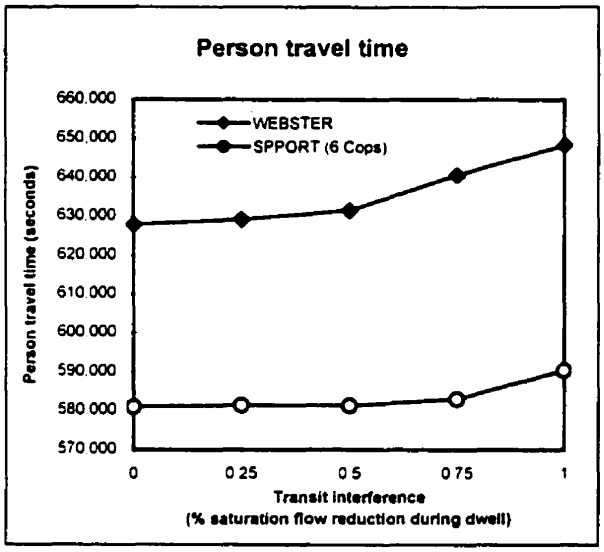


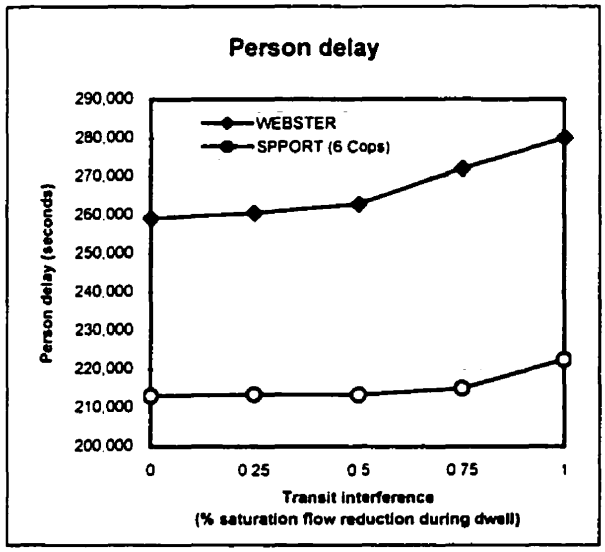


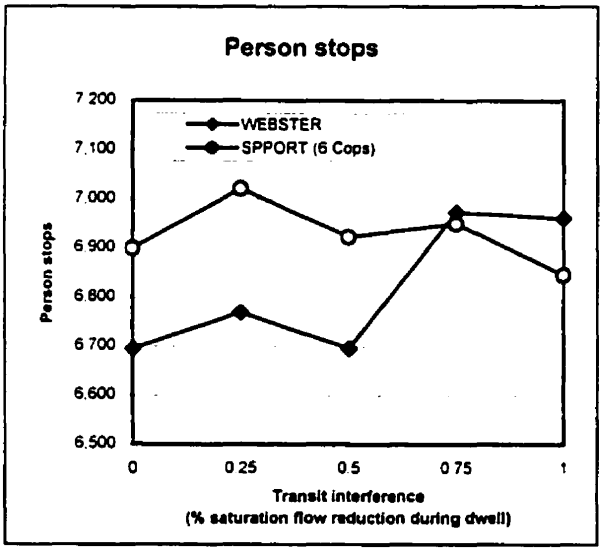


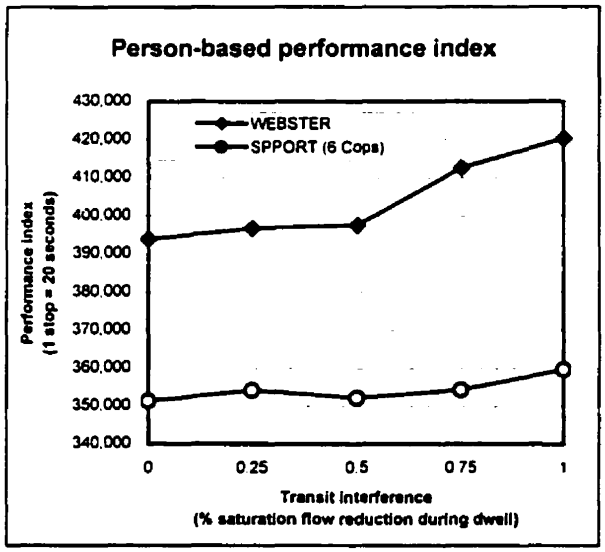


Isolated intersection scenario #6
(High constant demand with transit vehicles)

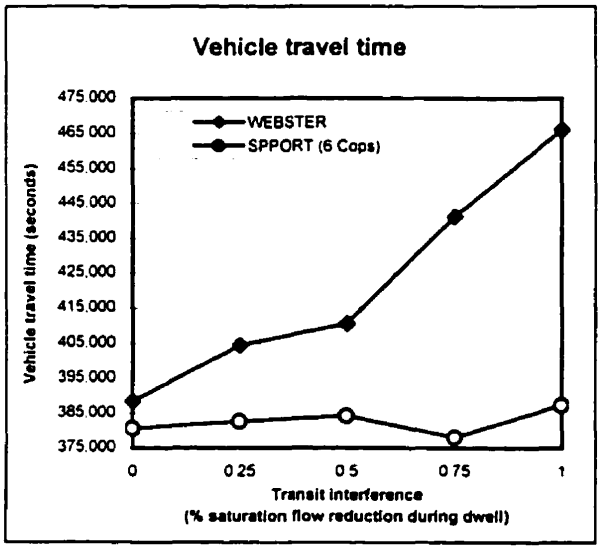


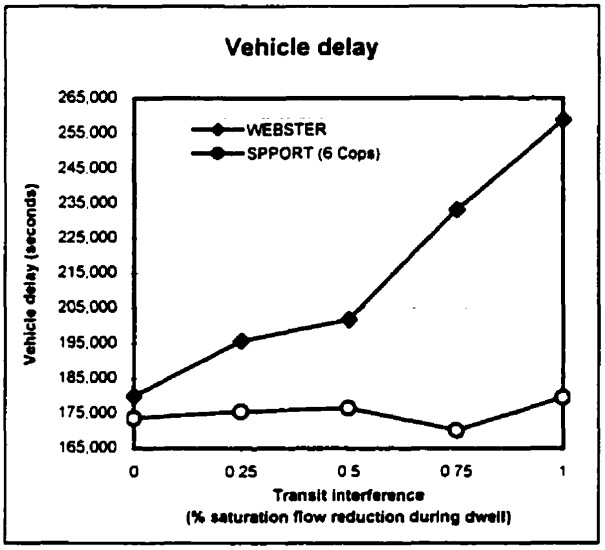


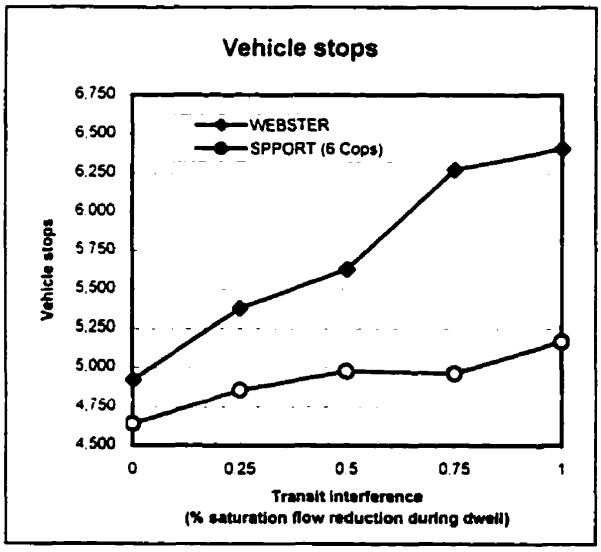


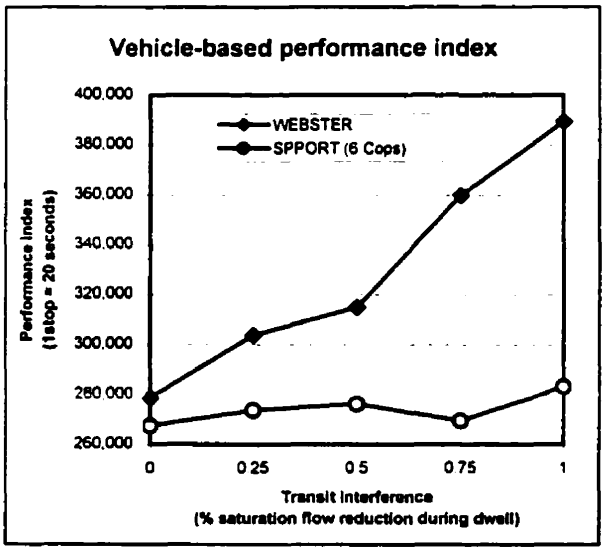


Isolated intersection scenario #6p
(High peaking demand with transit vehicles)

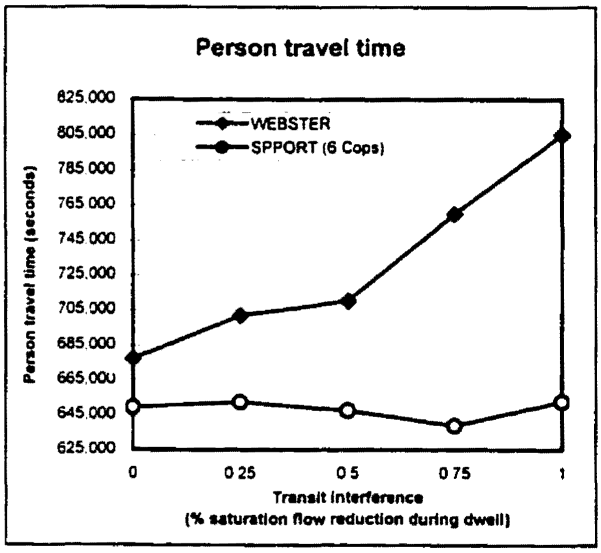


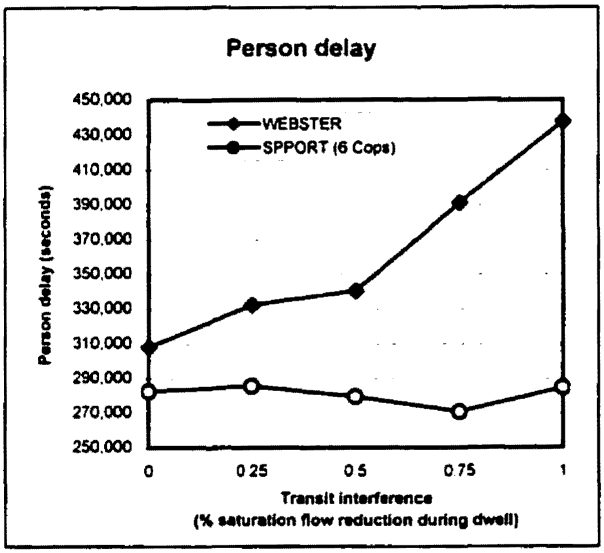


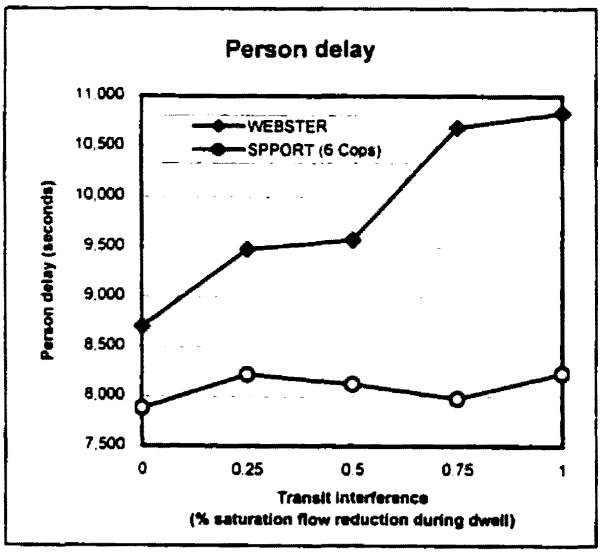


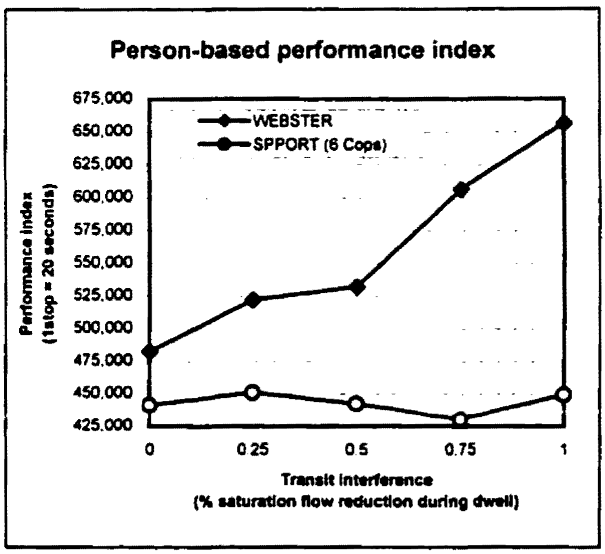


Isolated intersection scenario #6p
(High peaking demand with transit vehicles)





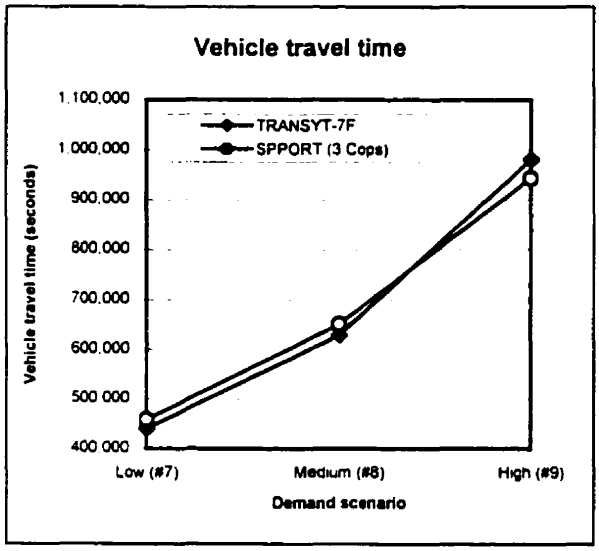


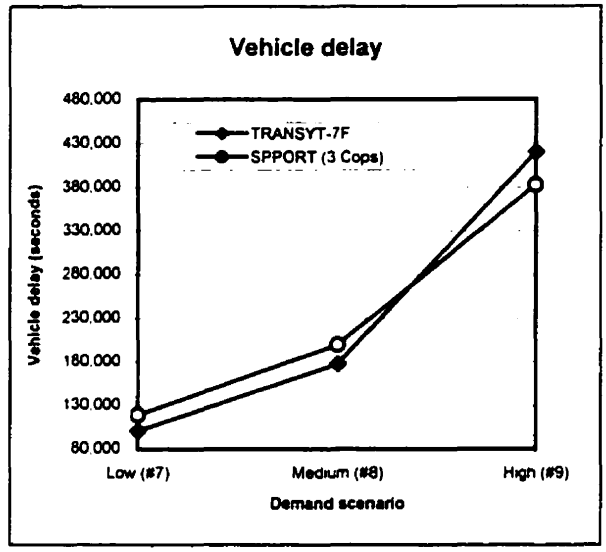


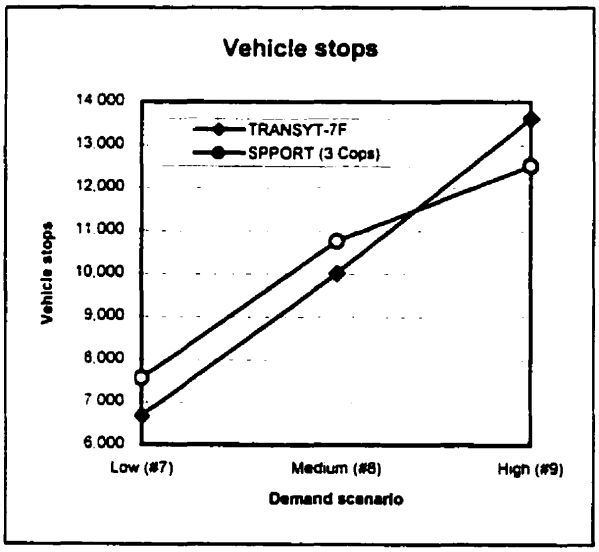
APPENDIX F

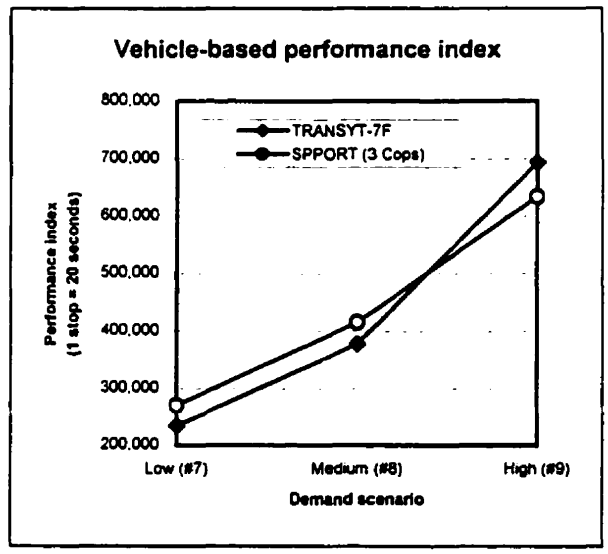
Comparison Results Between SPPORT and Fixed-Time Control for Arterial Scenarios

Arterial scenarios #7, #8, #9
(Constant demand without transit vehicles)

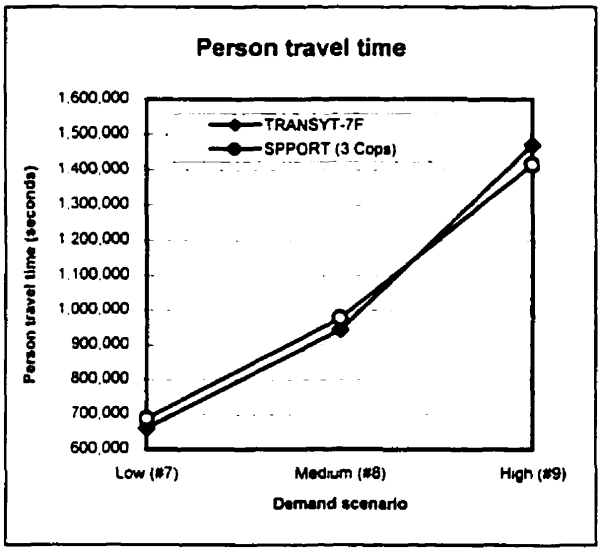


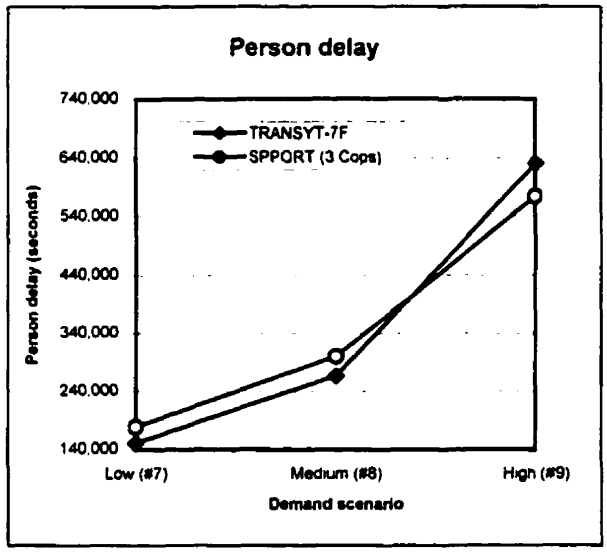


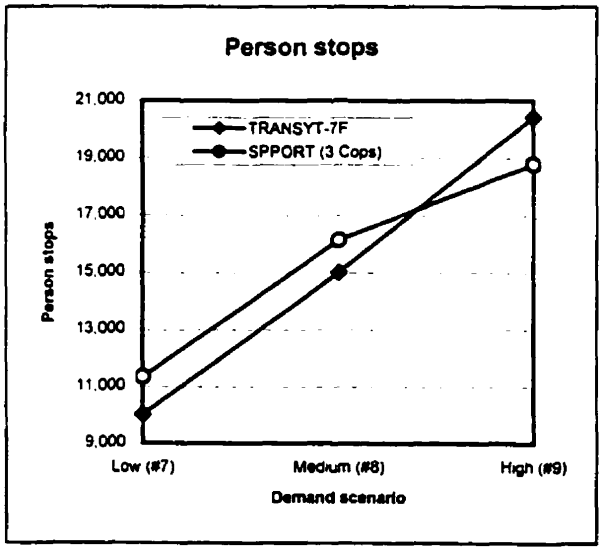


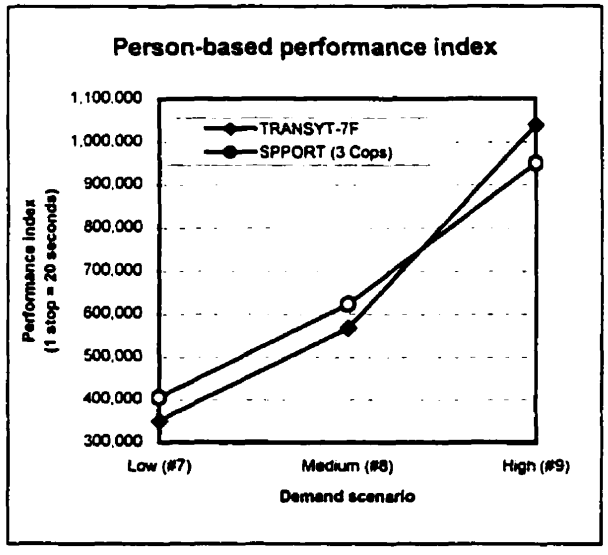


Arterial scenarios #7, #8, #9
(Constant demand without transit vehicles)

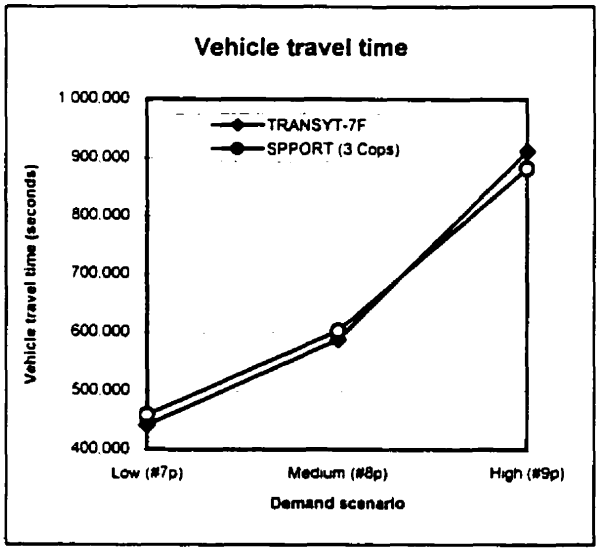


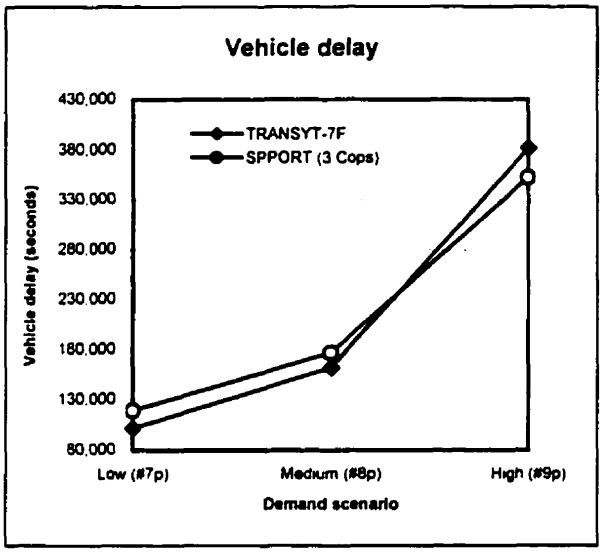


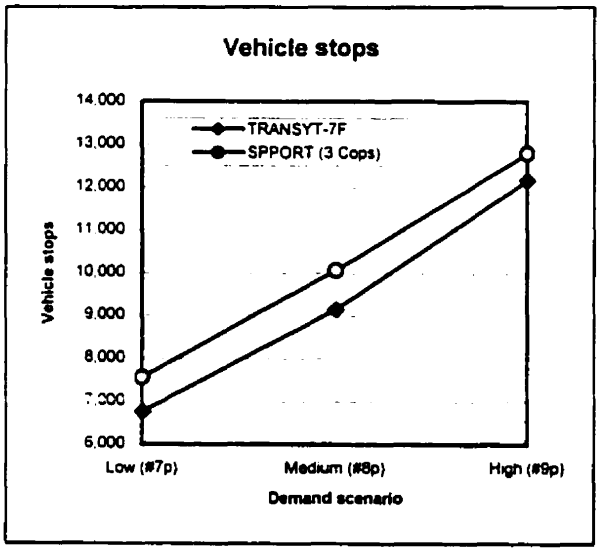


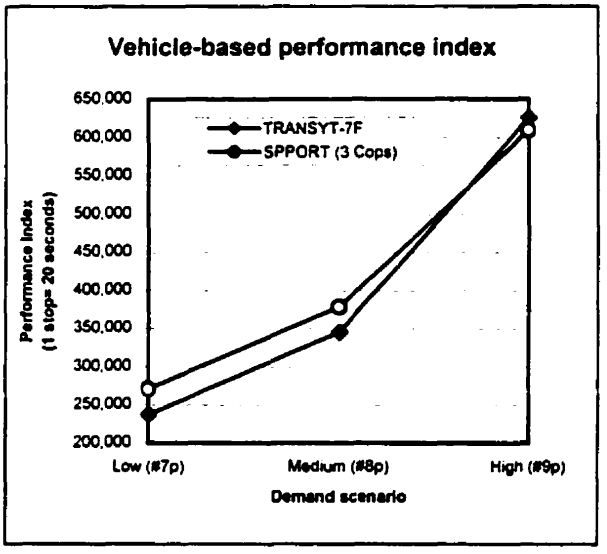


Arterial scenarios #7p, #8p, #9p
(Constant demand without transit vehicles)

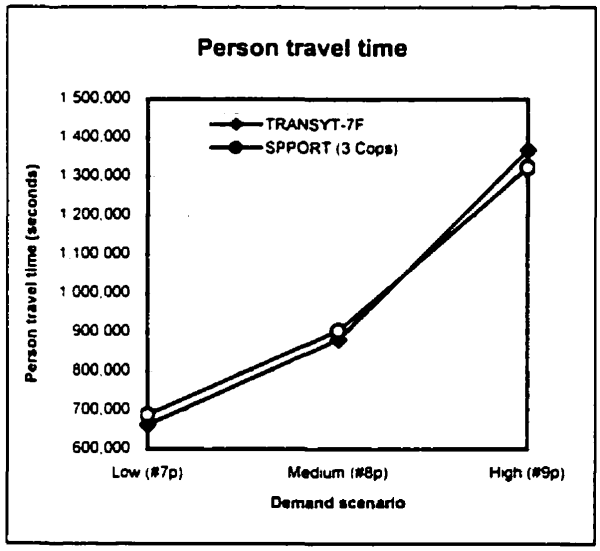


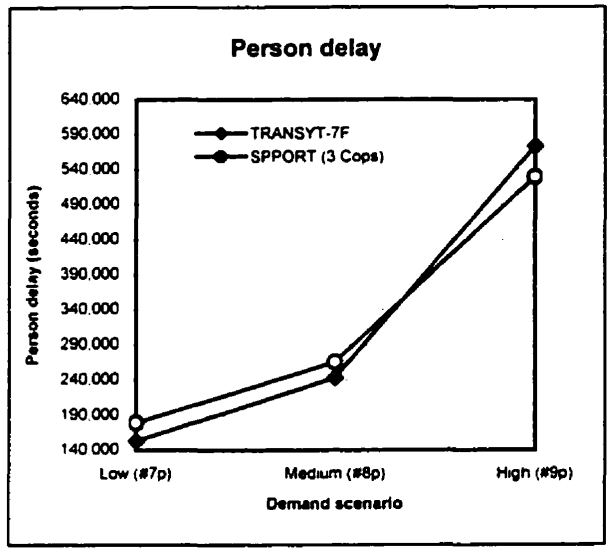


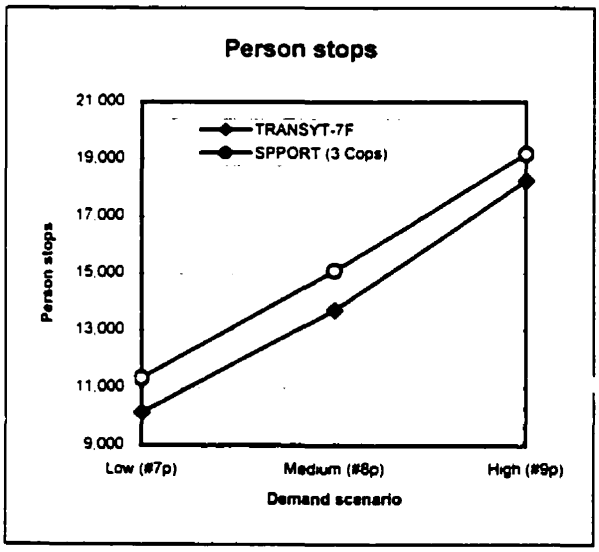


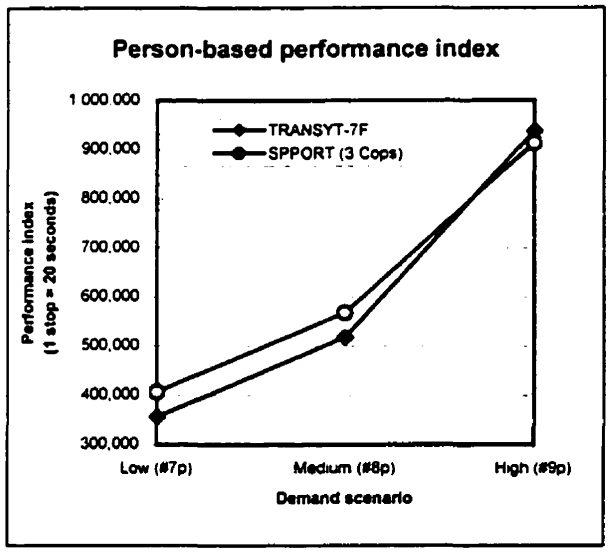


Arterial scenarios #7p, #8p, #9p
(Constant demand without transit vehicles)

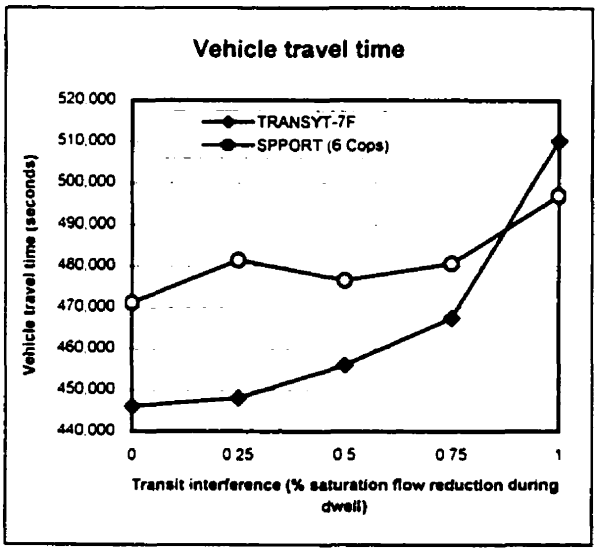


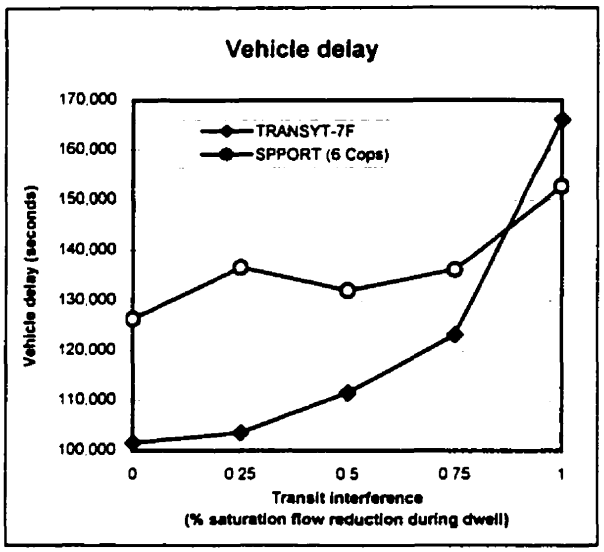


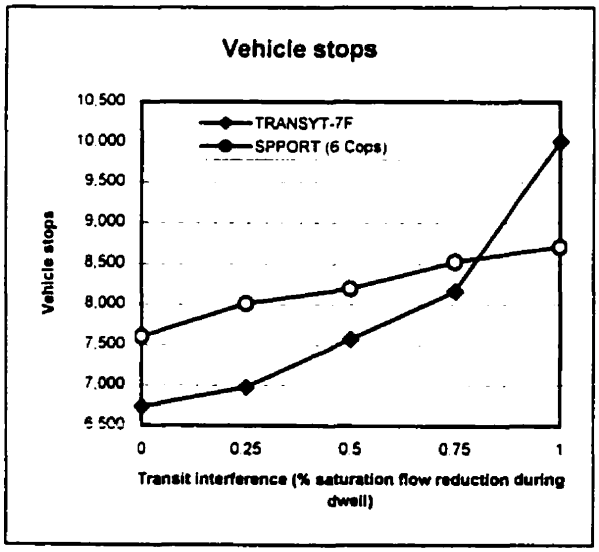


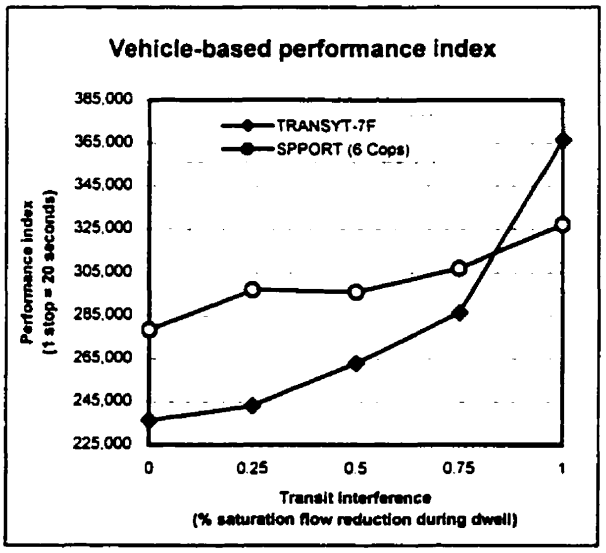


Arterial scenario #10
(Low constant demand with transit vehicles)

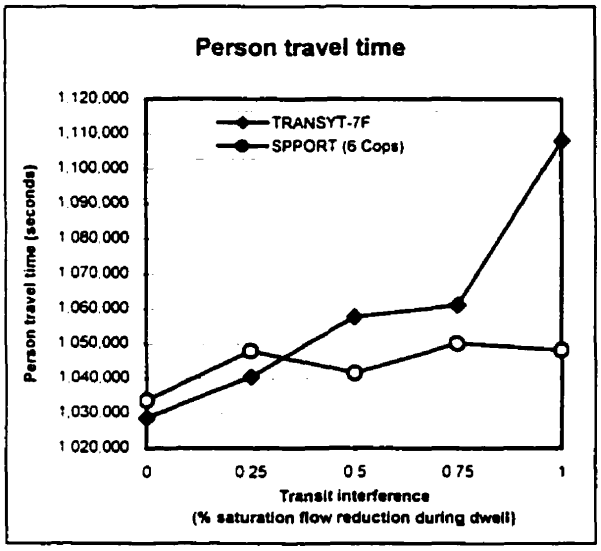


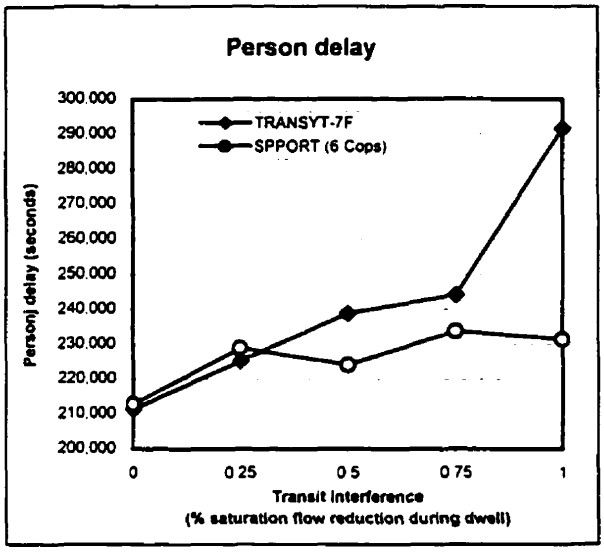


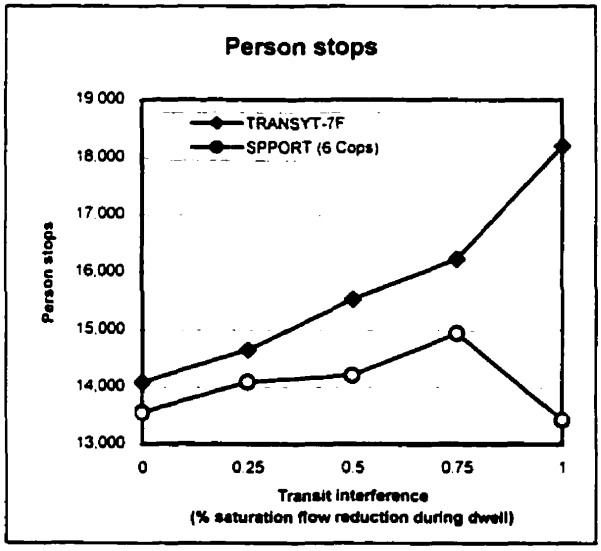


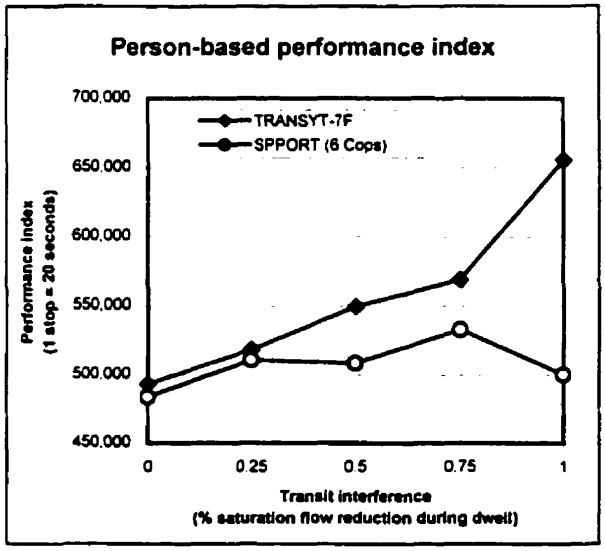


Arterial scenario #10
(Low constant demand with transit vehicles)

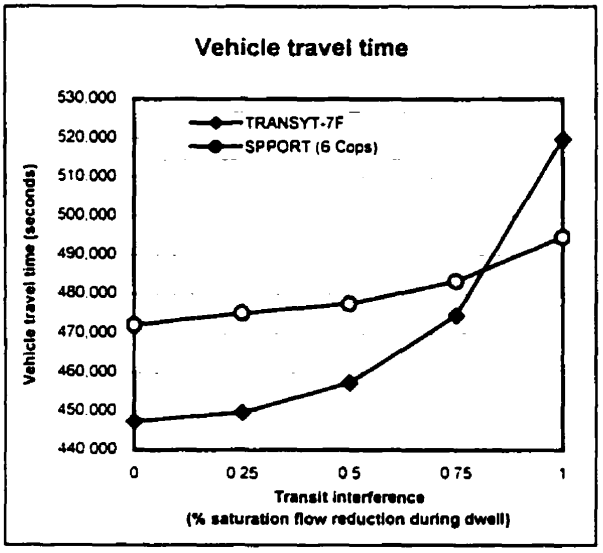


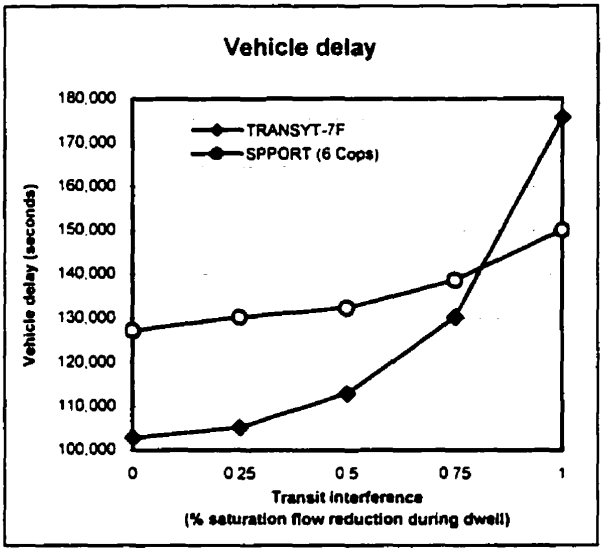


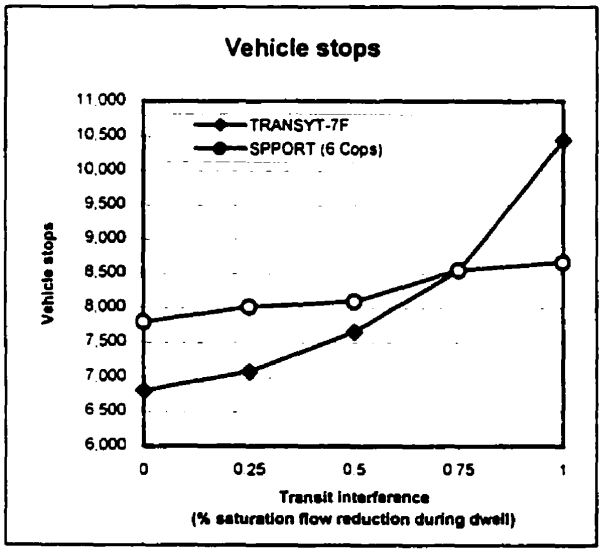


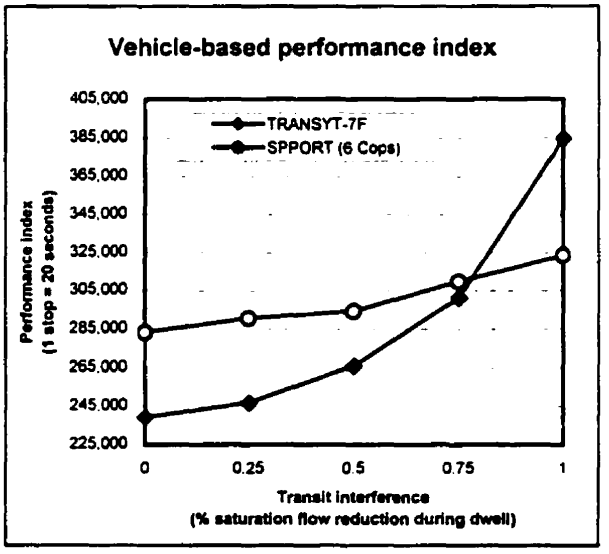


Arterial scenario #10p
(Low peaking demand with transit vehicles)

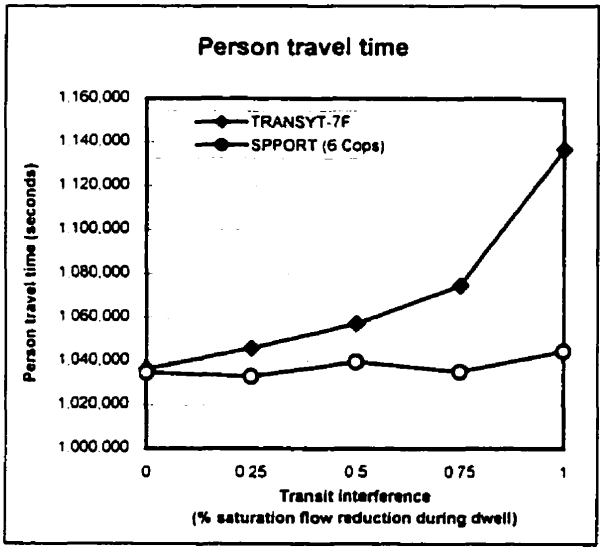


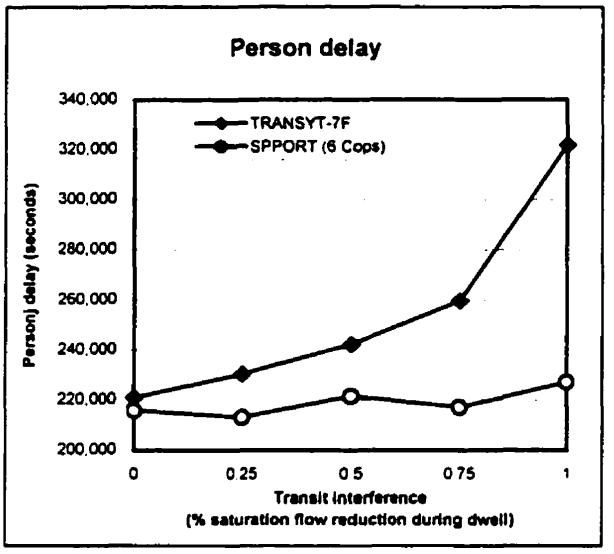


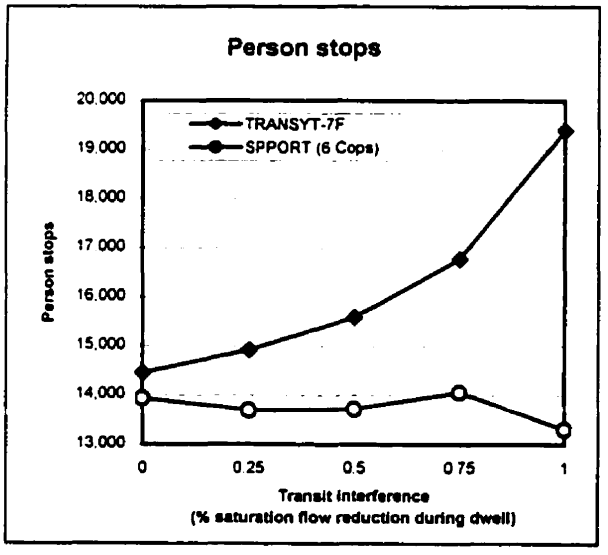


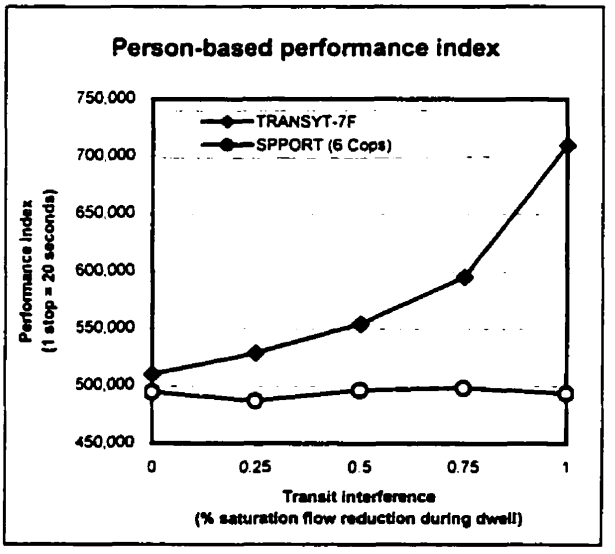


Arterial scenario #10p
(Low peaking demand with transit vehicles)

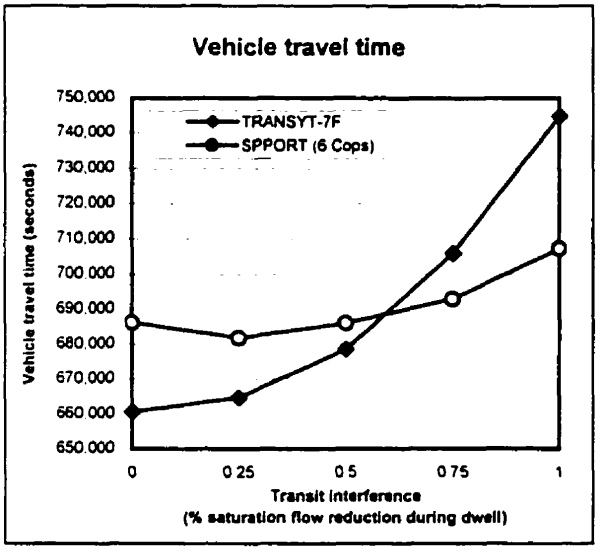


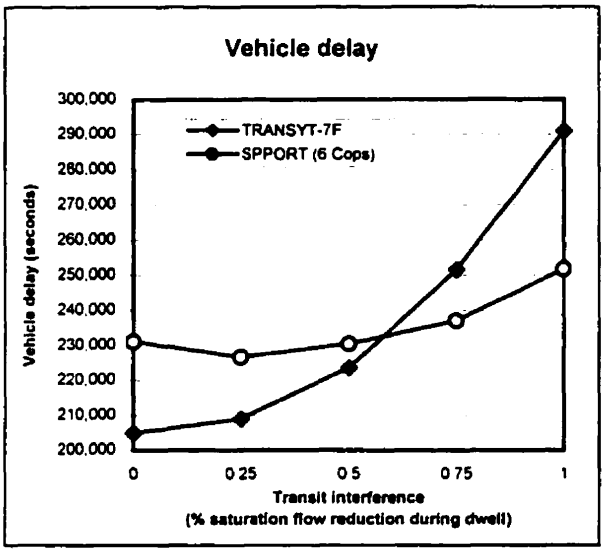


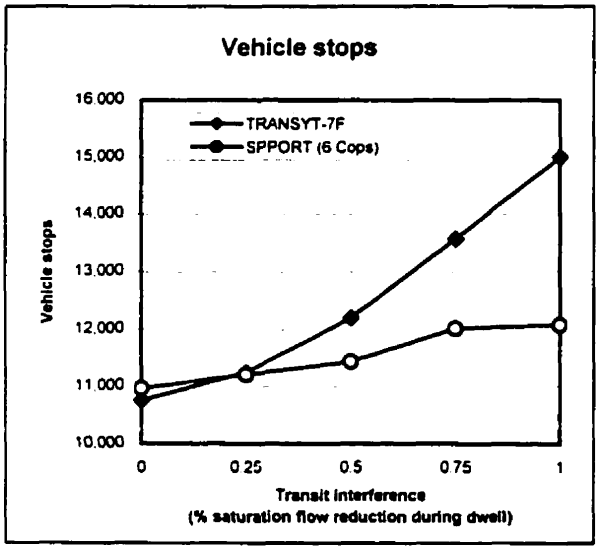


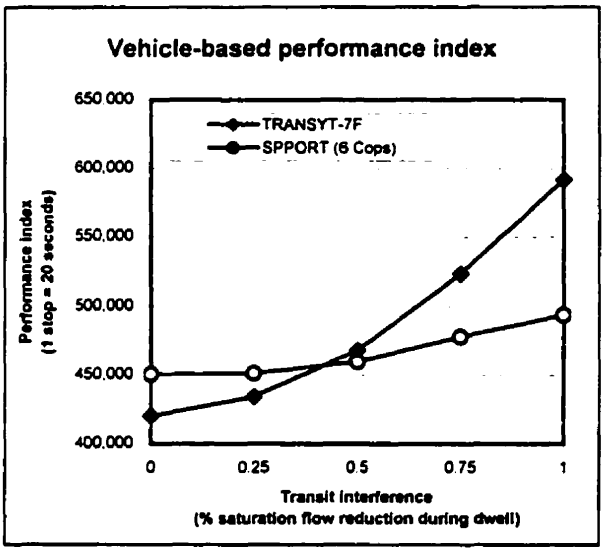


Arterial scenario #11
(Medium constant demand with transit vehicles)

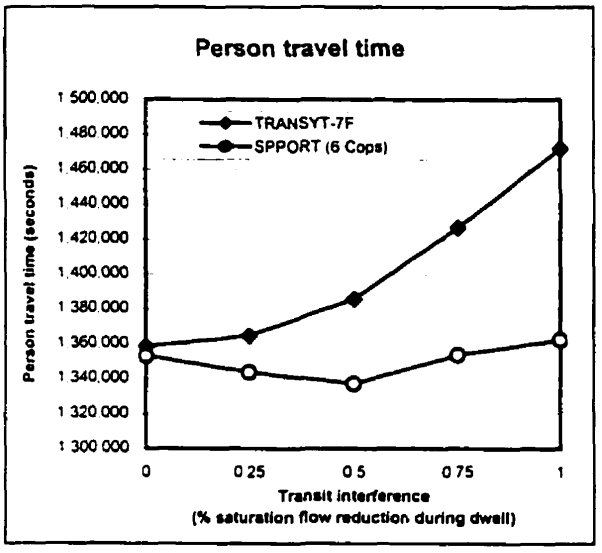


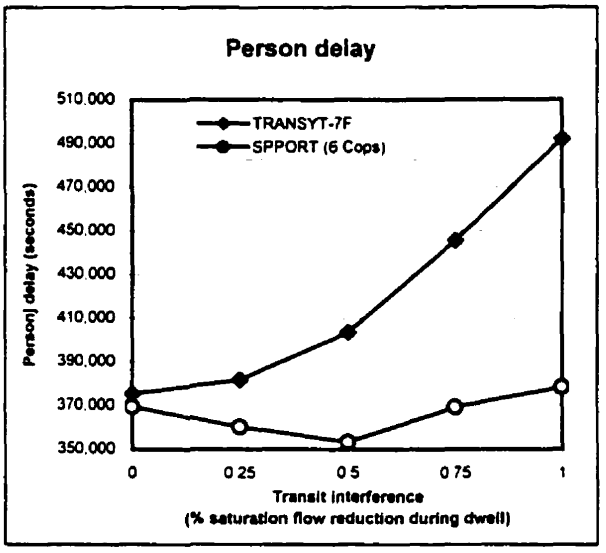


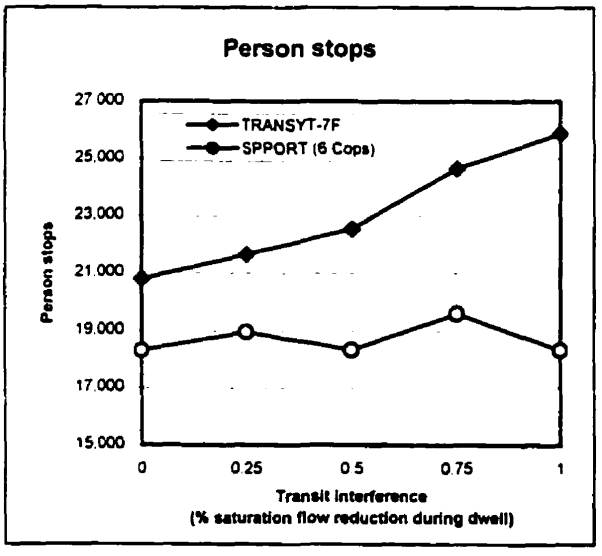


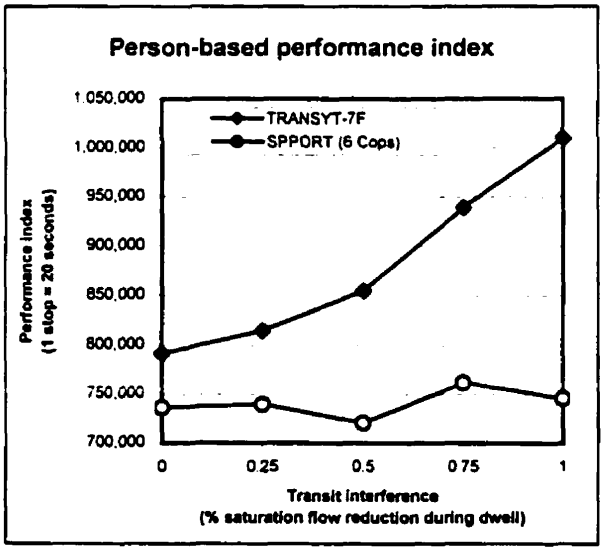


Arterial scenario #11
(Medium constant demand with transit vehicles)

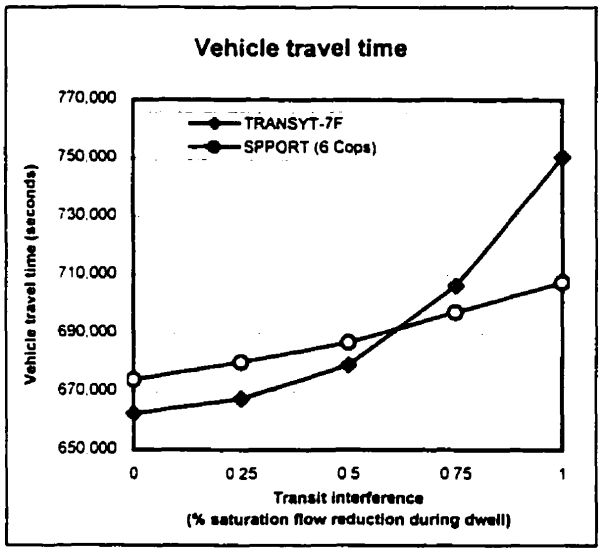


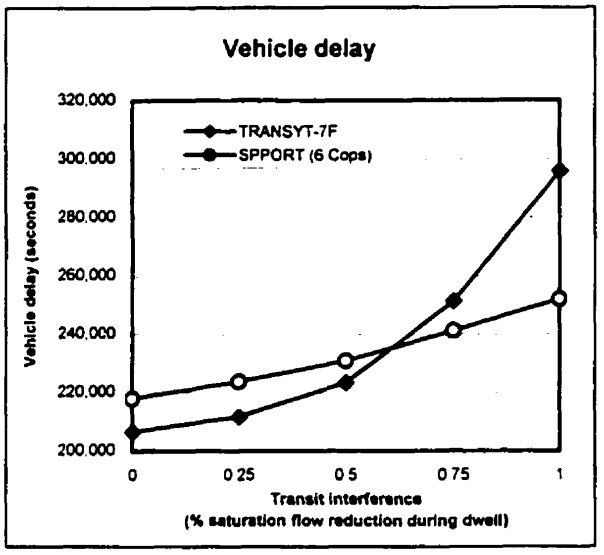


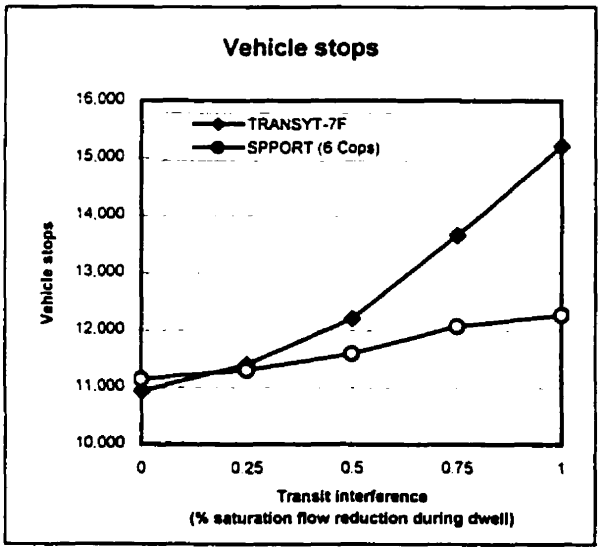


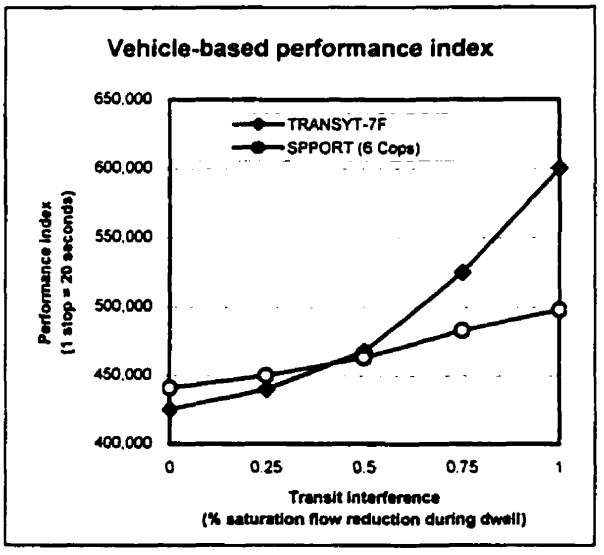


Arterial scenario #11p
(Medium peaking demand with transit vehicles)

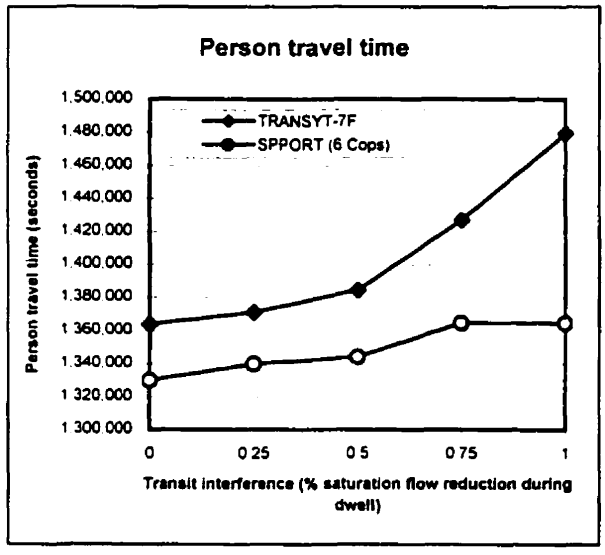


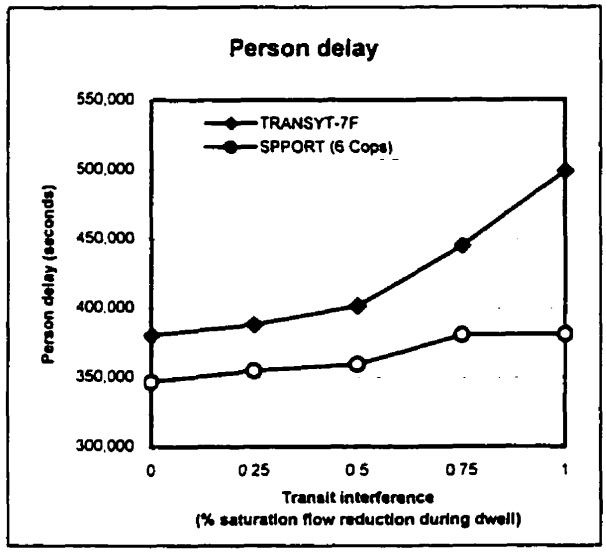


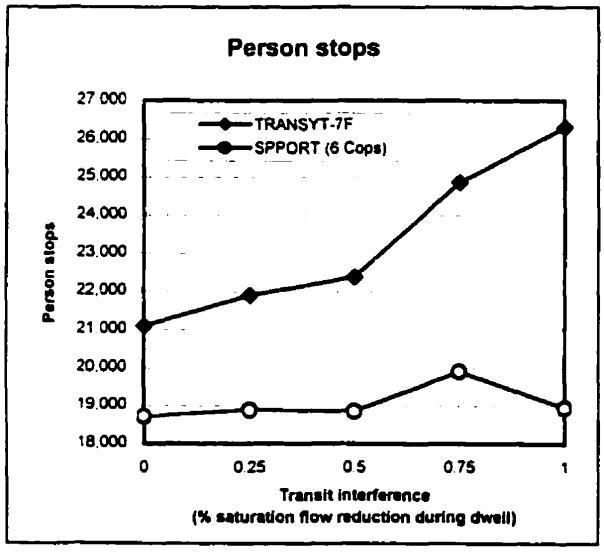


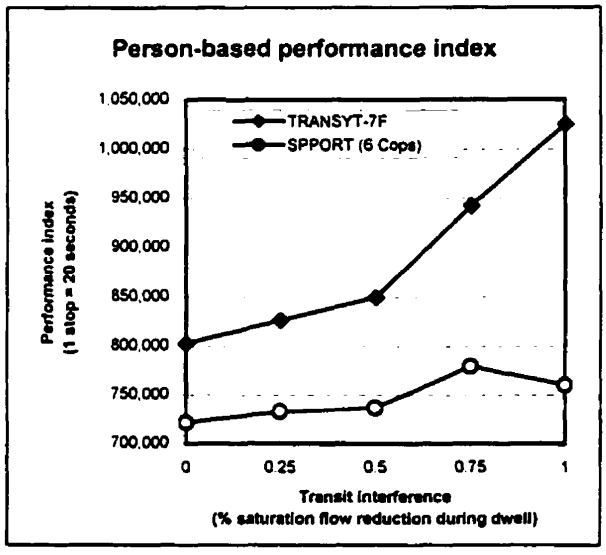


Arterial scenario #11p
(Medium peaking demand with transit vehicles)

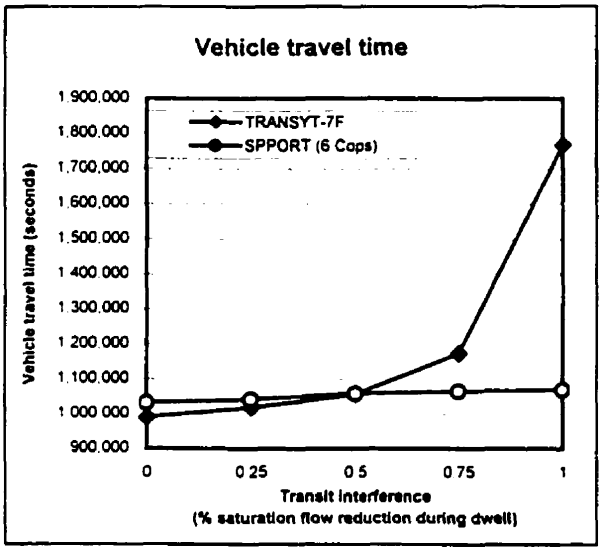


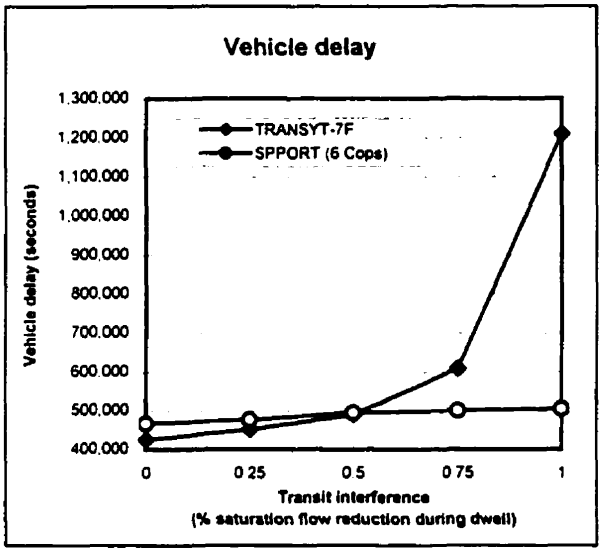


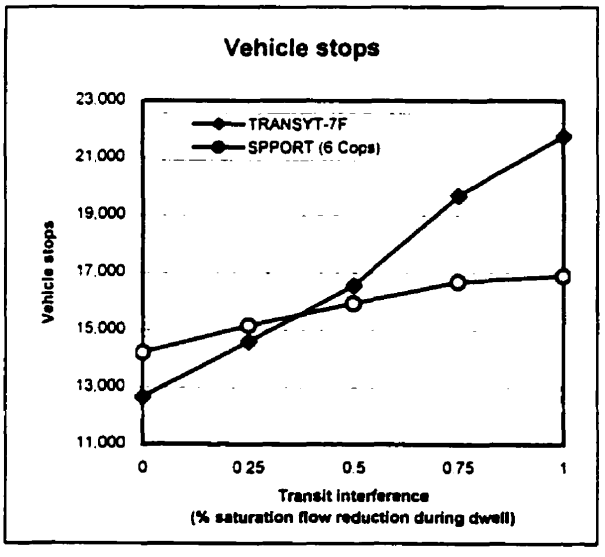


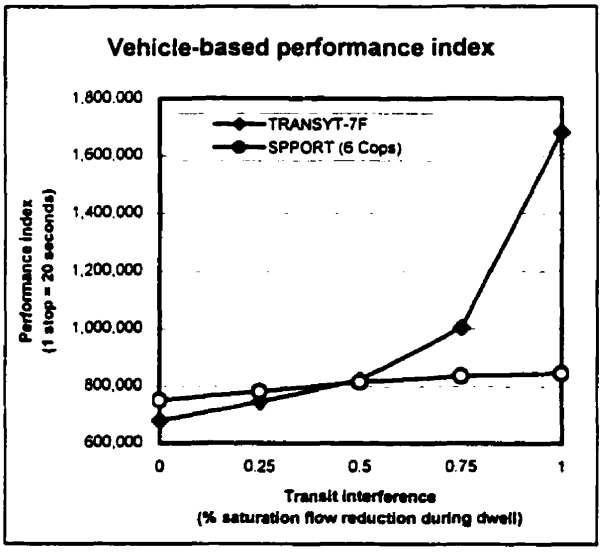


Arterial scenario #12
(High constant demand with transit vehicles)

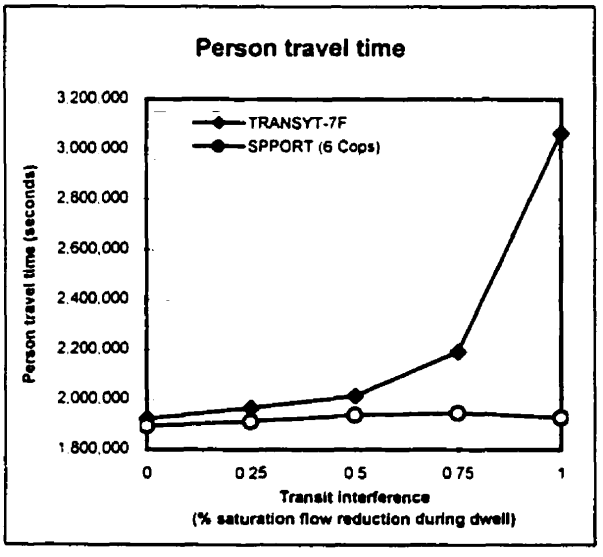


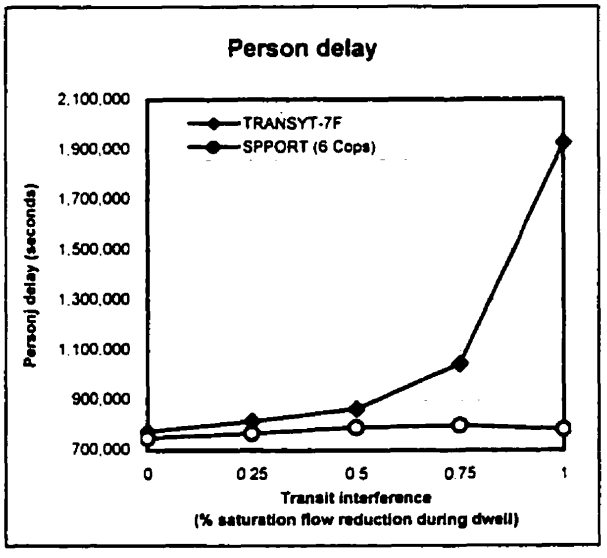


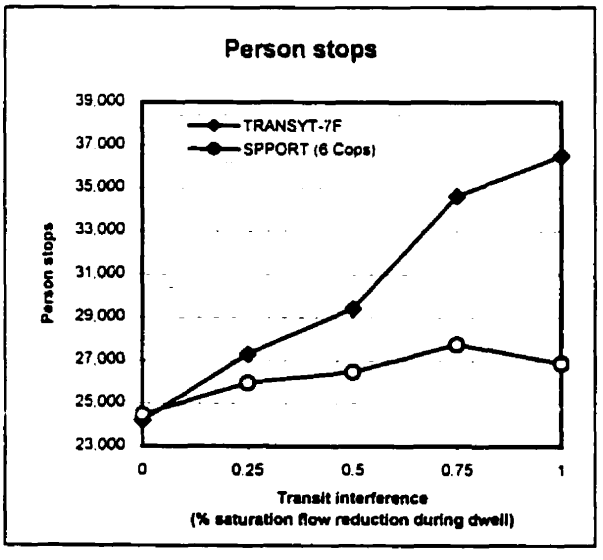


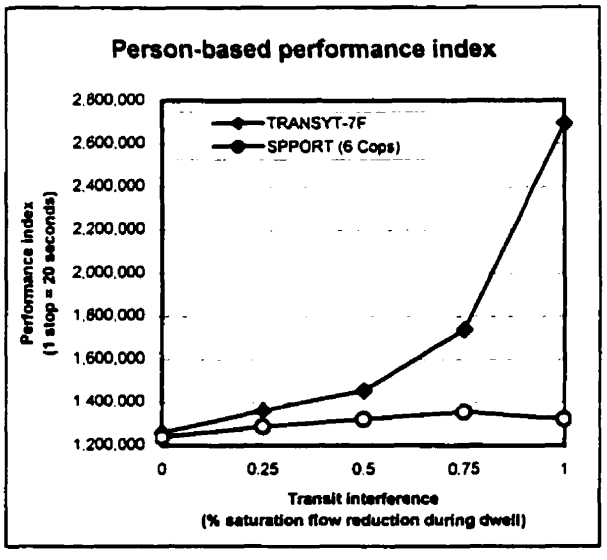


Arterial scenario #12
(High constant demand with transit vehicles)

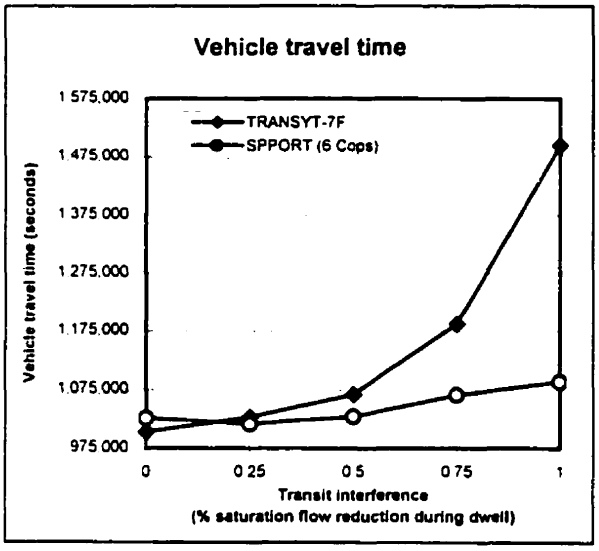


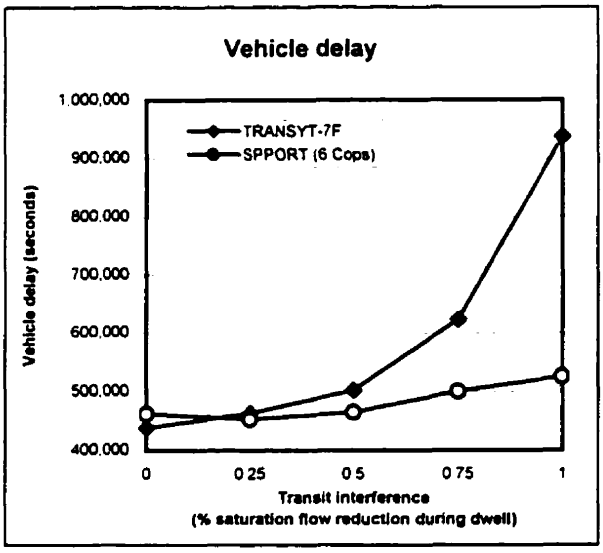


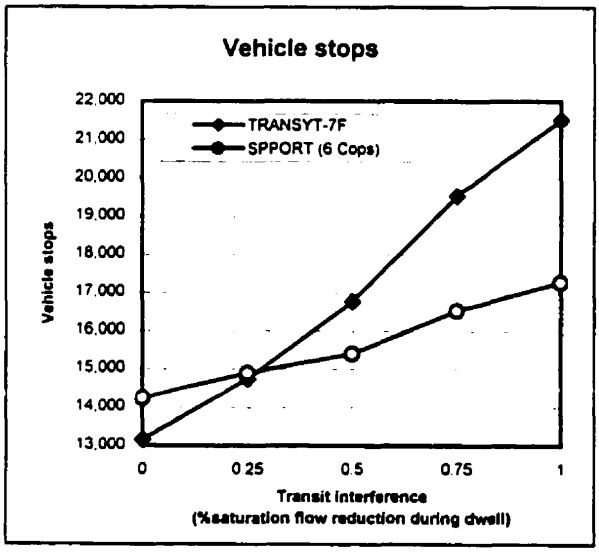


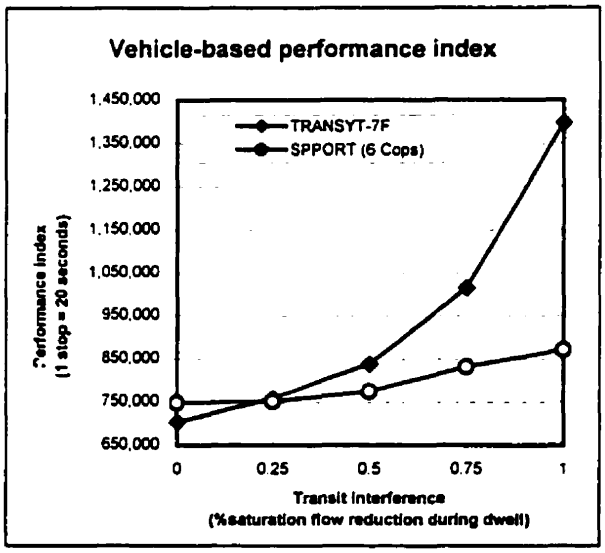


Arterial scenario #12p
(High peaking demand with transit vehicles)

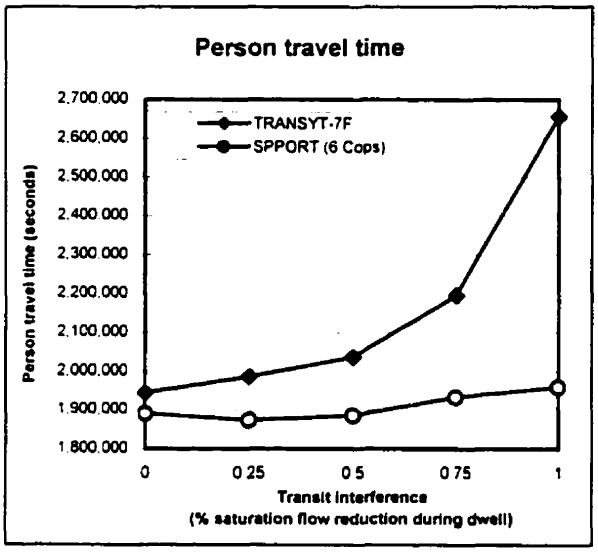


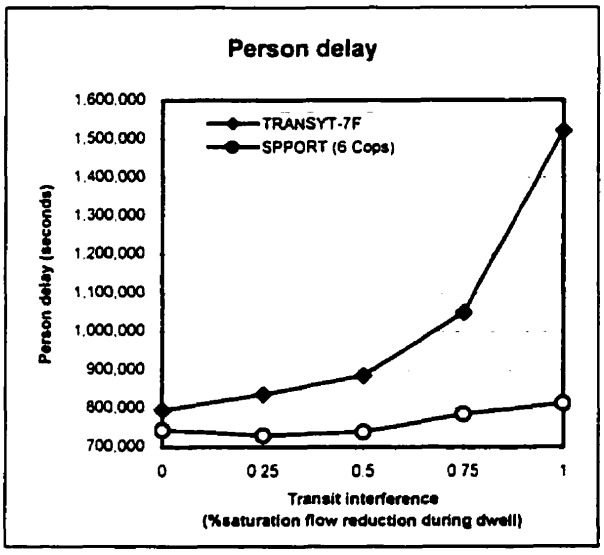


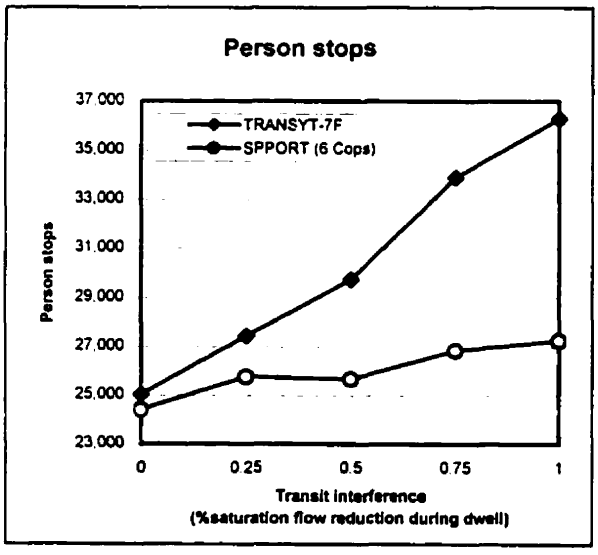


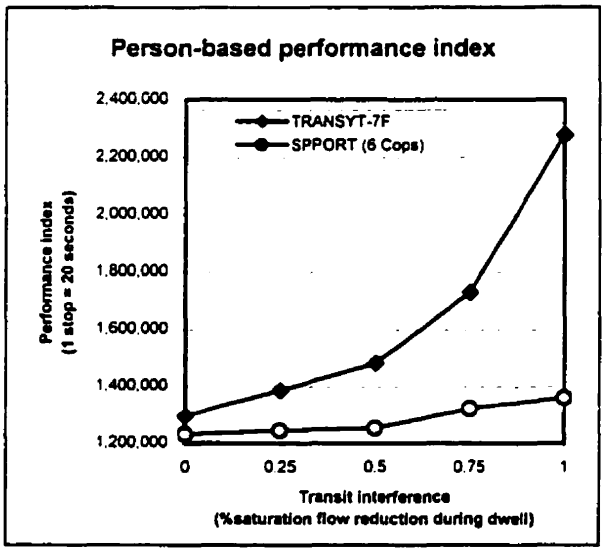


Arterial scenario #12p
(High peaking demand with transit vehicles)





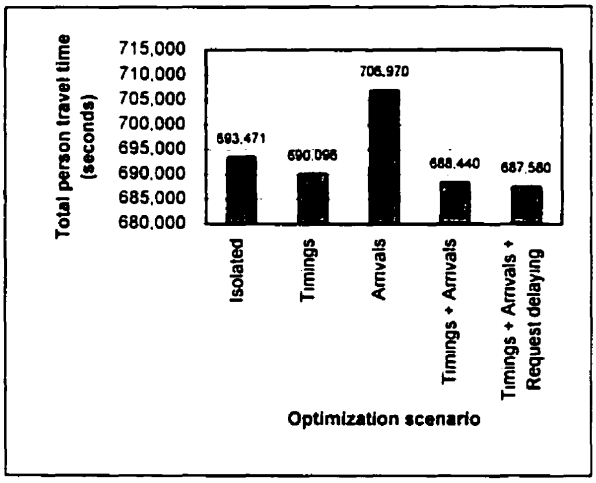


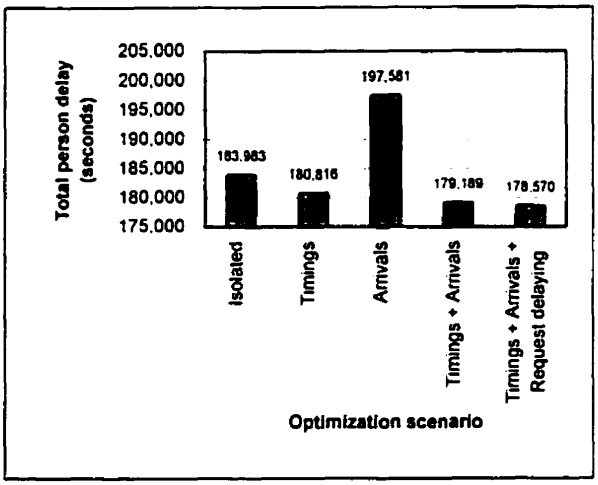


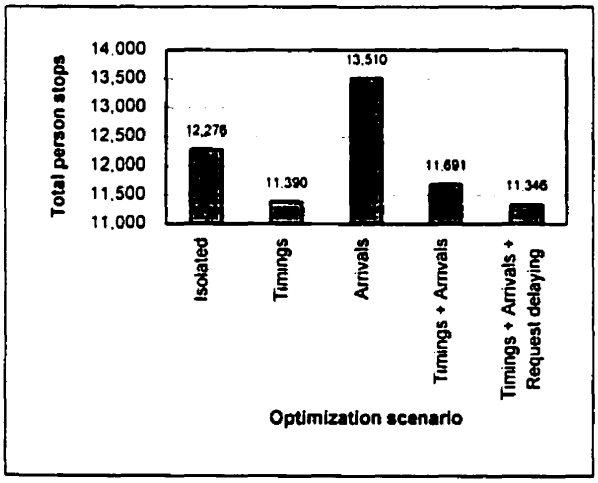
APPENDIX G

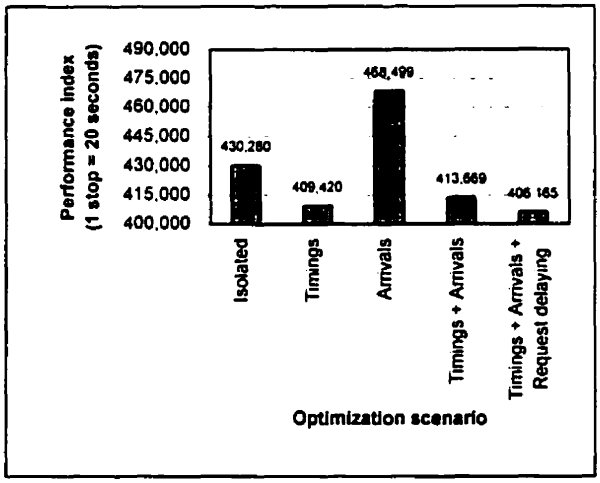
Coordination Logic Evaluation Results

Arterial scenario #7
(Low constant demand without transit vehicles)

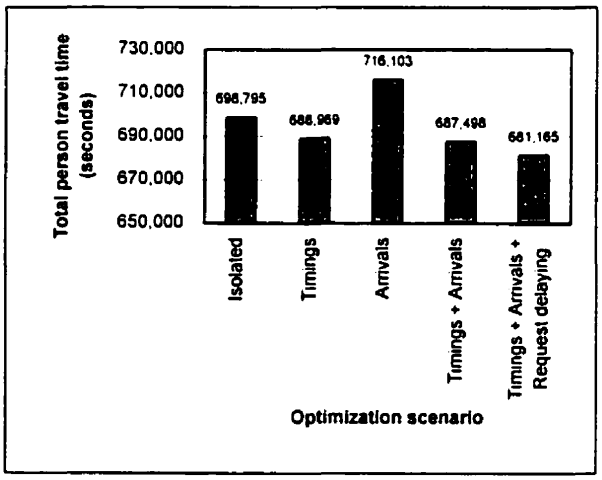


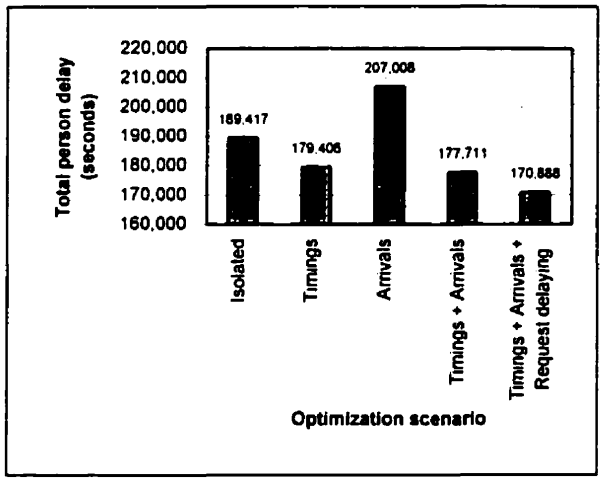


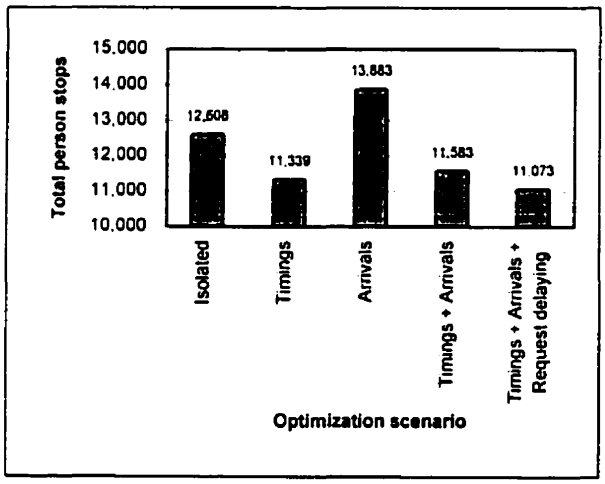


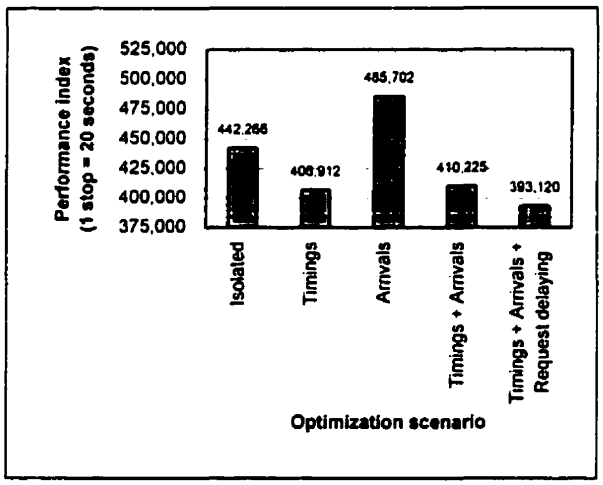


Arterial scenario #7p
(Low peaking demand without transit vehicles)

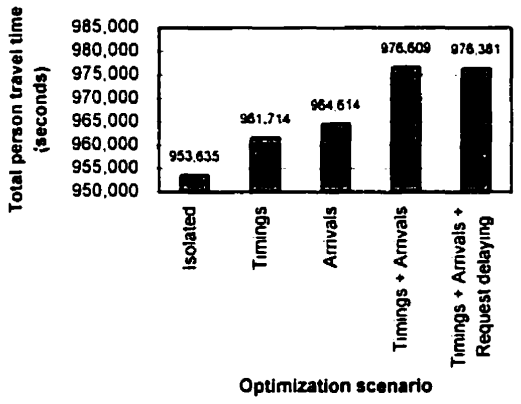


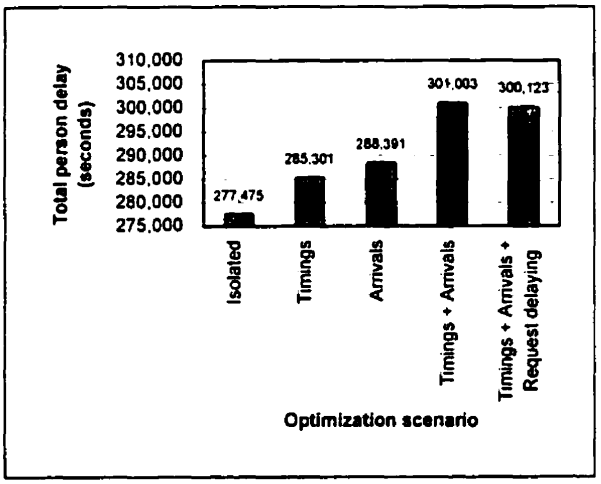


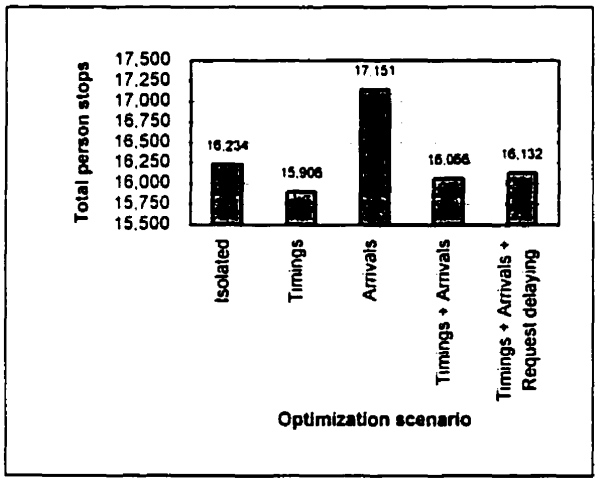


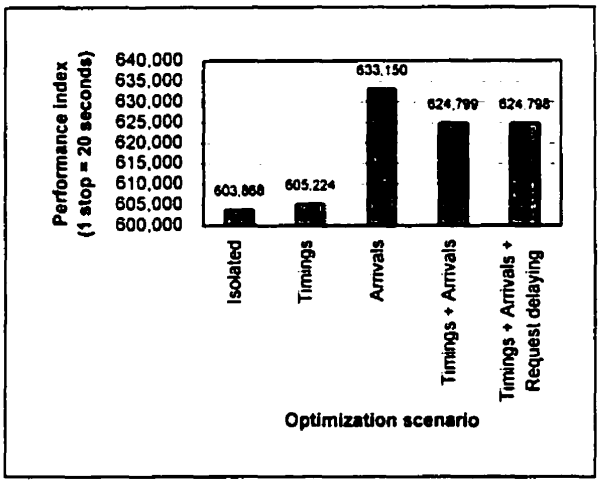


Arterial scenario #8
(Medium constant demand without transit vehicles)

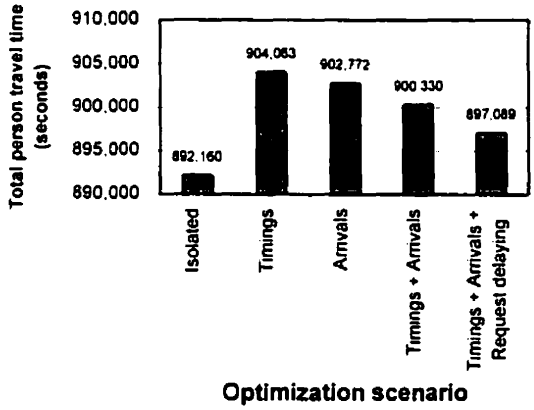


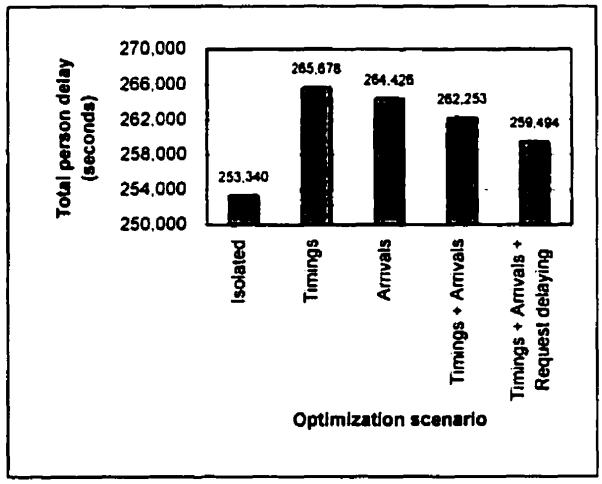


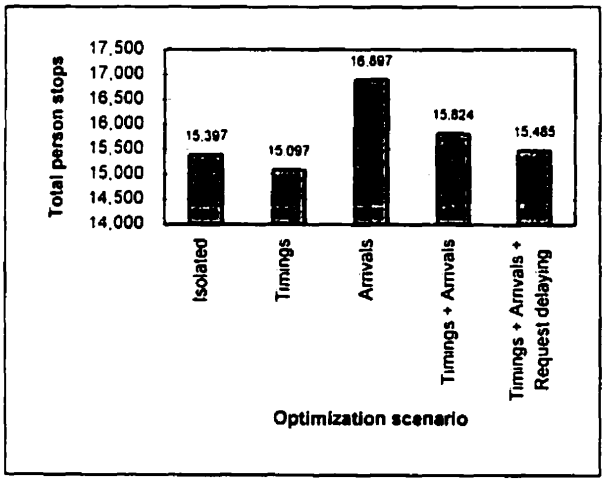


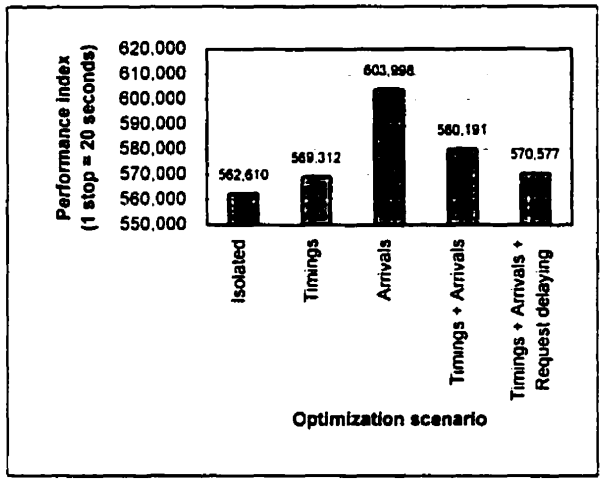


Arterial scenario #8p
(Medium peaking demand without transit vehicles)

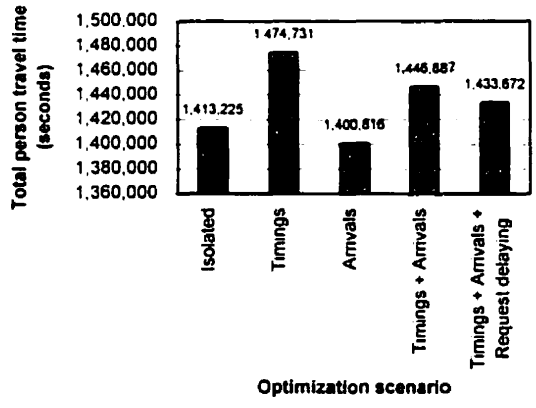


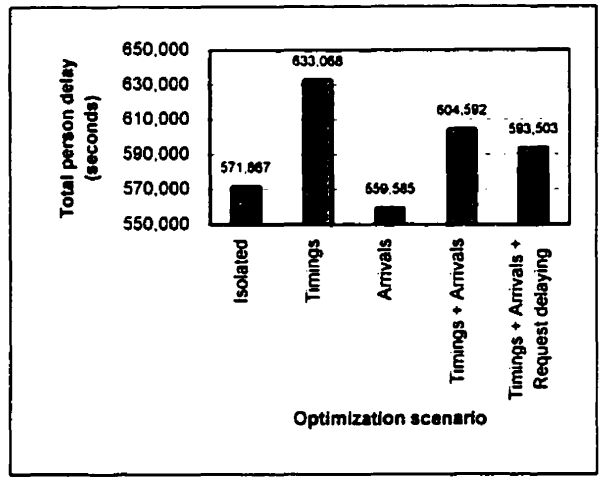


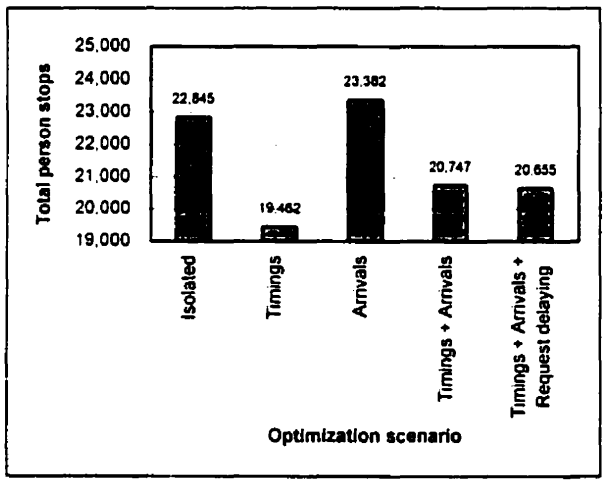


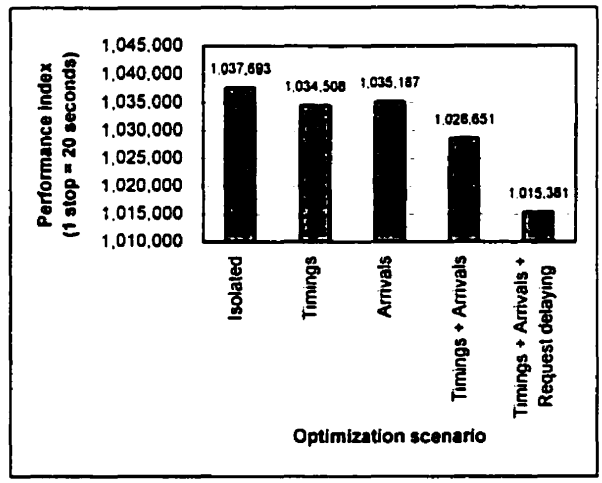


Arterial scenario #9
(High constant demand without transit vehicles)

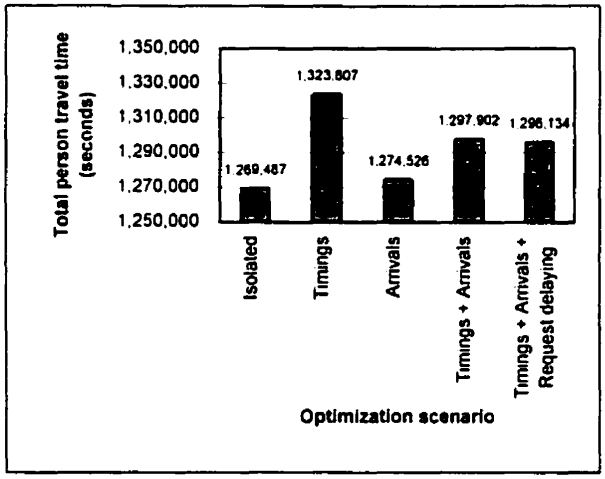


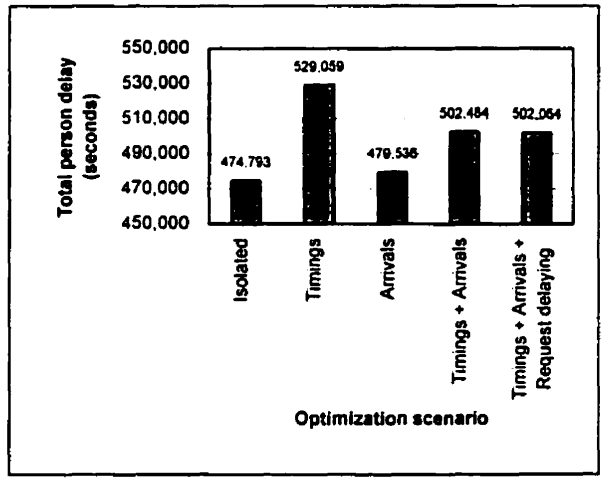


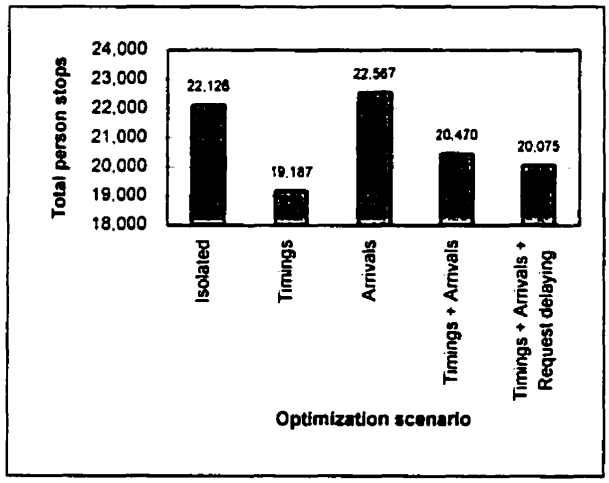


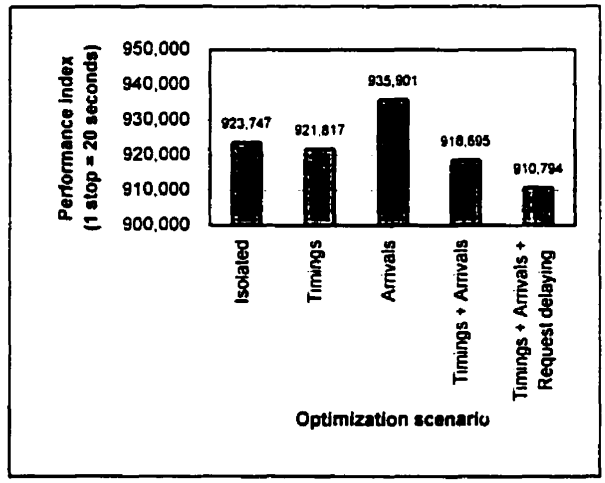


Arterial scenario #9p
(High peaking demand without transit vehicles)

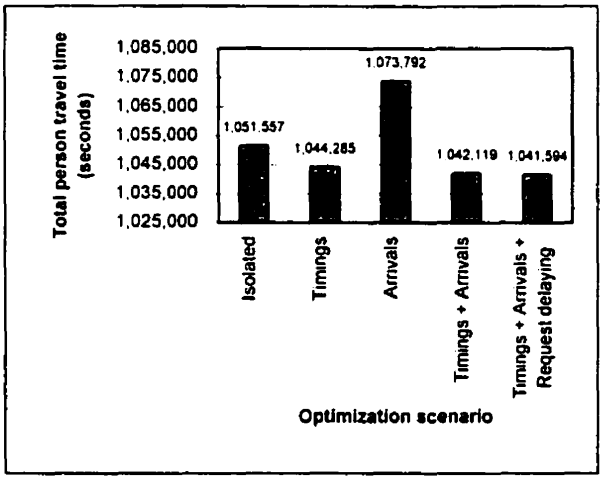


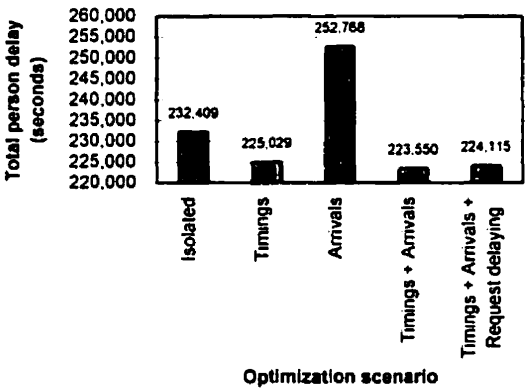


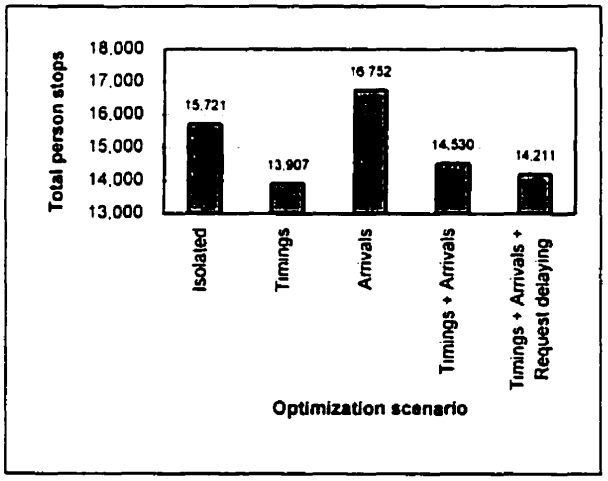


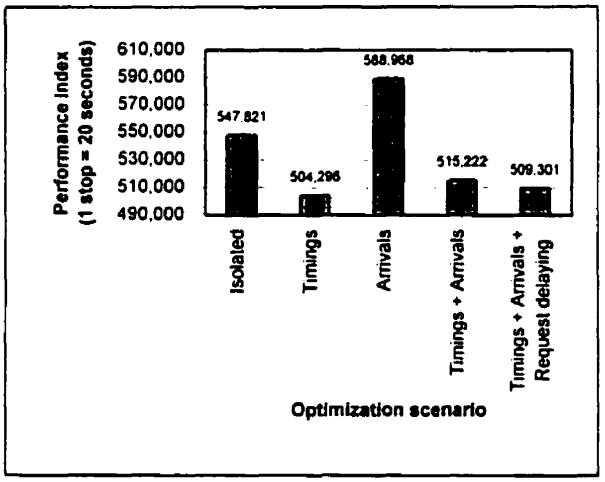


Arterial scenario #10
(Low constant demand with transit vehicles)

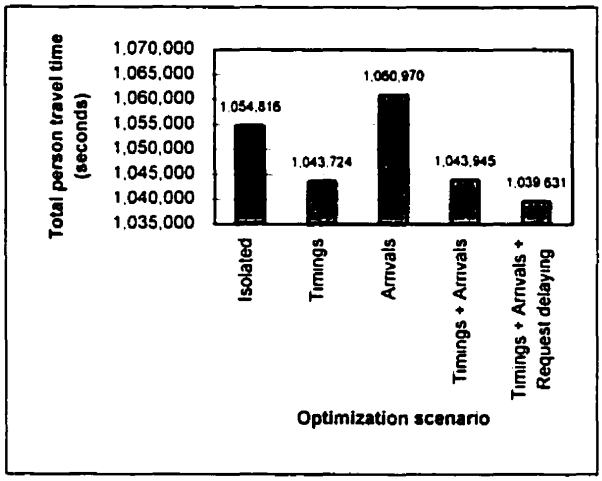


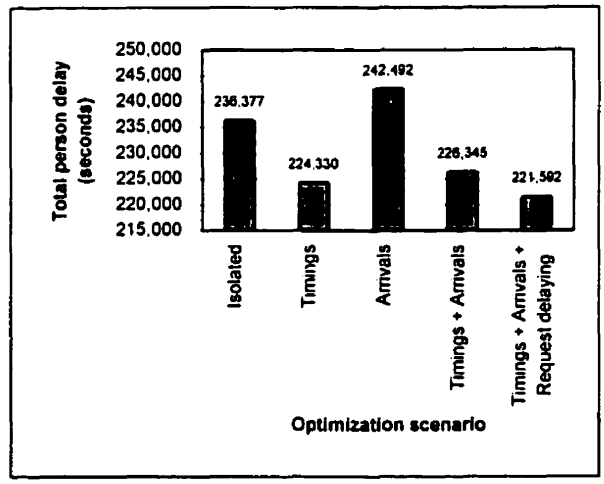


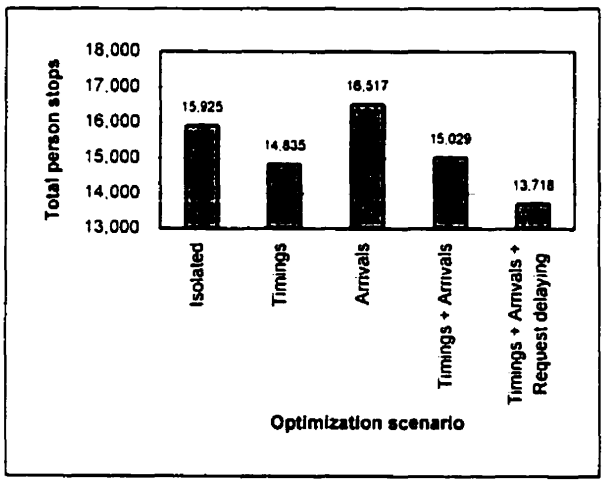


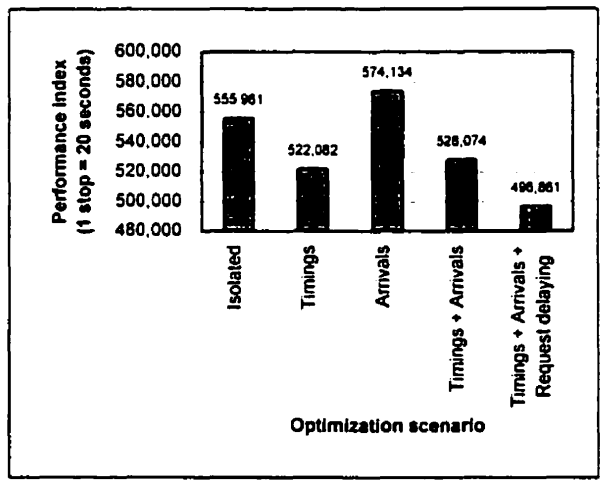


Arterial scenario #10p
(Low peaking demand with transit vehicles)

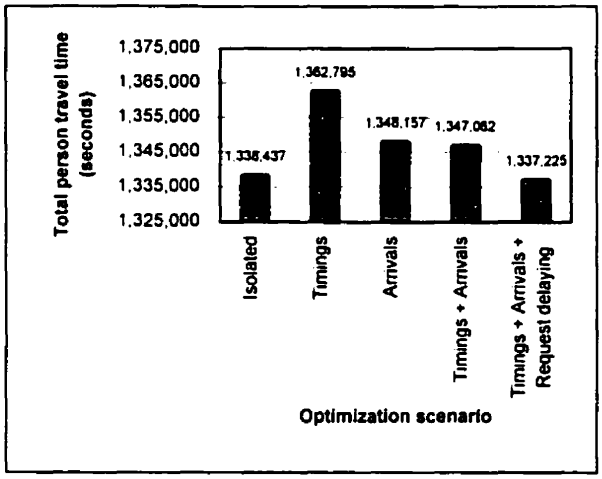


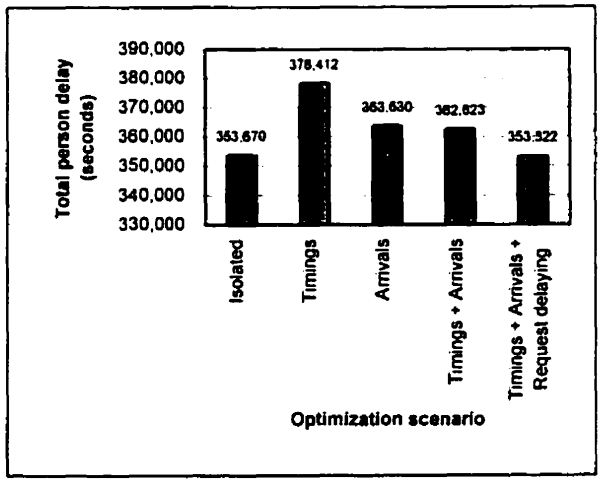


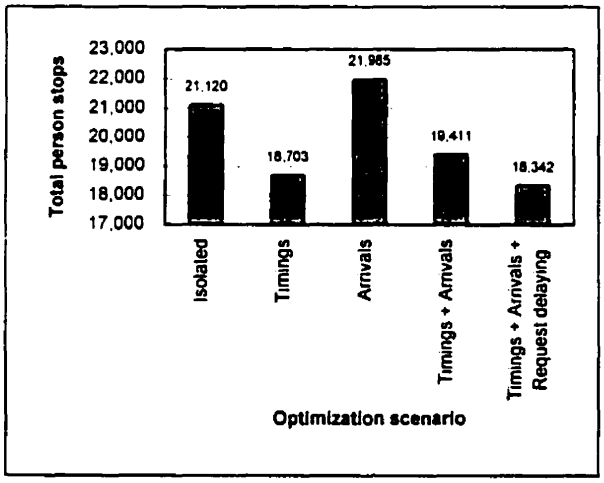


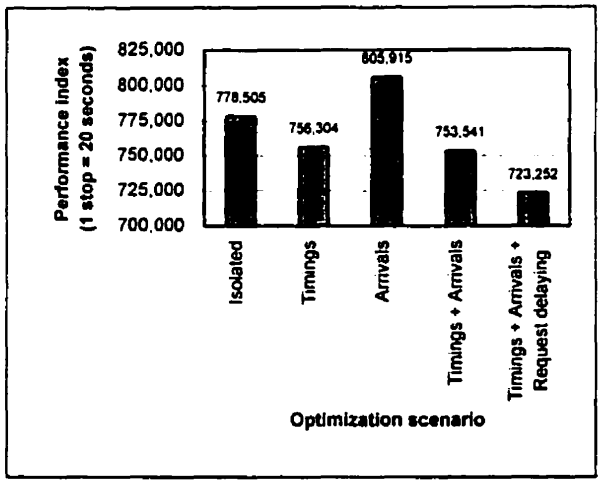


Arterial scenario #11
(Medium constant demand with transit vehicles)

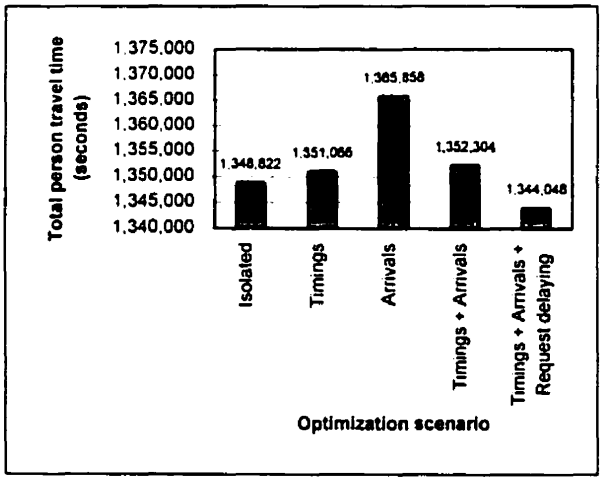


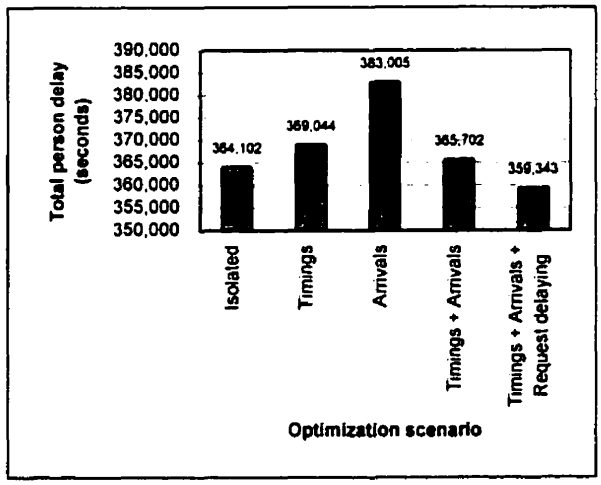


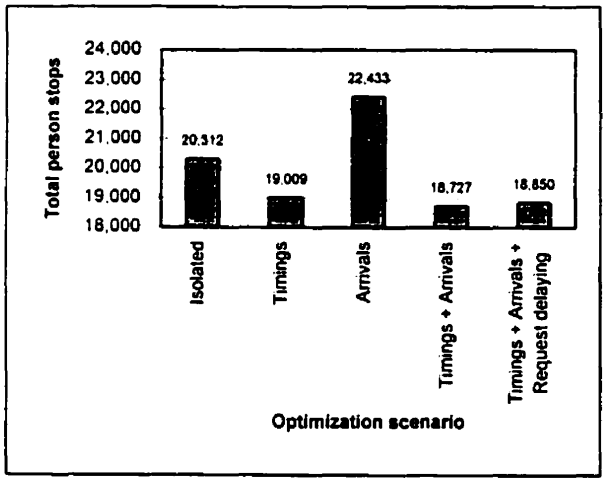


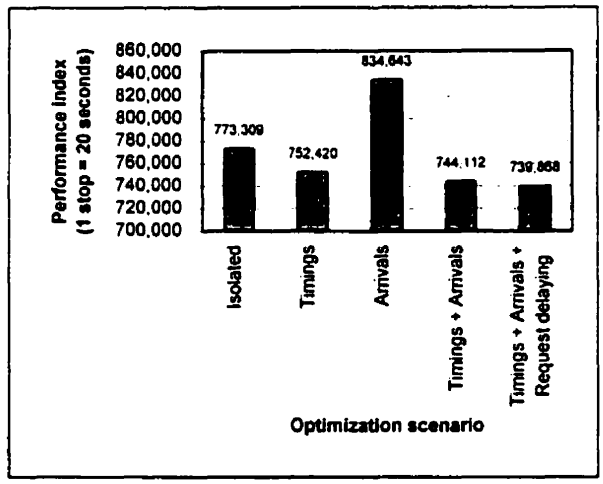


Arterial scenario #11p
(Medium peaking demand with transit vehicles)

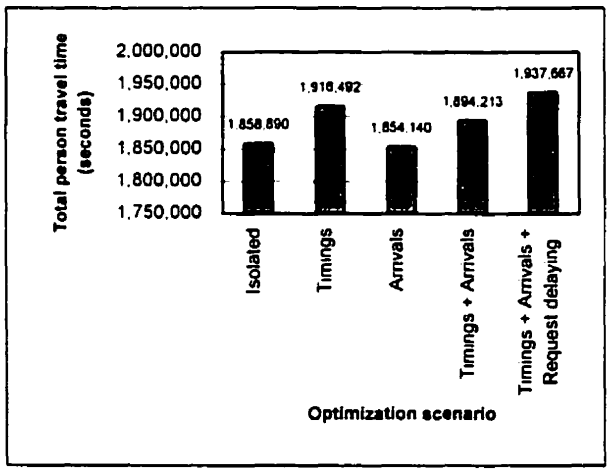


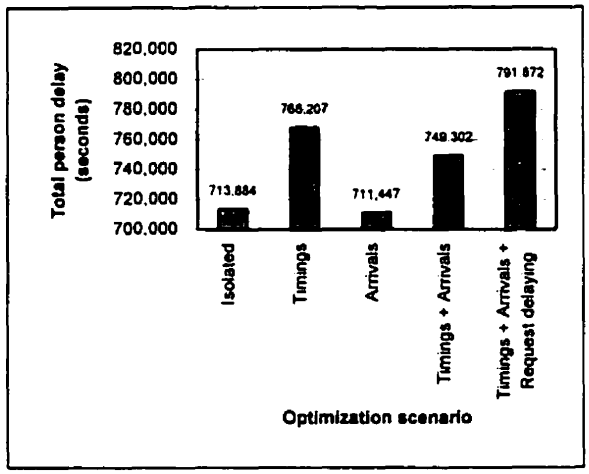


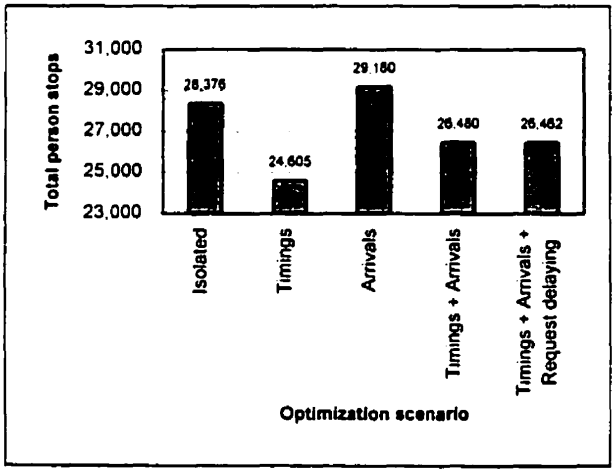


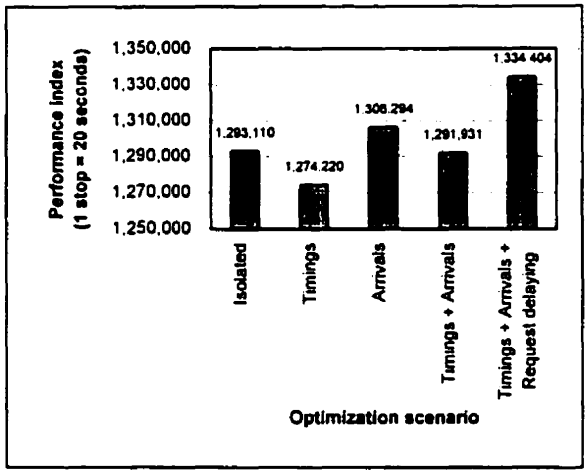


Arterial scenario #12
(High constant demand with transit vehicles)

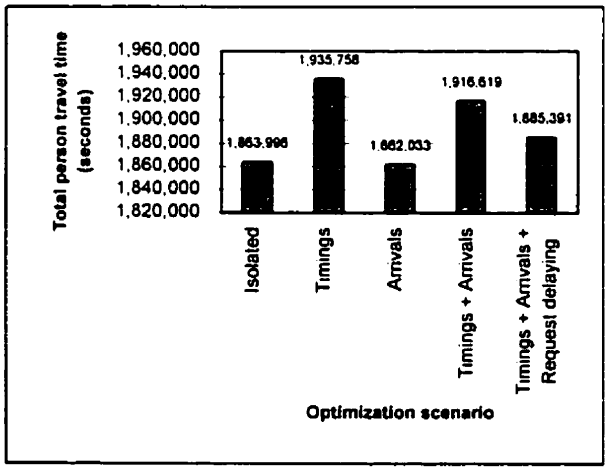


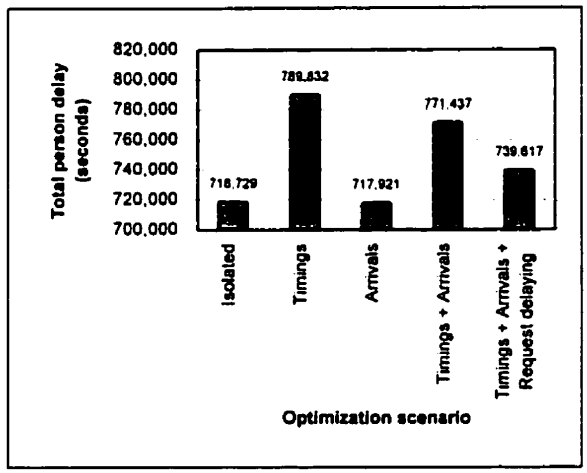


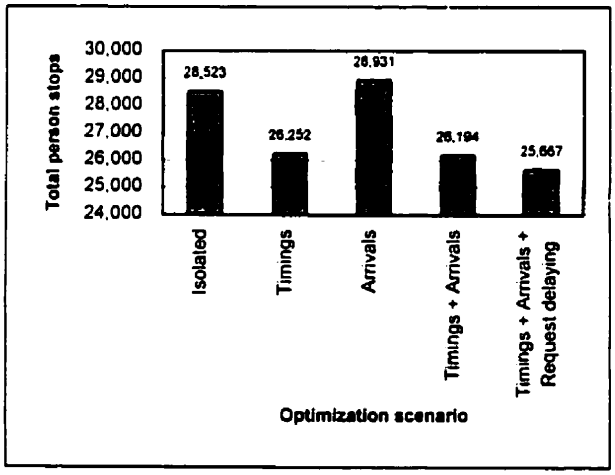


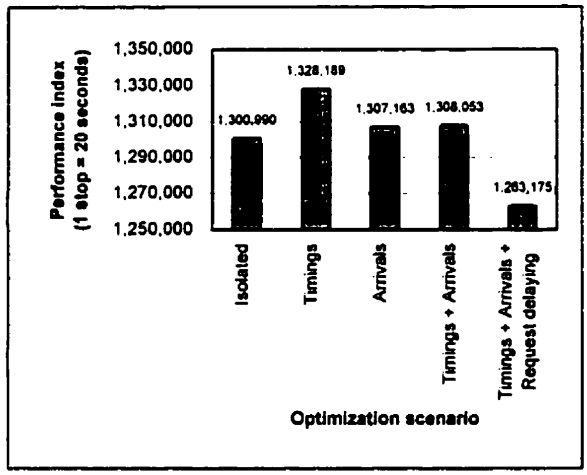


Arterial scenario #12p
(High peaking demand with transit vehicles)





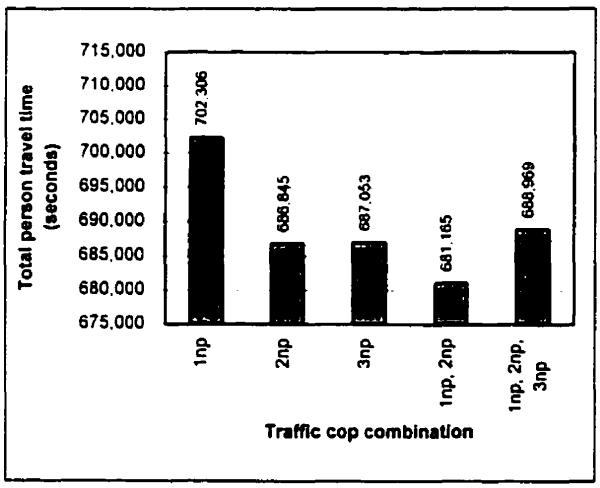


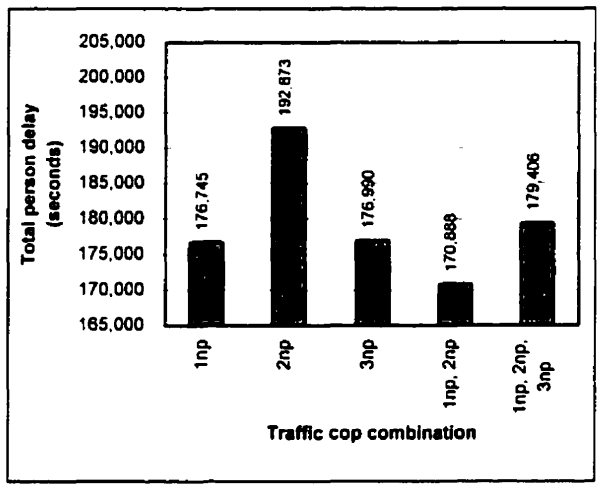


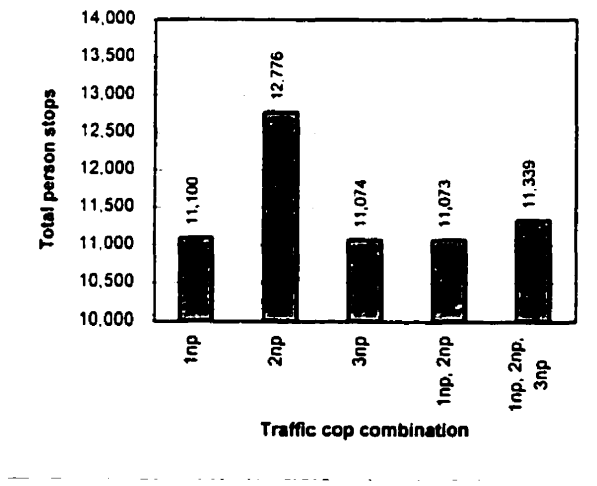
APPENDIX H

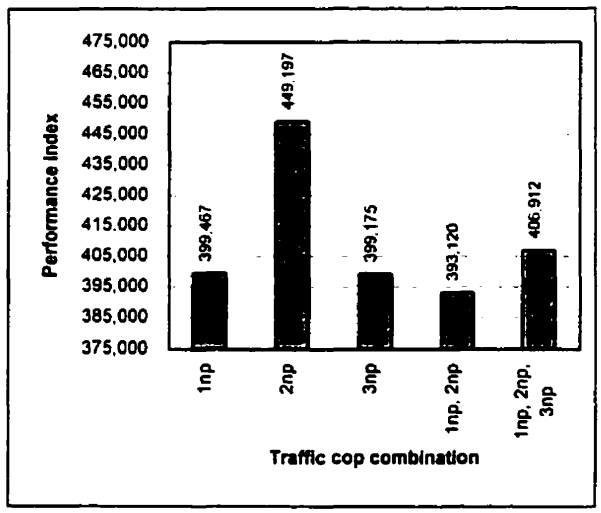
Multi-Objective Optimization Evaluation Results

Arterial Scenarios #7p
(Low peaking demand without transit vehicles)

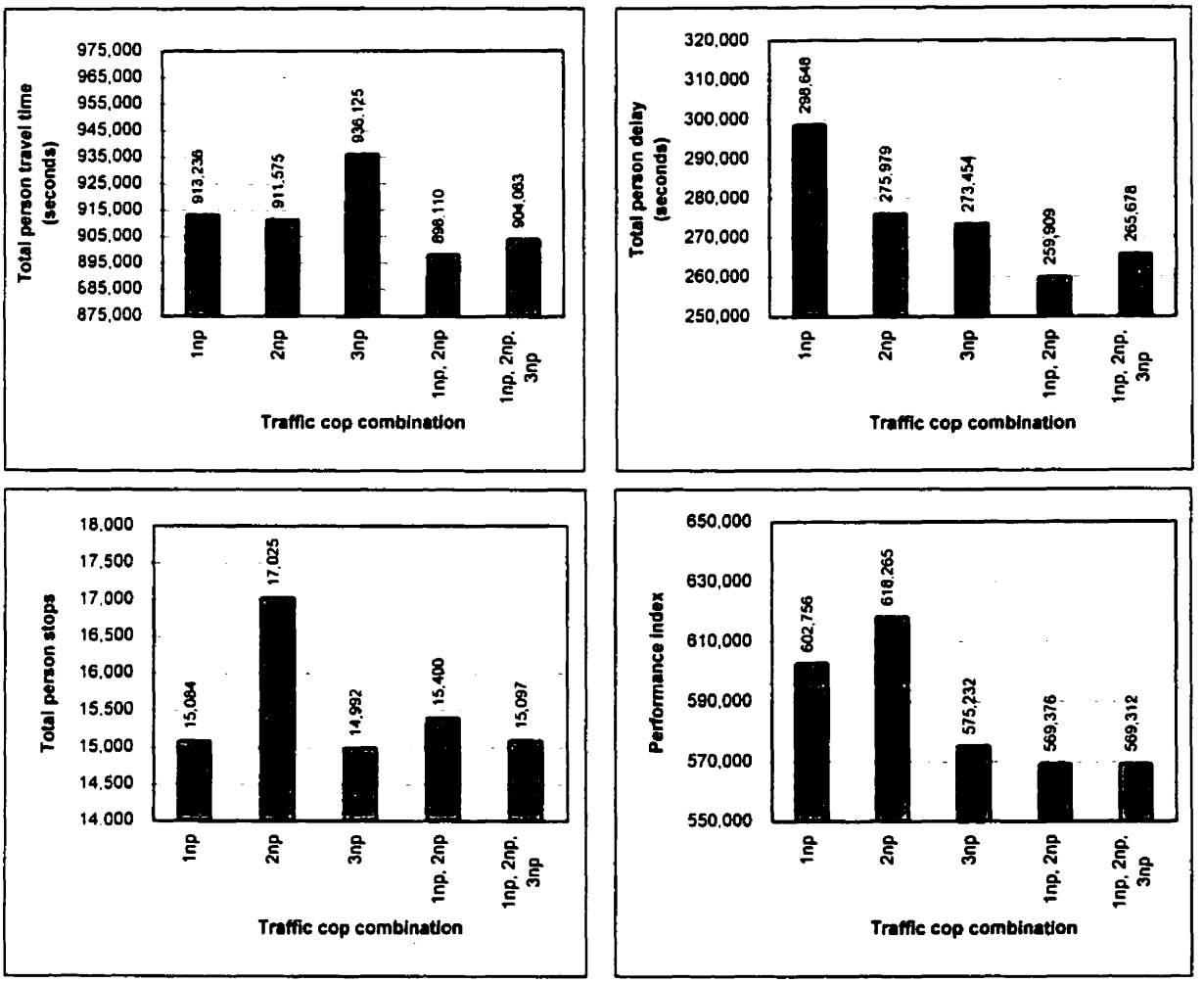




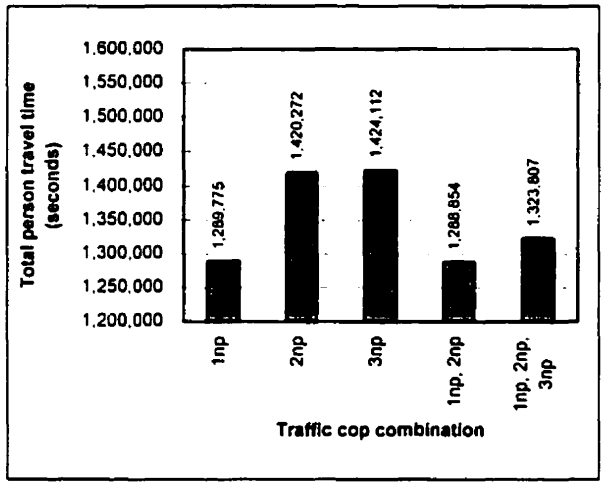


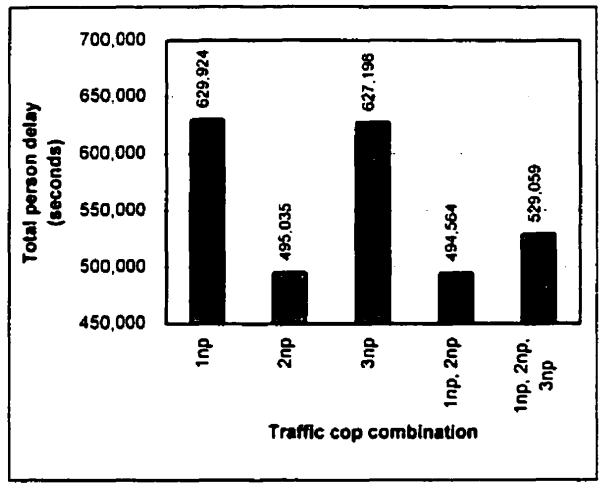


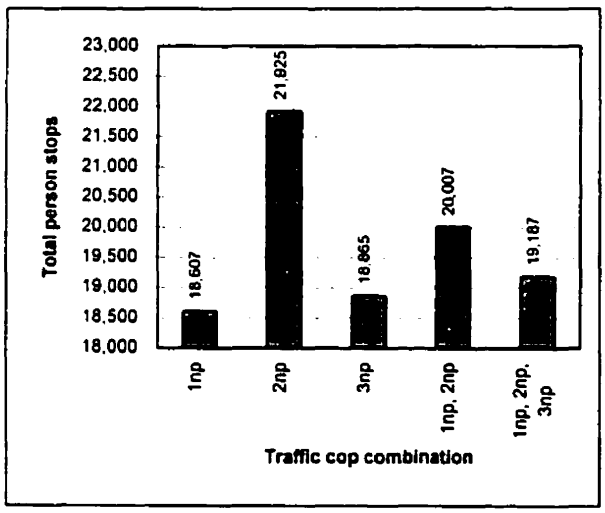
Arterial Scenarios #8p
(Low peaking demand without transit vehicles)

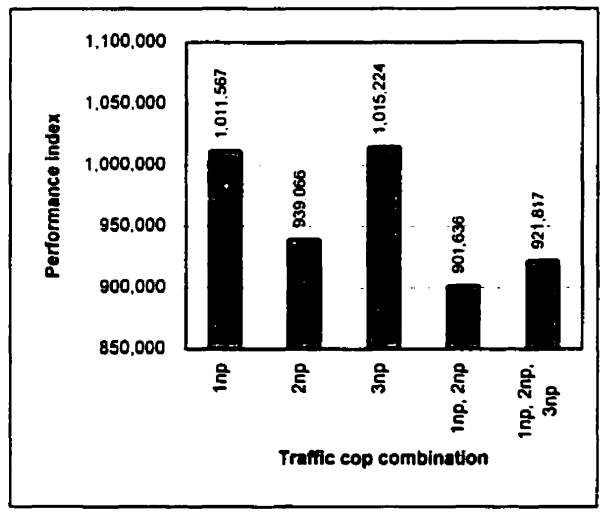


Arterial Scenarios #9p
(Low peaking demand without transit vehicles)

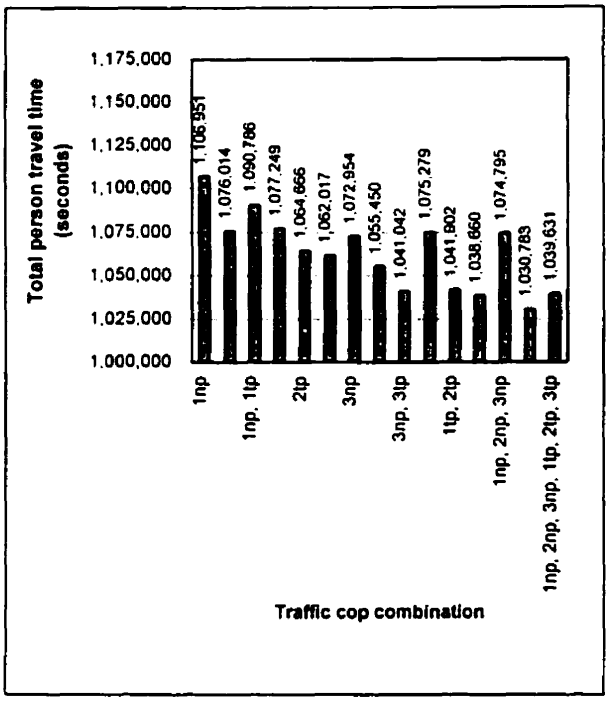


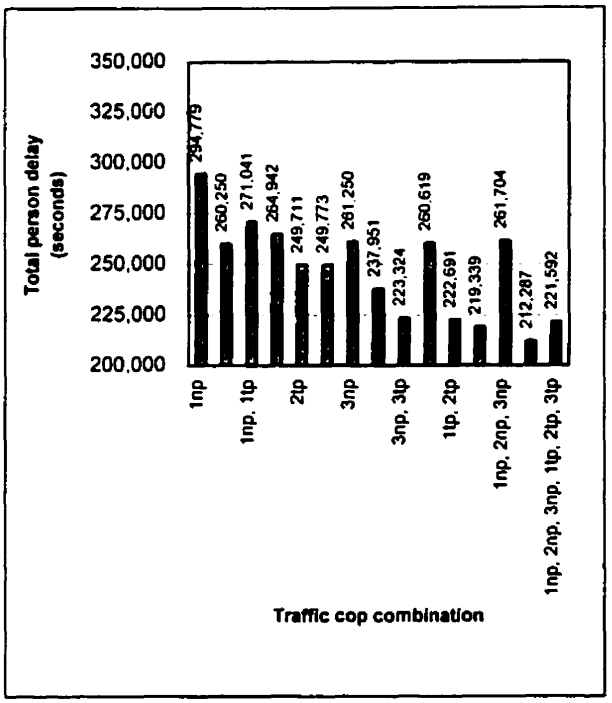


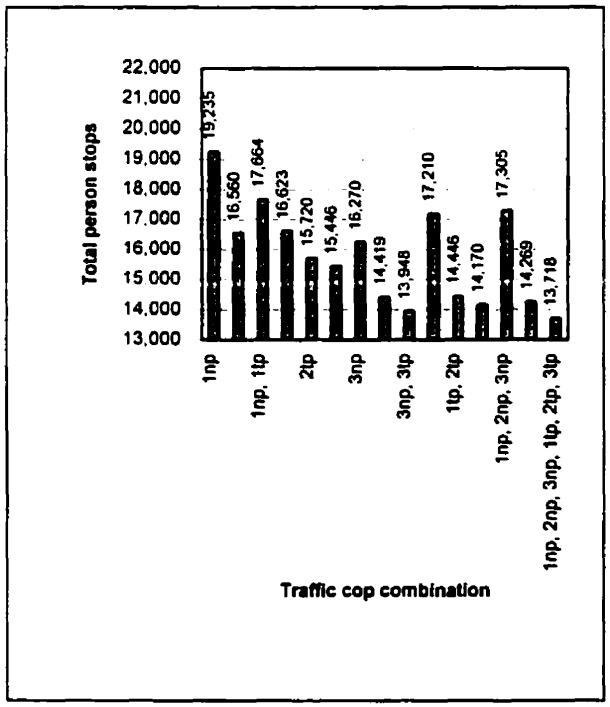


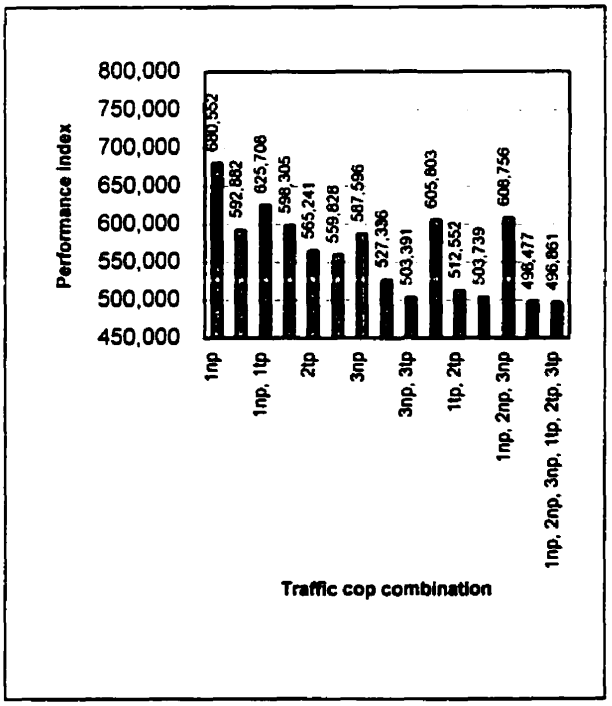


Arterial Scenarios #10p
(Low peaking demand with transit vehicles)

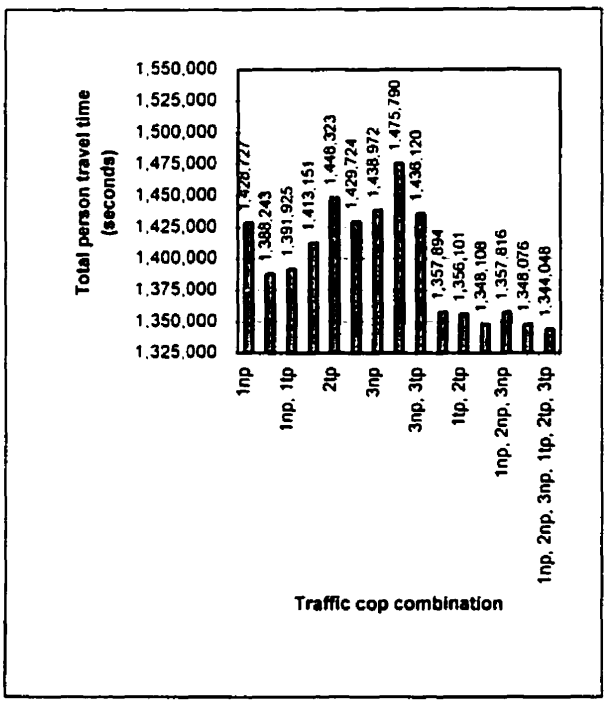


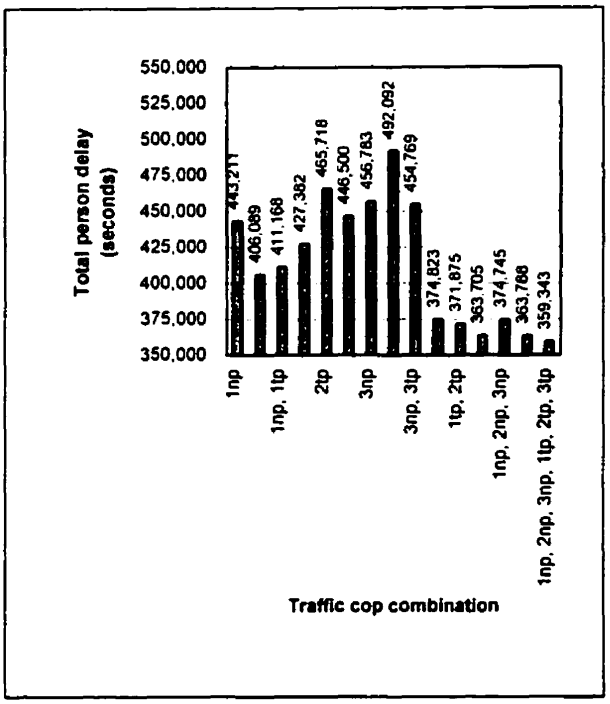


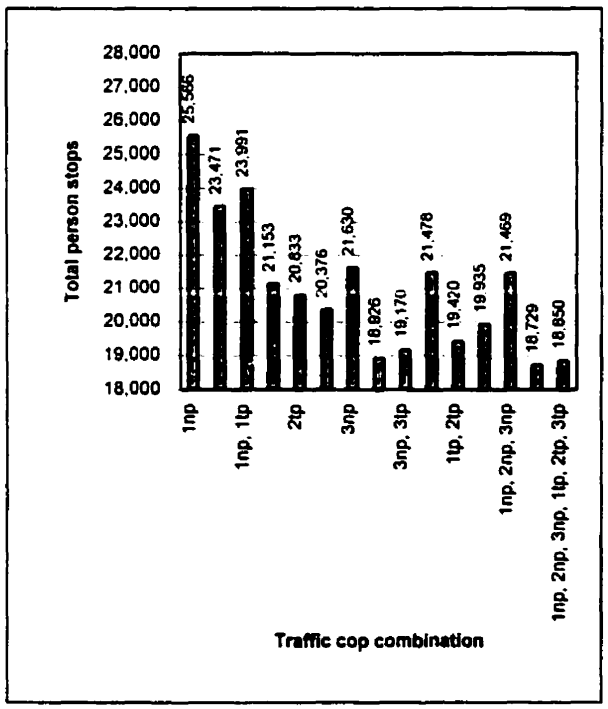


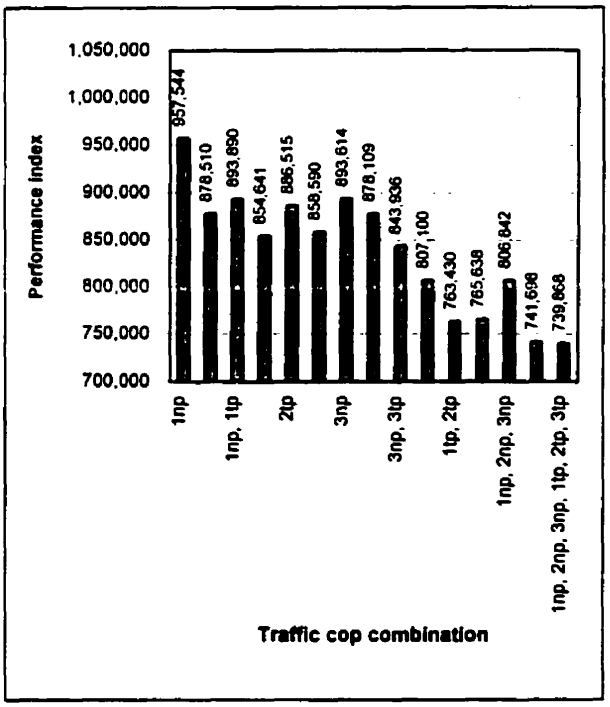


Arterial Scenarios #11p
(Medium peaking demand with transit vehicles)

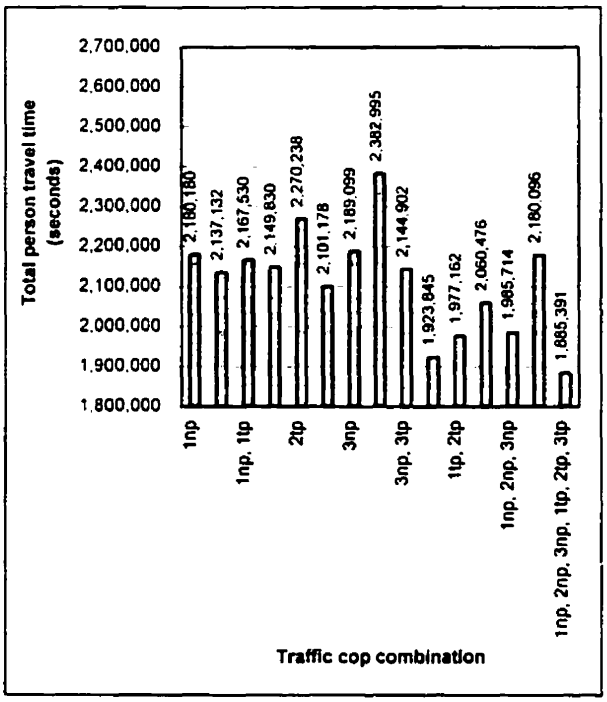


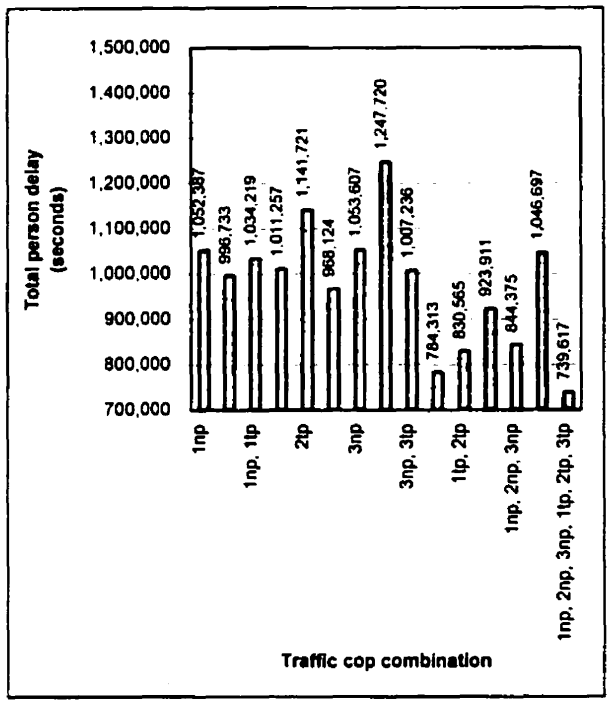


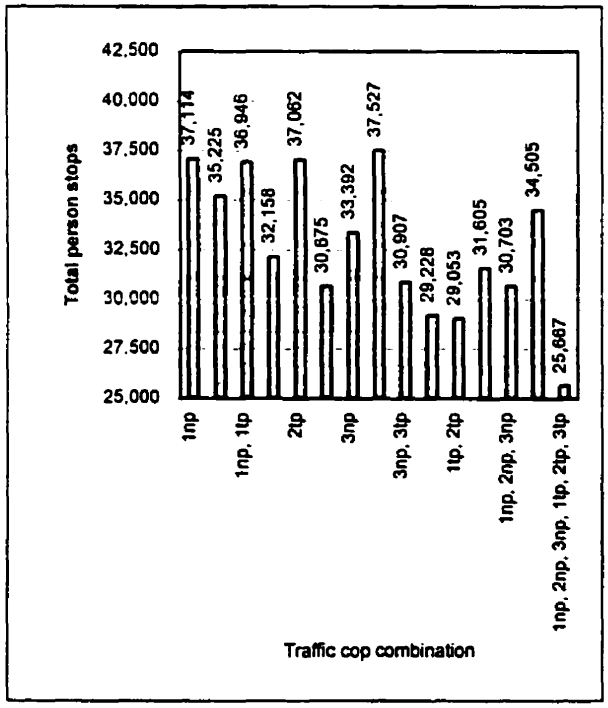


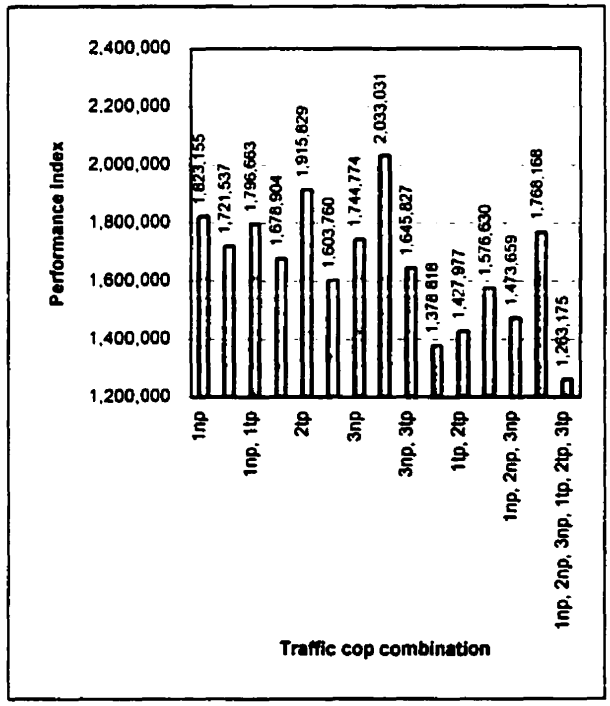


Arterial Scenarios #12p
(High peaking demand with transit vehicles)









References

Al-Sahili, K.A. and W.C. Taylor (1996). "Evaluation of Bus Priority Signal Strategies in Ann Arbor, Michigan." In Transportation Research Record 1554, Transportation Research Board, National Research Council, Washington, D.C., U.S.A., pp. 74-79.

Bang, K.L. (1976). "Optimal Control of Isolated Traffic Signals." In Traffic Engineering and Control, July 1976, pp. 288-292.

Barrière, J.F., J.L. Farges, and J.J. Henry (1986). "Decentralization vs Hierarchy in Optimal Traffic Control." Proceedings, 5th IFAC/IFIPIFORS International Conference on Control in Transportation Systems, Vienna, Austria, July 1986, pp. 321-326.

Bell, M.G.H., M.P.H. Cowell, and B.G. Heydecker (1989). "Traffic-Responsive Signal Control at Isolated Junctions." Proceedings, Engineering Foundation Conference (S. Yagar and E. Rowe Editors), Santa Barbara, California, United States, pp. 273-294.

Benevelli, D.A., A.E. Radwan and J.W. Hurley, Jr. (1983). "Evaluation of a Bus Preemption Strategy by Use of Computer Simulation." In Transportation Research Record 906, Transportation Research Board, National Research Council, Washington, D.C., U.S.A., pp. 60-67.

Bowen, G.T., R. Bretherton, J.R. Landles, and D.J. Cook (1994). "Active Bus Priority in SCOOT." Proceedings, 7th IEE International Conference on Road Traffic Monitoring and Control, Conference Publication No. 391, Institution of Electrical Engineers, London, UK, pp. 73-76.

Bowen, G.T., and R.D. Bretherton (1996). "Latest Developments in SCOOT -- Version 3.1." Proceedings, 8th IEE International Conference on Road Traffic Monitoring and Control, Conference Publication 422, Institution of Electrical Engineerings, London, UK, pp. 61-65.

Bretherton, D. (1996). "SCOOT Current Developments: Version 3." Presented at the 75th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Burton, R., and N. Hounsell (1993). "Bus Priority and UTC Systems - The PROMPT Project." Proceedings, 1993 IEEE/IEE Vehicle Navigation and Information Systems Conference, Ottawa, Canada, pp. 602-605.

Chang, E.C., and C.J. Messer (1990). "PASSER-II 90 User's Guide. Research Report 467-2F, Texas Transportation Institute, Texas A&M University, USA.

Chang, E.C., C.J. Messer and B. Marsden (1985). "Reduced-Delay Optimization and Other Enhancements in the PASSER II-84 Program." In Transportation Research Record 1005, Transportation Research Board, National Research Council, Washington, D.C., U.S.A., pp. 80-89.

Chang, E.C., S.L. Cohen, C.Liu, N.A. Chaudhary, and C. Messer (1989). "MAXBAND-86: Program for Optimizing Left-Turn Phase Sequence in Multiarterial Closed Networks. In Transportation

Research Record 1181, Transportation Research Board, National Research Council, Washington, D.C., USA, pp. 61-67.

Chang, G.L., M. Vadusevan, C.-C. Su (1995). "Bus Preemption under Adaptive Signal Control Environments." Paper presented at the 74th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Chaudhary, N.A., A. Pinnoi, and C.J. Messer (1991). "Proposed Enhancements to MAXBAND 86 Program." In Transportation Research Record 1324, Transportation Research Board, National Research Council, Washington, D.C., USA, pp. 98-104.

Chaudhary, N.A., and C.J. Messer (1993). "PASSER IV: A Program for Optimizing Signal Timings in Grid Networks." In Transportation Research Record 1421, Transportation Research Board, National Research Council, Washington, D.C., USA, pp. 82-93.

Cheng, L., J. Schnablegger and S. Teply (1979). "Transit Priority and Development of a Passive Bus Detection System in Edmonton." In Traffic Engineering and Control, Vol. 20, No. 11, pp 522-525.

Conrad, M. (1997). "Real-Time Traffic Signal Optimization with Transit Priority: An Object-Oriented Approach." Master of Applied Sciences Thesis, Department of Civil Engineering, University of Waterloo, Canada.

Conrad, M, F. Dion and S. Yagar (1998). "Real-Time Traffic Signal Optimization with Transit Priority: Recent Advances in the SPPORT Model." Presented at the 77th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Cornwell, P.R., J.Y.K. Luk, and B.J. Negus (1986). "Tram Priority in SCATS." In Traffic Engineering and Control, Vol. 27, No. 11, pp 561-565.

Davidson, F., and C.D. Di Taranto (1992). "Priority for Public Transport usign Advanced Transport Telematics." Proceedings, 3rd International Conference on Vehicle Navigation and Information Systems, Oslo, Norway, September 1992, pp. 530-536.

De Groot, P.W. (1981). "Advanced Traffic-Responsive Intersection Control Strategies." M.A.Sc. Thesis, Department of Civil Engineering, University of Waterloo.

Dell'Olmo, P., and P.B. Mirchandani (1995). "An Approach for Real-Time Coordination of Traffic Flows on Networks." Presented at the 74th Annual Meeting of the Transportation Research Board, Washington D.C., USA.

Donati, F., V. Mauro, G. Roncolini, and M. Vallauri (1984). "A Hierarchical-Decentralized Traffic Light Control System - The First Realization: Progetto Torino". Proceedings, 9th IFAC Triennal World Congress, Budapest, Hungary, pp. 2853-2858.

El-Reedy, T.Y. and R. Ashworth (1978). "The Effect of Bus Detection on the Performance of a Traffic Signal Controlled Intersection." In Transportation Research, Vol. 12, pp. 337-342.

Evans, H.K. and G.W. Skiles (1970). "Improving Public Transit through Bus Preemption of Traffic Signals." In Traffic Quartely, Vol. 24, No. 4, October 1970, pp. 531-543.

Farges, J.L., and Henry, J.J. (1988). "Réalisation et test d'une méthode de programmation dynamique des feux de circulation PRODYN" [in French]. Intermediary Report S.E.R.T./DERA No 1/7540, Centre d'Études et de Recherches de Toulouse, Toulouse, France.

Federal Highway Administration (1988). "TRANSYT-7F User's Manual - Release 6." Federal Highway Administration, U.S. Department of Transportation, Washington, D.C., USA.

Federal Highway Administration (1998). "Traffic Software Integrated System Version 4.2 - Corsim Run-Time Extension." Federal Highway Administration, U.S. Department of Transportation, Washington, D.C., USA.

Fehon, B.E., and S.E. Moore (1982). "Dynamic Control of a Medium Sized Traffic Signal Network. Proceedings." 11th ARRB Conference, Vol. 11, Part 4, Australian Road Research Board, Melbourne, Australia, August 1982, pp. 85-93.

Gartner, N.H. (1982a). "Prescription for Demand-Reponsive Urban Traffic Control." In Transportation Research Record 881, Transportation Research Board, National Research Council, Washington, D.C., USA, pp. 73-76.

Gartner, N.H. (1982b). "Demand-Responsive Decentralized Urban Traffic Control, Part I: Single Intersection Policies." Report DOT/RSPA/DPB-50/81/24, U.S. Department of Transportation, Washington D.C., U.S.A.

Gartner, N.H., M.H. Kaltenbach, and M.M. Miyamoto (1983). "Demand-Responsive Decentralized Urban Traffic Control, Part II: Network Extensions." Report DOT/OST/P-34/85/009, U.S. Department of Transportation, Washington D.C., U.S.A.

Gartner, N.H. (1983). "OPAC: A Demand Responsive Strategy for Traffic Signal Control." In Transportation Research Record 906, Transportation Research Board, National Research Council, Washington, D.C., USA, pp. 75-81.

Gartner, N.H. (1989). "OPAC: Strategy for Demand-Reponsive Decentralized Traffic Signal Control." Proceedings, 6th IFAC/IFIP/IFORS Symposium on Control, Computers, Communication on Transportation, Paris, September 1989, pp. 241-244.

Gartner, N.H., P.J. Tarnoff, and C.M. Andrews (1990). "Evaluation of Optimized Policies for Adaptive Control Strategy." In Transportation Research Record 1324, Transportation Research Board, National Research Council, Washington D.C., U.S.A., pp. 105-114.

Gartner, N.H., S.F. Assmann, F. Lasaga, and D.L. Hou (1990). "MULTIBAND -- A Variable-Bandwidth Arterial Progression Scheme." In Transportation Research Record 1287, Transportation Research Board, National Research Council, Washington, D.C., USA, pp. 212-222.

Gartner, N.H., S.F. Assmann, F. Lasaga and D.H. Hou (1991). "A Multi-Band Approach to Arterial Traffic Signal Optimization." In Transportation Research B, Vol. 25B, No. 1, pp. 55-74.

Gartner, N.H., C. Stamatiadis, and P.J. Tarnoff (1995). "Development of Advanced Traffic Signal Control Strategies for IVHS: A Multi-Level Design." Presented at the 74th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Hadi, A.H. and C.E. Wallace (1992). "Progression-Based Optimization Model in TRANSYT-7F." In Transportation Research Record 1360, Transportation Research Board, National Research Council, Washington, D.C., U.S.A., pp. 74-81.

Han, B. and S. Yagar (1991). "Real-Time Control for Transit Mixed with Traffic." Proceedings, 35th Annual PTRC Meeting, University of Sussex, England.

Han, B. and S. Yagar (1992). "Real-Time Control of Traffic with Bus and Streetcar Interactions." Proceedings, 6th IEE International Conference on Road Traffic Monitoring and Control, Conference Publication No. 355, Institution of Electrical Engineers, London, UK, pp. 108-112.

Head, K.L., P.B. Mirchandani, and D. Sheppard (1992). "Hierarchical Framework for Real-Time Traffic Control." In Transportation Research Record 1360, Transportation Research Board, National Research Council, Washington, D.C., U.S.A., pp. 82-88.

Head, K.L. (1995). "An Event-Based Short Term Traffic Flow Prediction Model." Presented at the 74th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Henry, J.J, J.L. Farges, and J. Tuffal (1983). "The PRODYN Real Time Traffic Algorithm." Proceedings, 4th IFAC/IFIP/IFORS International Conference on Control in Transportation Systems, Baden-Baden, Germany, pp. 305-310.

Henry, J.J, and J.L. Farges (1984). "Commande des feux en temps réel PRODYN" [in French]. In Transport Environnement et Control, No 64, May-June 1984, pp. 31-36.

Henry, J.J, and J.L. Farges (1989). "PRODYN." Proceedings, 6th IFOC/IFIP/IFORS Symposium on Control Computers, Communications in Transportation, Paris, September 1989, pp. 253- 255.

Heydecker, B.G. (1990). "A Continuous-Time Formulation for Traffic-Responsive Signal Control. Presented at the 11th International Symposium on Transportation and Traffic Theory, Yokohama, Japan.

Hounsell, N.B., F.N. McLeod, R.D. Bretherton, and G.T. Bowen (1996). "PROMPT: Field Trail and Simulation Results of Bus Priority in SCOOT." Proceedings, 8th IEE International Conference on Road Traffic Monitoring and Control, Conference Publication 422, Institution of Electrical Engineerings, London, UK, pp. 90-94.

Hunt, P.B., D.I. Robertson, R.D. Bretherton, and R.I. Winton (1981). "SCOOT - A Traffic Responsive Method of Coordinating Signals." Report LR1014, Transportation and Road Research Laboratory, Crowthorne, United Kingdom.

ITE (1995). "Canadian Capacity Guide for Signalized Intersections, Second Edition." S. Teply Editor, D.I. Allingham, D.B. Richardson and B.W. Stephenson, Institute of Transportation Engineers, District 7, Canada.

Jacques, M.A.P. (1993). "Modeling Transit Effects in Signalized Networks." Ph.D. Thesis, Department of Civil Engineering, University of Waterloo.

Kelman, W.L., J.C. Greenough, and B.Y. Quan (1993). "A Performance Report on the Metropolitan Toronto SCOOT System: An Advanced Traffic Adaptive Control System. Metropolitan Toronto Department of Transportation, Toronto.

Kessaci, A. (1988). "Estimation en ligne et gestion des capacités pour la commande du traffic urbain" [in French]. Ph.D. Thesis, École Nationale Superieure de l'Aéronautique et de l'Espace, France.

Kessaci, A., J.L. Farges J.L., and J.J. Henry (1989). "Upper Level for Real-TIme Urban Traffic Control Systems." Proceedings, 11th IFAC World Congress, Tallin, Estonia.

KLD Associates, Inc. (1994a). "Traffic Models for Testing Real-Time Traffic Adaptive Signal Control Systems - Software Modifications for Interfacing NETSIM with Real-Time Traffic Adaptive Control Systems." Working Paper, KLD Associates Inc., Huntington Station, NY, U.S.A.

KLD Associates, Inc. (1994b). "Traffic Models for Testing Real-Time Traffic Adaptive Signal Control Systems - Software Modifications to Support Testing of Real-Time Traffic Adaptive Control Systems." Working Paper, KLD Associates Inc., Huntington Station, NY, U.S.A.

Law, A. and W.D. Kelton (1991). "Simulation Modeling and Analysis." Second Edition, McGraw-Hill.

Lialias, J.A. (1998). "AUSCI goes Solo - Advanced Video Detection for Minneapolis." In Traffic Technology International, Dec 97/Jan 98, pp. 36-38.

Lin, F.B., N. Wang and S. Vijayakumar (1987). "Development of an Intelligent Adpative Signal Logic." Proceedings, Engineering Foundation Conference (S. Yagar Editor), Henniker, New Hampshire, pp. 257-279.

Little, D.C., M.D. Kelson, and N.H. Gartner (1981). "MAXBAND: A Program for Setting Signals on Arteries and Triangular Networks." In Transportation Research Record 795, Transportation Research Board, National Research Council, Washington, D.C., USA, pp. 40-46.

Lowrie, P.R. (1982). "The Sydney Co-Ordinated Adative Traffic System - Principles, Methodology, Algorithms." Proceedings, International Conference on Road Traffic Signaling, Conference Publication No. 207, Institution of Electrical Engineers, London, United Kingdom, pp. 67-70.

Lowrie, P.R. (1990). "SCATS - Sydney Co-Ordinated Adaptive Traffic System." Traffic Control Section, Road and Traffic Autority of North South Wales, Sydney, Australia.

Luk, J.Y.K. (1984). "Two Traffic-responsive Area Traffic Control Methods: SCAT and SCOOT." In Traffic Engineering and Control, January 1984, pp.14-18.

Luk, J.Y.K., and L.W. Cahill (1986). "On-Line Estimation of Platoon Delay." Proceedings, 13th ARRB/5th REAAA Conference, Part 7, Australian Road Research Board, Australia, pp. 133-148.

Luk, J.Y.K., M. Besley, and B.W. Doughty (1988). "The Box Hill Study - Queue Length Estimation in SCATS." Proceedings, 14th ARRB Conference, Part 2, Australian Road Research Board, Australia, pp. 157-172.

MacGowan, J., and I.J. Fullerton (1979-1980). "Development and Testing of Advanced Control Strategies in the Urban Traffic Control System" (3 articles). in Public Roads, Vol. 43., Nos. 2-4, September 1979-March 1980, pp. 41-52 (No. 2), 97-105 (No. 3), and 159-165 (No. 4).

Malakapali, M.P. and C.J. Messer (1993). "Enhancements to the PASSER II-90 Delay Estimation Procedures." In Transportation Research Record 1421, Transportation Research Board, National Research Council, Washington, D.C., USA, pp. 94-103.

Martin, P.T., and S.L.M. Hockaday (1995). "SCOOT - An Update." In ITE Journal, January 1995, pp. 44-48.

Mauro, V., and C. Di Taranto (1989). "UTOPIA." Proceedings, 6th IFAC/IFIP/IFORS Symposium on Control, Computers, Communication on Transportation, Paris, September 1989, pp. 245-252.

Mauro, V. (1991). "Road Network Control." In Concise Encyclopedia of Traffic & Transportation Systems, Pergamon Press, Oxford, pp. 361-366.

Messer, C.J., R.H. Wilson, C.L. Dudeck, and E.J. Romano (1974). "A variable-Sequence Multiphase Progression Optimization Program." in Highway Research Record 445, Highway Research Board, National Research Council, Washington, D.C., USA, pp. 24-33.

McShane, W.R., and K. Crowley (1976). "Regularity of Signal Detectors - Observed Arterial Traffic Volume Characteristics." In Transportation Research Record 596, Transportation Research Board, National Research Council, Washington, D.C., USA.

McShane, W.R., and R.P. Roess (1990). "Traffic Engineering." Prentice Hall, Englewoods Cliffs, New Jersey, USA.

McShane, W.R., R.P. Roess and E.S. Prassas (1998). "Traffic Engineering." Second Edition, Prentice Hall, Upper Saddle River, New Jersey, USA.

Metropolitan Toronto (1994). "SCOOT Demonstration Project - An Evaluation of Traffic Adaptive Control Using SCOOT." Evaluation Report, Metropolitan Toronto.

Metro Toronto Transportation, Toronto Transit Commission, Ministry of Transportation of Ontario (1996). "Mainline Traffic Signal Priority Study - Phase IV: Bus Priority Demonstration." Final Draft Report.

Michalopoulos, P.G., R.D. Jacobson, C.A. Anderson and J.C. Barbaresso (1993). "Integration of Machine Vision and Adpative Control in the FAST-TRAC Intelligent Vehicle Highway System Program." In Transportation Research Record 1408, Transportation Research Board, National Research Council, Washington, D.C., USA, pp. 108-115.

Middleman, F., H. Taale, J. Dibbits, and W. Fransen (1996). "The Assessment of the SCOOT System in Nijmegen". Proceedings, 8th International Conference on Road Traffic Monitoring and Control, Conference Publication 422, Institution of Electrical Engineerings, London, UK, pp. 66-70.

Miller, A.J. (1963). "A Computer Control System for Traffic Network." Proceedings, Second International Symposium on the Theory of Road Traffic Flow, Organisation for Economic Cooperation and Development, Paris, pp. 200-220.

NCHRP (1992). "Signal Timing Improvement Practice." Synthesis of Highway Practice 172, National Cooperative Highway Research Program, Transportation Research Board, National Research Council, Washington, D.C., USA.

Owen, L.E., C.M. Stallard, D.M. Curtis (1997). "An Evaluation of Real-Time Traffic Adaptive Signal Control Prototypes." Presented at the 76th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Peterson, A. (1994). "AUT - A Dynamic Tool for Traffic Control and Supervision." Proceedings, 7th IEE International Conference on Road Traffic Monitoring and Control, Conference Publication No. 391, Institution of Electrical Engineers, London, United Kingdom, pp. 64-67.

Richardson, A.J. and K.G. Ogden (1979). "Evaluation of Active Bus Priority Signals." In Transportation Research Record 718, Transportation Research Board, National Research Council, Washington, D.C., U.S.A., pp. 5-12.

Robertson, D.I. (1967). "TRANSYT: A Network Study Tool." Report LR 253, Road Research Laboratory, Crowthorne, England.

Robertson, D.I. (1986). "Research on the TRANSYT and SCOOT Method of Signal Coordination." In ITE Journal, January 1986, pp. 36-40.

Shriber, T.J. (1974). "Simulation using GPSS." John Wiley and Sons, Toronto.

Sims, A.G. (1978). "The Sydney Co-ordinated Adaptive Traffic System." Proceedinds, Workshop on Area Traffic Control, 9th ARRB Conference, Australian Road Research Board, Australia, pp. 12-27.

Sims, A.G., and K.W. Dobbinson (1979). "SCAT - The Sydney Co-Ordinated Adaptive Traffic System - Phylosophy and Benefits." Proceedings, International Symposium on Traffic Control Systems, Vol. 2B, University of California, Berkeley, USA, pp. 19-41.

Sims, A.G. and A.B. Finlay (1984). "SCATS – Splits and Offsets Simplified (S.O.S.)" Proceedings, 12th ARRB Conference, Part 4, Australian Road Research Board, Australia, pp. 17-33.

Sunkari, R.S., P.S. Beasley, T. Urbanick, and D.B. Fambro (1995). "Model to Evaluate the Impacts of Bus Priority to Signalized Intersections." Paper presented at the 74th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

TRB (1985). "Highway Capacity Manual." Special Report 209, Second Edition, Transportation Research Board. National Research Council. Washington, D.C., USA

TRB (1994). "Highway Capacity Manual." Special Report 209, Third Edition, Transportation Research Board, National Research Council, Washington, D.C., USA

Vincent, R.A., and C.P. Young (1986). "Self-Optimising Traffic Signal Control using Microprocessors - The TRRL MOVA Strategy for Isolated Intersections." In Traffic Engineering and Control, July/August 1986, pp. 385-387.

Vincent, R.A., B.R. Cooper and K. Wood (1978). "Bus Actuated Signal Control at Isolated Intersections: Simulation Studies of Bus Priority." Report 814, Transport and Road Research Laboratory, Crowthorne, Berkshire, United Kingdom.

Vincent, R.A., and C.P. Young (1988). "MOVA: Traffic-Responsive, Self-Optimising Signal Control for Isolated Intersections." Report RR170, Transportation and Road Research Laboratory, Crowthorne, United Kingdom.

Wallace, E.W and K.G. Courage (1982). "Arterial Progression – New Design Approach." In Transportation Research Record 881, Transportation Research Board, National Research Council, Washington, D.C., U.S.A., pp. 53-59.

Wallace, E.W., and K.G. Courage (1992). "TRANSYT-7F Users Guide." M.O.S.T. Volume 4, Federal Highway Administration, U.S. Department of Transportation, Washington, D.C., USA.

Wood, K. (1993). "Urban Traffic Control, Systems Review." Project Report 41, Transport Research Laboratory, Crowthorne, United Kingdom.

Yagar, S., B. Han and J. Greenough (1991). "A Real-Time Traffic-Responsive Signal Control Approach to Transit Priority." Report to the Department of Transportation of the Municipality of Metropolitan Toronto.

Yagar, S., B. Han and J. Greenough (1992). "Real-Time Signal Control for Mixed Traffic and Transit Based on Priority Rules." In Traffic Management, Engineering Foundation Press, pp. 147-166.

Yagar, S., B. Han, and J. Wang (1992). "SPPORT - A Real-Time Traffic-Responsive Signal Control Model for Transit Priority." Report to the Ministry of Transportation of Ontario.

Yagar, S. and B. Han (1994). "A procedure for Real-Time Signal Control that Considers Transit Interference and Priority." In Transportation Research B, Vol. 28B, No. 4, pp. 315-331.

Glossary

Amber interval – The portion of the signal cycle during which traffic facing a circular or arrow amber signal indication must stop before the stop line or other legally defined intersection boundary, unless such stop cannot be made safely. Please note that some jurisdiction may use different definitions.

All-red period – The period during which all traffic signal heads display red signal indications. Signal indications for movement continuing in the following stage may display green.

Approach – A section of the roadway upstream of the intersection stop line in which queues form.

Arrival flow – Flow rate on an intersection approach lane (exceptionally on a combination of lanes) upstream of the queue influence

Arterial – Signalized street that serves primarily through traffic an provides access to abutting properties as a secondary function, having signal spacing of 3.2 kilometers or less and turn movements that do not exceed 20 percent of total traffic.

Bottleneck – Street or highway segment where the volume of arriving traffic exceeds the capacity of the segment.

Capacity – Maximum departure flow that can discharge across the stop line of an intersection lane over an extended period of time, usually not less than 15 minutes (exceptionally across a stop line of an approach or a combination of approach lanes).

Change interval – The amber plus all-red intervals that occur between phases of a traffic signal to provide for the clearance of the intersection before conflicting movements are released.

Clearance lost time – The portion of the time at the end of signal phases during which an intersection is not used by any traffic movements.

Conflicting flow – The flow of traffic that is in a potential conflict with a specific movement.

Congestion period – The period of time (longer than just a few cycles) during which a continuous queue exists.

Control conditions – Prevailing conditions concerning traffic controls and regulations in effect, including the type, phase composition, cycle structure and timing of traffic control signals, stop and yield signs, permitted and prohibited movements, and similar measures.

Coordination of signals – Linking of traffic control signal timing at adjacent intersections in order to achieve specific operational objective, such as progressive movements of traffic or queue control..

Cycle – Any complete sequence of signal indication.

Cycle structure – The sequence and composition of phases in one cycle.

Cycle time – The total time for a signal to complete one sequence of indications.

Degree of saturation – Ratio of arrival flow to capacity.

Delay – The difference between the time required by an unimpeded passage of a flow unit through the intersection and the time actually needed under the prevailing geometric, traffic and control conditions.

Density – The number of vehicles occupying a given length of lane or roadway averaged over time, usually expressed as vehicles per kilometre or vehicle per lane per kilometre.

Departure flow – The rate at which the vehicles (or passenger car units) cross the stop line of a given lane during a given portion of the cycle time. Its maximum value is the saturation flow.

Display – Signal indication.

Downstream – The direction in which traffic is flowing.

Driver – A person who drives a private automobile or transit vehicle.

Dwell time – The time that a transit vehicle spends at a stop or station in order to discharge and board passengers.

Effective green interval – Duration of time equivalent to the period during which the departure flow of a fully saturated green interval can be represented by a uniform saturation flow. Also, the time allocated for a given movement (green plus amber) at a signalized intersection, less that start up and clearance lost time for the movement.

Effective red interval – The time during which a given traffic movement or set of movements is directed to stop, which correspond to the cycle length minus the effective green time.

Exclusive lane – An approach lane dedicated to only one direction of a departure movement (typically left-turn movement, straight-through movements or right-turn movement).

Fixed-time operation – A pre-programmed sequence and duration of signal indications. All signal intervals remain constant and are not affected by variations in traffic flow.

Flow – See arrival or departure flow.

Flow ratio – Ratio of arrival flow and saturation flow.

Free flow speed – The theoretical speed of traffic when no vehicles are present.

Gap – The time between two successive vehicles in a lane as they pass a point on the roadway , measured from the rear of the first vehicle to the front of the following vehicle.

Green band - Window of green signal indications that allows vehicle to travel uninterrupted along a set of adjacent intersections

Green split – The ratio of the effective green time for a given movement at a signalized intersection to the cycle length.

Green interval – The portion of the signal cycle during which traffic facing the circular or arrow green signal indication may proceed through the intersection in accordance with local laws and rules of the road.

Headway – The time between two successive vehicles in a lane as they pass a point of the roadway, measured from the front of one vehicle to the front of the successive vehicle.

High-occupancy-vehicle lane – A lane reserved for the use of vehicles with more than a preset number of occupants, such as transit vehicles, taxis and carpools.

Ideal conditions – Characteristics for a given type of facilities that are assumed to be the best possible from the point of view of capacity, that is, characteristics that if further improved will not result in increased capacity.

Intergreen period – The duration of time separating the end of the displayed green interval from the beginning of the next conflicting displayed green interval.

Interval – The duration of time during which a given indication is displayed.

Left-turn movement – A legally permitted movement of a vehicle which must cross the potential path of vehicles in the opposing direction.

Lost time in a phase – A portion of the phase (defined as the actual green interval plus the following intergreen period) that remains after the effective green interval has been subtracted.

Lost time in a cycle – The sum of lost times for all consecutive phases in the cycle.

Movement – Any legally permitted movement of a vehicle from a given lane.

Measures of effectiveness – Parameters describing the quality of service provided by a traffic facility to drivers, passengers or pedestrians; examples include speed, density, delay, and similar measures.

Motorist – The driver of a private automobile or transit vehicle.

Offset – The time, with respect to a given system reference point, at which the main street green interval is initiated at the intersection under control.

Passenger car – A motorized four-wheeled vehicle designed primarily for the transport of up to nine passengers. The term normally includes pickup trucks and vans with no more than four tires.

Passenger car unit – A unit of a completely homogeneous traffic, represented in practical terms by an average passenger car.

Pedestrian – A person afoot, in a wheelchair or pushing a bicycle.

Peak period – Period during which the maximum flow occurs during a given portion of the day.

Permitted left turns – A left-turn movement that takes place while the driver in the opposite direction of traffic face a circular green indication. Drivers making this left-turn must yield the right-of-way to the opposite flow.

Person – A driver or a passenger in a vehicle, but not a pedestrian.

Phase – That portion of a cycle during which the allocation of the right of way remains unchanged. This term normally includes the associated intergreen period.

Phase composition – The combination of vehicular, pedestrian and other movements legally permitted during a phase.

Phase plan – A sequence of phase implementation times over a given period of time.

Phase sequence – The order in which the phases follow each other in a cycle.

Platoon – A group of relatively fast moving vehicles with high traffic density.

Progression – A continuous movement of traffic in a given direction through two or more signalized intersections.

Protected left turns – A left-turn movement that takes place while the driver in the opposite direction of traffic face a circular red indication.

Queue – A line of traffic units waiting to be served at a signalized intersection. Slowly moving traffic joining the rear of the queue are usually considered as part of the queue. The interval queue dynamics may involve a series of starts and stops. A faster moving line of vehicles is often referred to as a moving queue or platoon.

Queue dissipation period – Period during which the number of vehicles departing from a queue exceeds the number of vehicle joining the back of queue at its tail and during which the total number of stopped vehicles within the queue diminishes.

Queue length – The number of traffic units in a queue, or the distance which is covered by the queue.

Queue reach – Used for vehicular traffic only, defined as the distance between the stop line of a lane and the point upstream at which vehicles are joining the queue, expressed in metres or as the number of vehicles that would fill that distance. The front of the queue may be some distance upstream from the stop line.

Real-time control – Traffic signal control in which the signal timing parameters are determined on line as new traffic conditions warrant it.

Reserved transit lane – A lane restricted to transit vehicle usage by special regulation and markings.

Right of way – Portion of a street where vehicles have the right to travel.

Saturation flow – The departure rate from a queue during the green interval, measured at the stop line.

Shared lane – A lane from which vehicles may discharge in more than one downstream direction. Also a lane shared by both passenger cars and transit vehicles.

Signal cycle – Any complete sequence of signal indication.

Signal indication – The following signal indication are defined in detail in the Manual for Uniform Traffic Control Devices for Canada, Part B, Traffic Control Signals (Transportation Association of Canada, 1993): green, amber, red, circular, arrow, steady, flashing, walk, don't-walk, flashing don't-walk and transit priority signal indications.

Signalized intersection – Intersection where traffic movements are controlled by traffic signals.

Signal operation – A term to describe the function of a signalized intersection, such as fixed-time operation and traffic-responsive operation.

Speed - A rate of motion expressed as a distance per unit of time.

Start lag – The interval of time between the moment a green signal indication is displayed and the moment a first vehicle crosses the intersection stop line.

Start-up lost time – Additional time consumed by the first few vehicles in a queue at a signalized intersection above and beyond the saturation headway because of the need to react to the initialization of the green phase and to accelerate to ambient speed.

Stop lag – The portion of the amber indication during which vehicles continue to flow across the intersection before traffic stops.

Stop line – Position along a given approach where vehicles must start to queue during a red interval, usually indicated by a large white mark extending across a traffic lane perpendicular to the direction of travel.

Traffic – The movement of motorized and non-motorized vehicles and pedestrians.

Traffic conditions – The combination of pedestrians and vehicle types at the intersection and on adjacent roadways, sidewalks, bicycle and other traffic facilities, including the temporal, directional and lane use distribution of traffic, and the type of driver or other user population.

Traffic flow – See arrival or departure flow.

Traffic-responsive operation – A control mode that adjusts the signal timings to prevailing traffic conditions.

Transit priority – A control mode in which transit vehicles receive a signal indication that provides for some advantage (usually shorter delay) to transit operations.

Transit stop – A position for a transit vehicle to pick up and discharge passengers, including curb stops and all other types of boarding and discharge facilities.

Transit vehicle – A heavy vehicle involved in the transit of passengers on a franchised transit basis.

Truck – A heavy vehicle designed primarily for the transport of goods.

Upstream – The direction from which traffic is coming.

Volume – The number of persons or vehicles passing a point on the roadway lane in a given time period.