FORMULATION AND EVALUATION OF A METHODOLOGY FOR NETWORK-WIDE SIGNAL OPTIMIZATION

By

SHIOW-MIN LIN

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1999

To my parents

ACKNOWLEDGMENTS

I would like to take this opportunity to thank my advisor, Professor Kenneth G. Courage, for honoring me by chairing my supervisory committee during the entire course of my doctoral research. His professional competence and hands-on experience in traffic signal control has inspired me tremendously. He recommended me to the Dwight David Eisenhower Transportation Fellowship Program, which has given me an opportunity to participate in research activities on a national level.

I would like to express my sincere gratitude to Dr. Henry C. Lieu for his invaluable advice and constructive criticisms in my doctoral research. He was also my technical advisor during the first three years of my assignments in the Turner-Fairbank Highway Research Center. His passion and knowledge in traffic simulation has taught me since day one when I came to Washington, D.C. For years, he has been more than a good friend to me.

I would also like to thank the rest of my supervisory committee, Drs. Charles E. Wallace, Jonathan J. Shuster and Scott S. Washburn, for their comments, suggestions, and interest in my research.

I wish to thank Dr. Joseph A. Wattleworth for providing me with financial support and serving on my supervisory committee during the early stage of my doctoral research. I enjoyed his classes and working with him.

I wish to express my appreciation to Dr. Boghos D. Sivazlian, my master advisor in Industrial and Systems Engineering, for his advice, encouragement, and serving on my supervisory committee during the early stage of my doctoral research.

I would like to extend my thanks to Dr. Rocco Ballerini for serving on my supervisory committee during the early stage of my doctoral research.

My appreciation is also due to Dr. Ilene D. Payne for her inspiration and assistance during my entire Dwight David Eisenhower Transportation Fellowship Program.

A special note of appreciation is due to the support provided by the Federal Highway Administration and the ITT Industries, Systems Division.

My appreciation is also extended to Dr. Nadeem A. Chaudhary for providing me with timely assistance in running the PASSER IV program.

Special thanks goes to Ms. Becky Hudson, Ms. Christine D. Ritch and Mr. Min-Tang Li for their timely support during the final stage of preparing my dissertation.

Finally, I wish to express my heartfelt thanks to my wife, Yiling. Without her love, understanding and constant support, I would have already given up my doctoral research long time before.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1-GC	First Generation Control
1.5-GC	One-and-a-Half Generation Control
2-GC	Second Generation Control
3-GC	Third Generation Control
ATIS	Advanced Traveler Information Systems
ATMS	Advanced Traffic Management Systems
BDP	Backward Dynamic Programming
CALIFE	Computer Based Traffic Control System (Translated from French)
CBD	Central Business District
CERT	Centre d'Etudes et de Recherches de Toulouse
CFP	Cyclic Flow Profile
CIC	Critical Intersection Control
COP	Controlled Optimization of Phases
CORSIM	Corridor Microscopic Simulation
CYRANO	Cycle-Free Responsive Algorithm for Network Optimization
DLL	Dynamic Link Library
DOW	Day-of-Week
DP	Dynamic Programming
DRIVE	Dedicated Road Infrastructure for Vehicle Safety in Europe
DYNAMIT	Dynamic Network Assignment for the Management of Information to Travelers

DYNASMART	Dynamic Network Assignment Simulation Model for Advanced Road Telematics
EVIPAS	Enhanced Value Iteration Process for Actuated Signals
FDP	Forward Dynamic Programming
FHWA	Federal Highway Administration
FORCAST	An On-line Signal Control Strategy of 1.5-GC
FRESIM	Freeway Microscopic Simulation
HOV	High Occupancy Vehicle
INTEGRATION	An Integrated Simulation and Traffic Assignment Model
IOB	Intersection Optimization Block
IP	Integer Programming
MATGEN	A Module in MAXBAND
MAXBAND	Maximal Bandwidth Traffic Signal Setting Optimization Program
MCODE	A Mathematical Programming System
MILP	Mixed Integer Linear Programming
MINOS	A Package for Solving the MILP Problem
MITSIM	Microscopic Traffic Simulator
MOE	Measure-of-Effectiveness
MOVA	Microprocessor Optimised Vehicle Actuation
MPCODE	An Optimization Routine in MAXBAND
MULTIBAND	An Extension of MAXBAND
NEMA	National Electronics Manufacturers' Association
NETSIM	Network Microscopic Simulation Model
OPAC	Optimized Policies for Adaptive Control
OSCO	Optimal Sequential Constrained Method
PARAMICS	Parallel Microscopic Simulation

PASSER	Progression Analysis and Signal System Evaluation Routine
PI	Performance Index
PRODYN	A French Real-time Traffic Algorithm
PROS	Progression Opportunities
QMC	Queue Management Control
RHODES	Real-time, Hierarchical, Optimized, Distributed and Effective System
ROW	Right-of-Way
RTOR	Right-Turn-on-Red
RT-TRACS	Real-time, Traffic Adaptive Control System
SCATS	Sydney Coordinated Adaptive Traffic System
SCOOT	Split, Cycle and Offset Optimization Technique
SIGOP	Traffic Signal Optimization Model
SO	Simultaneous Optimization
SOAP	Signal Operations Analysis Package
TANSTP	Traffic Adaptive Network Signal Timing Program
TOD	Time-of-Day
TRANSYT-7F	Traffic Network Study Tool, Version 7, Federal
TRRL	Transport and Road Research Laboratory
TRSP	Traffic Responsive
TRUSTS	Traffic Responsive and Uniform Surveillance Timing System
UTCS	Urban Traffic Control System
UTOPIA	Urban Traffic Optimization by Integrated Automation
VISSIM	Traffic in Towns - Simulation (Translated from German)

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

FORMULATION AND EVALUATION OF A METHODOLOGY FOR NETWORK-WIDE SIGNAL OPTIMIZATION

By

Shiow-Min Lin

December 1999

Chairman: Kenneth G. Courage
Major Department: Civil Engineering

There has been no doubt about the significance of what an optimized traffic signal system can do to our society today and beyond. With state-of-the-art technologies in computer, communication and surveillance becoming available, feasible and affordable, more and more accurate traffic information can be obtained and processed in real-time. In search of the literature, however, there exists no methodology that can simultaneously optimize network-wide signal phasing, timing and coordination in response to real-time traffic information. As a result, global optimality has been compromised and a traffic signal control system implemented in the field is constantly not operating in its best settings. Besides, it is found that a traffic flow model employed in an existing signal optimization method is too limited to provide realistic and reliable performance indices of concern. Consequently, the results generated by such a method are very questionable. Moreover, the computational difficulty identified in an existing signal optimization method suggests more room for improvement.

This dissertation revisits the generally recognized network-wide signal optimization problem in dynamic urban traffic control, and provides a methodology that can generate traffic signal control plans in such a way that signal phasing, timing and coordination are simultaneously considered, formulated and then optimized based on real-time traffic information. The methodology aims at a decentralized, non-parametric and cycle-free signal control, and it fully exploits a microscopic simulation technology so that performance indices of concern can be both realistically and reliably calculated in its optimization process.

The solution algorithm for the methodology developed in this study is based entirely on dynamic programming, and it is capable of performing network-wide signal optimization. It has been shown more computationally efficient without compromising global optimality. In addition, a heuristic search procedure has been developed. It can significantly reduce the computation and still generate comparable results. Both solution algorithms have been implemented and evaluated in a simulation testing environment, and the simulation results indicate significant improvements compared to a well-timed fixed-time control and an actuated signal.

The methodology developed in this study provides a feasible computational framework that can be applied to a dynamic urban traffic control in conjunction with Advanced Traffic Management Systems and Advanced Traveler Information Systems for network-wide signal optimization.

CHAPTER ONE
INTRODUCTION

Background

Since 1914 when the first electronic traffic signal in the U.S. was erected in Cleveland, Ohio (FHWA, 1996), there has been no doubt about the significance of what a smart and efficient traffic signal control system can do to our society. It is because of this belief that millions of dollars have been invested worldwide in improving traffic signal control, which can be characterized in twofold. With advances of hardware technologies in the areas of control, computer and communication, traffic signal control systems have evolved from simple electromechanical controls to sophisticated microprocessor-based systems. On the other hand, with the general premise that increased responsiveness would lead to improved traffic performance (Gartner, 1985), lots of research efforts have helped expedite the advances of traffic signal control from the First, then the Second, to the Third Generation Control, which were actually envisioned more than 30 years ago in the original Urban Traffic Control System (UTCS) Project (MacGowan & Fullerton, 1979-1980).

In searching the literature for what has been achieved so far in terms of state-of-the-art traffic signal control, it has been reported that only a few systems implemented in the field apply to their daily traffic operations the signal control plans that are somewhat responsive to real-time traffic measures (Tarnoff & Gartner, 1993). These systems are basically of the Second Generation Control such as SCOOT (Split, Cycle and Offset

Optimization Technique) and SCATS (Sydney Coordinated Adaptive Traffic System). These are centralized systems that adjust signal control parameters (e.g., cycle time, splits and offsets, but not phasing) in a cyclic fashion; however, they actually deviate from the commonly acceptable trend of having a decentralized, non-parametric and cycle-free signal control that is envisioned nowadays (Gartner, 1983; Sen & Head, 1997).

Yet, not all the advanced signal control strategies have always produced superior performance results as they should have conceptually (Gartner et al., 1995). For example, it has been reported (FHWA, 1976a) that in the evaluation of UTCS the overall performance of the Second and the Third Generation Control was actually inferior to that of the First Generation Control. Besides, when compared with traditional fixed-time control, the performances of SCOOT in floating car surveys were sometimes inferior or indistinguishable (Hunt et al., 1981; Bretherton, 1989). The causes contributing to these problems can be mainly summarized as follows:

- In the optimization processes of those systems, signal phasing, timing and coordination are not simultaneously considered, formulated and then optimized based on real-time traffic information to generate signal control plans. As a result, global optimality has been compromised and those systems are constantly not operating in their "best" settings.
- The traffic flow models employed in those systems are so limited that they are unable to provide their signal optimization processes with both realistic and reliable performance indices of concern. Consequently, the results generated by the signal optimization processes are questionable.

Although still under development, several control strategies of the Second and the Third Generation Control have already had deficiencies found in their methodologies. For example, the optimization procedure adopted by OPAC (Optimized Policies for Adaptive Control) is an exhaustive search for determination of only two phase changes to further reduce its computational difficulty (Gartner, 1983). Moreover, COP (Controlled Optimization of Phases) requires a desired phase sequence to start with (Sen & Head, 1997). It can be shown that not all signal phasing possibilities are being considered by COP unless enough cycles of phases have been added in the initial phase sequence; however, doing so will consequently decrease the computational efficiency of COP. Both OPAC and COP apply to single intersections only due to the fact that their computational difficulties prevent them from being expanded to network-wide signal optimization. Besides, they share the same problem of employing a simplistic traffic flow model.

With greater proliferation of computer, communication, surveillance and simulation technologies that are becoming available, feasible and affordable, more and more accurate traffic information can be obtained and processed in real-time. Therefore, it is the motivation of this study that there is a definite need for improvements in dynamic urban traffic control by fully exploiting the technologies.

Objectives and Scope

The main objective of this study is to develop a methodology for network-wide signal optimization so that signal control plans are generated in such a way that signal phasing, timing and coordination are simultaneously considered, formulated and then optimized based on real-time traffic information. The methodology is designed to

provide a feasible computational framework that can be applied to a dynamic urban traffic control in conjunction with Advanced Traffic Management Systems (ATMS) and Advanced Traveler Information Systems (ATIS). It aims at a decentralized, non-parametric and cycle-free signal control, and fully exploits a microscopic simulation technology so that performance indices of concern can be realistically calculated in its optimization process. However, it is not the purpose of this study to address congestion management issues, which are subject to the discretion and policies of each governing agency.

Figure 1.1 depicts an operational environment of dynamic urban traffic control that is envisioned in this study and generally recognized by the broad traffic research community. However, it is assumed available in this study the so-called "dynamic network partitioner, dispatcher and optimizer," whose functionalities are:

- To process real-time traffic measures provided by the surveillance system;
- To share real-time information with ATMS and ATIS systems;
- Based on both prevailing and predicted traffic conditions, to dynamically decompose the traffic network being considered into logical and reasonable sizes of subnetworks, such as isolated intersections, diamond interchanges, urban arterial streets, central business districts (CBDs), or a combination of the above;
- To provide the distributed signal control system with all necessary real-time information (including traffic predictions) and specified objectives and constraints so that signal control plans for each individual subnetwork can be optimized by the methodology being developed; and

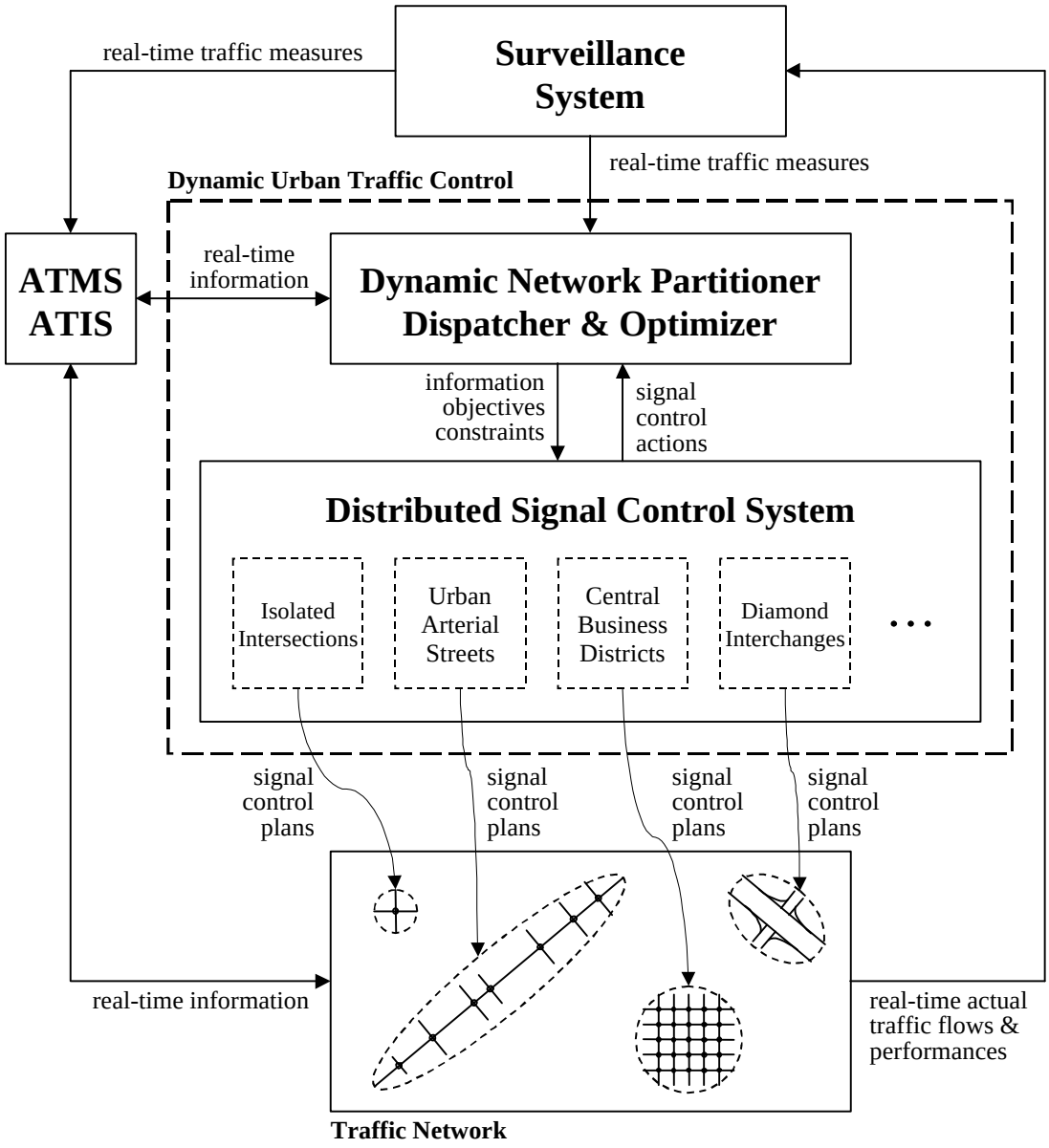


Figure 1.1. The Operational Environment of Dynamic Urban Traffic Control Envisioned in this Study.

- To integrate the subnetworks based on both overall network-wide objectives and optimized signal control actions taken by each individual subnetwork to assure smooth traffic progression across the subnetwork boundaries.

In order to develop such a methodology suitable for the dynamic urban traffic control environment envisioned in this study, specifically the following objectives are necessary:

- To review literature on the existing methodologies of signal control plan generation;
- To identify the functional requirements of a distributed signal control system;
- Based on the functional requirements being identified, to develop a modeling framework for the distributed signal control system;
- Based on the modeling framework being established, to develop a mathematical formulation for the generally recognized network-wide signal optimization problem;
- Based on the mathematical problem formulation being established, to develop a solution algorithm for the methodology being developed to solve the problem; and
- To evaluate the efficiency of the methodology being developed by virtue of a case study.

Organization

The dissertation is organized based on the objectives stated previously. The next chapter reviews the relevant literature on signal phasing, timing and coordination in the areas of off-line signal optimization models and on-line signal control strategies. In

addition, a discussion on microscopic traffic simulation models is presented given the fact that microscopic simulation is playing an increasingly important role in traffic signal optimization. A summary of findings from the literature search is provided as a result.

Chapter Three first identifies the operational characteristics of a distributed signal control system in dynamic urban traffic control, followed by the presentation of the functional requirements for the distributed signal control system. Then, a modeling framework is developed so that necessary information dynamics can be captured and modeled for network-wide signal optimization.

In Chapter Four, a mathematical problem formulation of integer programming is developed based on the modeling framework established in Chapter Three. The integer programming problem formulation serves as a reasonably faithful representation of the generally recognized network-wide signal optimization problem. Besides, it addresses many more real-world issues, which have not yet been attempted before.

Chapter Five develops a solution algorithm of dynamic programming for solving the network-wide signal optimization problem formulated in Chapter Four. A numerical example is provided to illustrate the dynamic programming calculations of the solution algorithm. In addition, a heuristic search procedure is developed that can significantly reduce the dynamic programming computation.

Chapter Six presents a case study to assess the efficiency of the methodology developed in this research. The case study is performed in a simulation testing environment, where the two solution algorithms developed in Chapter Five are implemented and evaluated against well-timed fixed-time controls and actuated signals. A summary of the findings is presented as a result.

Finally, Chapter Seven concludes this study with the summary of the findings and suggests directions for future research.

CHAPTER TWO
LITERATURE REVIEW

Introduction

Basic issues concerning controlling a traffic signal are phasing, timing and coordination. To a signal control strategy, those can be boiled down to one single issue as to how it satisfies the basic elements of signal control under what circumstances. In literature, there are four basic signal control elements identified: phase sequence, cycle length, split and offset. Hereinafter, these four basic elements may be collectively referred to as traffic signal control plan, or signal control plan for simplicity.

A signal control plan generating method, on- or off-line, basically contains two major components: an optimization process and a traffic flow model. An optimization process is nothing but a search procedure that explores a solution space for an optimality. Its computational efficacy is of main concern to the underlined method, especially for real-time applications. On the other hand, a traffic flow model is used along the optimization process to assess performance indices (or optimization criteria) of concern. Usually, it is either analytical or simulation-based (macroscopic, microscopic or mesoscopic) with deterministic and/or stochastic characteristics in nature. Its usefulness depends mainly on its capability of how realistically it can replicate real-world traffic flows subject to signal control decisions being explored, geometric configurations being examined, traffic operations being considered, and incidents being investigated.

The traffic engineering profession has more than 30 years of experience with computer control of traffic signals and the development and testing of various types of control strategies. Therefore, it is the purpose of this chapter to retrospect so that that body of knowledge can be fully exploited in the development of a better methodology for network-wide signal optimization. Failure to do so may cause yesterday's mistakes to become tomorrow's.

Next, several existing on- and off-line methods of signal control plan generation are reviewed. For each method, the main interests, if applicable, are as follows:

- Its ways to generate signal control plans;
- The effectiveness of its optimization process;
- The capabilities of its traffic flow model; and
- Its performances in terms of strengths and weaknesses.

In addition, several microscopic traffic simulation models are also reviewed given the fact that microscopic simulation is playing an increasingly important role in traffic signal control. Finally, a summary of findings from the literature review is provided as a result.

Off-line Signal Optimization Models

As commonly known, so far none of the off-line signal optimization models can explicitly and simultaneously optimizes all four basic elements of signal control. Nevertheless, it is certainly beneficial to learn the experiences from those models.

In this section, several popular models in off-line signal optimization are reviewed. First, the famous Webster method is introduced, followed by the discussions on several existing off-line signal optimization models that can be generally fit into the

following two categories: the disutility-oriented methods and the progression-based methods. Other models are addressed next. Finally, a summary of findings is presented as a result.

Webster Method

Due to the fact that a theoretical calculation of delay is very complex and that direct observation of delay in the field is complicated by uncontrollable variations, Webster (1958) pioneered in using computer simulation to help derive the well-known Webster's delay formula, assuming unsaturated, random arrivals for isolated, fixed-time signals. He detailed a procedure of how to calculate optimum cycle length and green times (splits) based on minimum overall intersection delay estimated by his delay formula (Webster, 1958; Webster & Cobbe, 1966).

Webster designated for each signal phase a critical lane (or critical movement) as the one with the highest ratio of flow to saturation flow (or flow ratio). Suppose y_i be the critical flow ratio for phase i. By definition,

$$y_i = \underset{a}{\text{Max}} \left\{ \frac{v_a}{s_a} \right\} \tag{2.1}$$

where v_a and s_a are the flow rate and saturation flow rate, respectively, for the designated critical lane a for phase i.

The optimum cycle length, C_o, can be approximated as:

$$C_o = \frac{1.5L + 5}{1 - Y} \tag{2.2}$$

where L is the total intersection lost time per cycle and Y is the summation of all critical flow ratios corresponding to each of the n phases in the cycle:

$$Y = \sum_{i=1}^{n} y_i \tag{2.3}$$

For each phase *i*, its optimum green time, g_i, is calculated by distributing the total available green time, i.e., C_o - L, in proportion to its critical flow ratio:

$$g_i = \frac{y_i}{Y}(C_o - L) \tag{2.4}$$

Notice that Webster's optimum allocation of green times is based on the equal degree of saturation for all phases in the cycle. It is dependent solely on Y and independent of L.

Disutility-Oriented Methods

Two models fit into this category are recognized, and they are TRANSYT-7F (Traffic Network Study Tool, Version 7, Federal) and SIGOP-III (Traffic Signal Optimization Model, Version III).

TRANSYT-7F (FHWA, 1998) is a macroscopic network optimization and simulation model, whose structure consists of two main parts:

- A macroscopic, deterministic traffic flow model which is used to compute the value of a specified performance index for a given traffic network and a given set of signal control settings; and
- A hill-climbing optimization procedure which makes changes to signal control settings (splits and offsets) and determines whether or not the specified performance index is improved.

The foundation of the traffic flow model used in TRANSYT-7F is the so-called "cyclic flow profile" (or CFP). A CFP is a measure of the average one-way flow of vehicles past any chosen point on the road during each part of the cycle time for the

upstream signal. The cycle time is divided into short time steps, each having a duration of about one to five seconds. Once a CFP is known, TRANSYT-7F can determine the size of the queue and the time required to clear it.

The primary objective of the optimization model used in TRANSYT-7F is to minimize the sum of the average queues in the network. This criterion is expressed as a performance index (PI). The PI also takes into consideration the number of times that vehicles have to stop and the so-called "progression opportunities" (or PROS). Using an optimization procedure based on an iterative gradient search technique, also known as hill-climbing, TRANSYT-7F explicitly optimizes splits and offsets to determine whether or not the specified PI is improved. By adopting only those changes that improve the PI, the optimizer tries to find a set of signal control settings which optimize the PI, subject to any limits placed on the process. The model also performs cycle range evaluation to determine the best cycle in conjunction with the optimal signal control settings.

TRANSYT-7F has the following known weaknesses and limitations that have been reported:

- It cannot optimize phase sequence, which is treated as an input to the model. Recently, Hadi and Wallace (1994) proposed using the so-called "Cauchy Simulated Annealing" technique to simultaneously optimize cycle lengths, phase sequences and offsets on the basis of PROS.
- It does not actually optimize cycle length. Instead, it evaluates a range of cycle lengths and selects the one with the best PI.
- Its hill-climbing optimization algorithm does not guarantee that a global optimum will be achieved and consequently does not guarantee that the "best"

signal control plan will be found either (Cohen, 1983). This is because the signal control problem in general has a solution space for the PI, which consists of a number of local optima. It is just computationally infeasible, when using the hill-climbing technique, to search through all local optima to find the best one.

- Its solution is mainly dependent on the quality of the starting solution, which is not always available (Chaudhary & Messer, 1993). Again, this is due to the nature of the hill-climbing technique just explained earlier.

Like TRANSYT-7F, SIGOP-III (FHWA, 1983) is a macroscopic signal timing model. It contains two main modules: a traffic flow module and a signal optimization module. Similar to its predecessors, SIGOP and SIGOP-II (MacGowan & Fullerton, 1979-1980), SIGOP-III optimizes signal control settings by embodying the interaction of split and offset in a sense that it trades off good offsets and green time in exchange of the optimal signal control. However, unlike its predecessors, it has a better traffic flow model so that vehicle delay, stops and queue length can be explicitly computed in a considerable detail during the optimization process.

SIGOP-III has the capability to evaluate multiple cycle length situations; however, each signal cycle can accommodate only a maximum of four phases (FHWA, 1985b). Besides, it neither models permissive and unprotected turns explicitly nor considers bus operations.

Progression-Based Methods

Four models fit into this category are recognized, and they are MAXBAND (Maximal Bandwidth Traffic Signal Setting Optimization Program), MULTIBAND, and

PASSER II and IV (Progression Analysis and Signal System Evaluation Routine, Model II and IV).

MAXBAND (Little et al., 1981; FHWA, 1985b) is a bandwidth-based signal optimization model that finds signal control settings on arterial streets and triangular networks to achieve maximal progression bandwidths. The model was developed by Little (1966 & 1977) based on mixed integer linear programming (MILP) formulation. Later, it was expanded by Messer (FHWA, 1987) to address general grid networks.

Recent enhancements to MAXBAND in terms of an arterial model were made by Tsay and Lin (1988), and Gartner et al. (1990). However, Stamatiadis and Gartner (1996) pointed out that Tsay and Lin's so-called "inverted funnel" progression model could not adequately adapt the bandwidth to the variations of flow since the band could only increase in width along the arterial street, and that its applicability is too limited as a result. Instead, they promoted the approach of MULTIBAND, an extension of MAXBAND, which will be discussed later.

As for the enhancements to MAXBAND as a network model, Chaudhary et al. (1991) proposed two heuristic methods to improve the computational efficiency of MAXBAND and to speed up its optimization process. Moreover, Stamatiadis and Gartner (1996) promoted the idea of MULTIBAND-95, which is the network version of MULTIBAND.

Due to the numerical instability of MAXBAND resulting in suboptimal or no solutions for network problems with a range of variable cycle times, Pillai et al. (1998) presented the development of a fast and numerically stable heuristic for the maximum bandwidth signal setting problem. The heuristic is based on restricted search of the

integer variables in the solution space. It is computationally efficient, and can generate optimal/near-optimal solutions.

MAXBAND computes offsets, cycle time, progression speeds, and left-turn phase sequence, which maximizes the weighted sum of bandwidths subject to cycle time constraint, bandwidth ratio constraint, interference constraints, loop integer constraints, and speed and speed-change constraints. It also has the capability to allow small deviations from the arterial-wide progression speed on individual links, a process referred to as speed search. MAXBAND uses the MCODE mathematical programming system, written by Land and Powell (1973), to solve the MILP problem formulation produced by the MATGEN module (FHWA, 1987). Notice that Land and Powell's method is a branch-and-bound search (Pillai et al., 1998).

The strengths of the MAXBAND model have been reported as follows:

- Compared with the disutility-oriented methods (to minimize delay, stops, etc.), it requires relatively little input and produces progression bands that can be easily visualized and understood by both traffic engineers and drivers (Little et al., 1981).
- Compared with PASSER II, it has a more rigorous mathematical programming model, wider set of decision capabilities (e.g., cycle time, progression speeds, etc.), expandability to network formulations, and capability to handle other objective functions.
- Unlike TRANSYT-7F, it achieves a global optimum, requires no starting solution, and optimizes cycle length and phase sequence (Cohen, 1983).

On the other hand, the weaknesses of the MAXBAND model have been reported as follows:

- One must invest substantial effort in learning how to formulate and solve problems in the way of MAXBAND.
- Its traffic flow model is extremely simplistic, which is one of the two reasons (the other is its computational inefficiency) that it has not gained acceptance in the traffic engineering community (Chaudhary et al., 1991). For this reason, it is not generally true that maximizing bandwidth minimizes such measures-of-effectiveness (MOEs) as stops or delay (Cohen, 1983).
- It requires extensive computation time since it is based on the MILP formulation and employs branch-and-bound techniques for its solution, which is infeasible for realistic network problems (Little et al., 1981; Chaudhary et al. 1991; Pillai et al., 1998).
- It does not optimize green splits, and has no capability of reporting traffic measures such as delay, stops and level of service (Chaudhary & Messer, 1993).
- Its generated progression schemes are of uniform width, which does not always hold. The effect of using average through-moving volume for allocating the total bandwidth is that the green band may be either wasted at intersections with lower than the average through-moving volume, or be deficient at intersections with higher than the average through-moving volume. This has been a major drawback of progression methods and, for

varying traffic volumes, optimum results cannot be guaranteed (Stamatiadis & Gartner, 1996).

- The numerical instability results in suboptimal or no solutions for network problems with a range of variable cycle times (Pillai et al., 1998).

The MULTIBAND model (Gartner et al., 1990; Stamatiadis & Gartner, 1996) is designed to overcome the limitations of MAXBAND by relaxing the assumption of a uniform platoon moving through all the signals. It enables the traffic engineer to specify a variety of flow-dependent objective functions, and optimizes the same variables as in MAXBAND (Gartner, 1991). The MULTIBAND model optimizes all the signal control variables, including phase lengths, offsets, cycle time and phase sequences, and generates variable bandwidth progressions on each arterial in the network that correspond to the specified objective. The mathematical programming package used for solving the MILP problem in MULTIBAND is the MINOS package (Murtagh & Saunders, 1993). As a result, a global optimality can be attained.

PASSER II (FHWA, 1991) is a macroscopic, deterministic model, which is based on an iterative, gradient search procedure to determine the combination of offsets over a range of specified cycle lengths that will result in the widest bands in both directions of the arterial street. The PASSER II model combines Brook's Interference Algorithm with Little's Optimized Unequal Bandwidth Equation, and extends them to multiphase arterial signal operations (Messer et al., 1973; FHWA, 1985b; Tsay & Lin, 1988). It uses an efficient heuristic optimization technique. The model first determines the optimal demand-to-capacity ratios and uses them to determine the splits. Trial cycle lengths, phase patterns, and offsets are varied to determine the optimal set of signal control

settings which maximizes the progression bandwidth (or minimizes the total interference to the progression bandwidth).

The weaknesses of the PASSER II model have been reported as follows:

- It is unable to handle multi-arterial networks with closed loops (Chaudhary et al., 1991).
- Its resulting control strategies generated often deteriorate performance on the cross streets, while ensuring the enhancement of traffic flow on the main arterial street. This deficiency becomes more evident in networks consisting of several intersecting arterial streets. A basic limitation of such a model is that the progression schemes generated are of uniform width (Stamatiadis & Gartner, 1996).
- Its heuristic optimization technique does not produce the widest possible green bands as MAXBAND does, and consequently it does not guarantee a global optimum (Tsay & Lin, 1988).
- Similar to other progression-based models, it does not model traffic flow explicitly. Rather, it uses discrete, deterministic models to estimate the traffic performance measures including permitted movements (FHWA, 1991).

PASSER IV (Chaudhary & Messer, 1993 & 1996) is so far the only progression-based model that can optimize signal control settings for large multi-arterial networks. It simultaneously maximizes progression bandwidth on all arterial streets (one-way and two-way) in closed networks. PASSER IV determines all four signal control parameters. It optimizes cycle length, offsets and phase sequences to maximize progression bandwidth; however, the green splits are determined by using the Webster method

(Webster & Cobbe, 1966) as discussed previously. In addition, it allows link to link speed variations, and arterial and directional priority options. However, like other progression-based signal optimization models, PASSER IV suffers from the fact that it employs a simplistic traffic model. Consequently, its applicability is limited. Besides, it does not have the capability to estimate traffic performance measures such as delay, stops and fuel consumption.

PASSER IV retained the MPCODE optimization routine used in MAXBAND (Chaudhary & Messer, 1993). In addition, it implemented two heuristic optimization techniques developed by Chaudhary et al. (1991): two- and three-step methods. It has been reported that the two-step method is 10 times faster than the simultaneous optimization (SO) of all variables and produces the same results as the SO method, and that the three-step method is up to 99 percent faster than the SO approach but it does not guarantee the absolute maximum bandwidth.

Other Signal Optimization Methods

Three models are discussed in this category, and they are EVIPAS (Enhanced Value Iteration Process for Actuated Signals), SOAP (Signal Operations Analysis Package), and PASSER III.

EVIPAS (Bullen et al., 1987; Halati & Torres, 1992) is an optimization model that is capable of analyzing and optimizing the settings for a variety of traffic signal controllers, including NEMA (National Electronics Manufactures' Association) and Type 170 controllers (one- or two-ring) in fixed-timed, semi-actuated, fully-actuated, or volume-density modes of operations. It is able to determine the optimum values of minimum green time, maximum green time, unit extension time, minimum gap time, time

before reduction, time to reduce, added initial, and maximum initial for each phase. The optimum timing value is defined as the timing setting which results in the minimum MOEs, including delay, fuel consumption, depreciation, other vehicle costs, and pollutant emissions.

The EVIPAS model has two major components: an optimization module and a traffic flow module. The optimization module is a multivariate procedure that uses quasi-Newton methods to find the optimal timing settings. Its numerical procedures compute first and second derivatives of an objective function for a given vector of parameter values. The values of these derivatives are used to determine the direction and size of steps from one parameter vector to the next, and to determine whether the minimum is reached. The algorithms and numerical methods are based on pseudo codes and subroutines for unconstrained optimization reported by Dennis and Schnabel (1983). The algorithms are designed to avoid certain types of local minima, although there exist conditions for which these methods fail to find the optimum solution. In general, however, these methods are among the best available for the solution of the same class of problems. On the other hand, EVIPAS makes the objective function out of its simulation model, which is an important feature in its structure. The simulation is a second-by-second vehicle scanning procedure considering car-following, vehicle queues and queue discharges.

The weaknesses of the EVIPAS model are basically due to the time needed for the optimization module to converge and the fact that it fails to find the optimum solution under some conditions.

SOAP (FHWA, 1985a) was designed to generate signal control plans for isolated intersections by evaluating a wide range of alternatives, including fixed-time or multiphase actuated control. It determines the optimum signal timing and phasing by three computational functions: design, analysis and evaluation. The design function examines all legitimate phase sequences, and selects the one that can be executed by using the minimum amount of green time. Optimum cycle length is determined based on minimum intersection delay subject to constraints on the amount of queuing that can be tolerated. The green is then allocated based on critical movements.

PASSER III (Fambro et al., 1991) was designed to analyze fixed-time (or traffic-responsive), fixed-sequence signalized diamond interchanges. The model evaluates existing or proposed signal control strategies to generate, based on the one that minimizes the average delay per vehicle, signal control plans for interconnecting a series of interchanges on one-way frontage roads.

Summary

In addition to the Webster method, a total of 9 off-line signal optimization models are reviewed in this section: two disutility-oriented methods, four progression-based methods, and three other models. Table 2.1 summarizes the ways each model generates signal control plans in terms of phase sequence, cycle length, splits and offset. In Table 2.2, the application areas for each model are listed along with the optimization criteria, traffic flow model and optimization process that are adopted by the model. Finally, in Table 2.3 the performances of each model are summarized in terms of its strengths and weaknesses.

Table 2.1. Summary of the Off-line Signal Optimization Models: Phase Sequence, Cycle Length, Splits and Offset.

Model Name	Phase Sequence	Cycle Length	Splits	Offset
EVIPAS	Optimum based on minimum delay & stops	Optimum based on minimum delay & stops	Optimum based on minimum delay & stops	Not applicable
MAXBAND	Evaluating all allowed for one with maximum bandwidth	Evaluating a range for one with maximum bandwidth	Analytical method without optimization	Optimum based on maximum bandwidth
MULTIBAND	Evaluating all allowed for one with maximum bandwidth	Evaluating a range for one with maximum bandwidth	Optimum based on maximum bandwidth	Optimum based on maximum bandwidth
PASSER II	Evaluating all allowed for one with maximum bandwidth	Evaluating a range for one with maximum bandwidth	Optimum based on minimum delay	Optimum based on maximum bandwidth
PASSER III	Evaluating all allowed for one with minimum delay	Evaluating a range for one with minimum delay	Evaluating all for one with minimum delay	Evaluating all for one with minimum delay
PASSER IV	Optimum based on maximum bandwidth	Optimum based on maximum bandwidth	Using Webster method	Optimum based on maximum bandwidth
SIGOP-III	Evaluating based on user-specified input	Evaluating a range for one with minimum delay, stops & excessive queue	Optimum based on minimum delay, stops & excessive queue	Optimum based on minimum delay, stops & excessive queue
SOAP	Evaluating all allowed for one with minimum amount of greens	Optimum based on minimum delay	Allocating greens based on critical movements	Not applicable
TRANSYT-7F	Evaluating based on user-specified input	Evaluating a range for one with best PI	Optimum based on best PI	Optimum based on best PI

Table 2.2. Summary of the Off-line Signal Optimization Models: Applications, Criteria, Traffic Flow Models and Optimization Processes.

Model Name	Applications	Criteria	Traffic Flow Model	Optimization Process
EVIPAS	Single intersections	Disutility (delay, stops)	Microscopic simulation	A multivariate procedure using quasi-Newton methods to find optimal timing settings.
MAXBAND	Networks	Progression (bandwidth)	Analytical	Solving MILP for offsets, cycle length, progression speeds & left-turn phase sequences, with fixed splits, to maximize weighted sum of bandwidths subject to constraints of cycle length, bandwidth ratio, interference, loop integers, and speed & speed-change.
MULTIBAND	Networks	Progression (bandwidth)	Analytical	Using MINOS package for solving MILP to simultaneously optimize offsets, cycle length, splits & phase sequences based on maximum weighted sum of outbound & inbound link-specific bandwidths on each arterial street in network, subject to constraints of cycle length, bandwidth ratio, interference, loop integers, and speed & speed-change.
PASSER II	Arterial streets	Progression (bandwidth)	Analytical	Using iterative, gradient search procedure over specified range of cycle lengths to determine best combination of phase sequence, cycle length & offsets for minimum-delay splits per cycle that will produce highest bandwidth efficiency in both directions of arterial street, while allowing progression speed variation; and then fine-tuning offsets to further minimize delay at non-critical intersections.

Table 2.2. -- continued.

Model Name	Applications	Criteria	Traffic Flow Model	Optimization Process
PASSER III	Diamond interchanges	Disutility (delay)	Analytical	Using interactive, trial-and-error method to evaluate all possibilities.
PASSER IV	Networks	Progression (bandwidth)	Analytical	Using one global & two heuristic branch-and-bound methods for solving MILP to simultaneously maximize progression bandwidths on all arterial streets in network.
SIGOP-III	Networks	Disutility (delay, stops, excessive queue length)	Macroscopic simulation	Using iterative, gradient search over collection of possible signal settings for optimality to minimize weighted sum of link-specific delay, stops & excessive queue lengths.
SOAP	Single intersections	Disutility (delay)	Analytical	Using interactive "Design" procedure to determine optimum cycle length that minimizes total delay subject to queue constraints, and to allocate green based on critical movement analysis.
TRANSYT-7F	Networks	Disutility (delay, stops), PROS	Macroscopic simulation	Using iterative gradient search (hill-climbing) to explicitly optimize splits & offsets for best PI given phase sequence & range of cycle lengths.

Table 2.3. Summary of the Off-line Signal Optimization Models: Strengths and Weaknesses.

Model Name	Model Strengths	Model Weaknesses
EVIPAS	Capable of analyzing & optimizing signal settings for NEMA & Type 170 controllers in fixed-timed, semi-actuated, fully-actuated, or volume-density modes of operations; Equipped with a microscopic simulation model in optimization process.	No guarantee on global optimality; Time needed for optimization procedure to converge; Only applicable to single intersections.
MAXBAND	Rigorous mathematical programming model; Global optimum achieved; Wide set of decision capabilities; Capable of performing signal control optimization in multi-arterial, closed-loop networks; No initial solution required; Less input required.	Unable to optimize offsets, cycle length, splits & phase sequences simultaneously; Unable to optimize green splits; Unable to model more than 4 approaches per intersection; Numerical instability resulting in suboptimal or no solutions; Computationally inefficient; Assuming uniform bandwidths; Substantial learning effort required for MILP; Oversimplified traffic flow model; Not necessarily minimizing delay & stops; No MOEs other than bandwidth data reported.
MULTIBAND	General case of MAXBAND model; Capable of optimizing offsets, cycle length, splits & phase sequences; Global optimum achieved; Producing variable-bandwidth; Sensitive to traffic variations along each arterial street in network while optimizing progression on crossing arterial streets as well; Significant benefits in terms of delay, stops & fuel consumption when compared with conventional uniform bandwidth models.	Not yet operational; Substantial learning effort required for MILP; Not necessarily minimizing delay & stops; Oversimplified traffic flow model; Unable to model more than 4 approaches per intersection; No MOEs other than bandwidth data reported.

Table 2.3. -- continued.

Model Name	Model Strengths	Model Weaknesses
PASSER II	Capable of providing optimal timing plans (phase sequence, splits, offsets & cycle lengths) for signal operations ranging from simple two-phase signal to dual-ring concurrent control; Computationally efficient.	Unable to handle multi-arterial networks with closed loops; Unable to optimize offsets, cycle length, splits & phase sequences simultaneously; No guarantee on global optimality; No optimization for left-turns; Assuming uniform bandwidth; Small set of decision capabilities; Simplistic traffic flow model.
PASSER III	Able to evaluate diamond interchanges & parallel frontage roads.	Unable to optimize offsets, cycle length, splits & phase sequences simultaneously; No guarantee on global optimality; No optimization for phase sequence; Assuming uniform bandwidth; Small set of decision capabilities; Simplistic traffic flow model.
PASSER IV	Capable of optimizing signal control settings for large multi-arterial networks; Simultaneously maximizing progression bandwidth on all arterial streets.	Unable to optimize offsets, cycle length, splits & phase sequences simultaneously; Unable to optimize green splits; Unable to model more than 4 approaches per intersection; Assuming uniform bandwidths; Substantial learning effort required for MILP; Not necessarily minimizing delay & stops; Simplistic traffic flow model; No MOEs other than bandwidth data reported.
SIGOP-III	Capable of evaluating multiple cycle lengths	Maximum of 4 phases per cycle; Unable to consider buses; Permissive & unprotected turns not addressed explicitly; Macroscopic traffic flow model.
TRANSYT-7F	Capable of optimizing network-wide signal control settings; Explicitly optimizing splits & offsets.	Unable to optimize phase sequence; Not actually optimizing cycle length; No guarantee on global optimality; A starting solution required; Macroscopic traffic flow model.

On-line Signal Control Strategies

After several decades of evolvement in computerized signal control, traffic engineering profession today still refers state-of-the-art on-line signal control strategies to those three generations of signal control that were envisioned more than 30 years ago in the original UTCS Project (MacGowan & Fullerton, 1979-1980).

In this section, several existing on-line signal control strategies are reviewed in the nomenclature of UTCS. They are categorized based on the definitions of the three generation controls in UTCS. Other signal control strategies are also reviewed. Finally, a summary of findings is presented as a result.

First Generation Control Strategies

The First Generation Control (1-GC) employs a library of prestored signal control plans, which are developed off-line based on historical traffic data using popular signal optimization models such as TRANSYT-7F and MAXBAND discussed in the previous section. Basically, the criteria to select a plan that drives a signal control system are threefold: on the basis of time-of-day (TOD) and day-of-week (DOW), by direct operator selection, or by matching from an existing library a plan best suitable for recently measured traffic conditions (e.g., volumes and occupancies). Two on-line signal control strategies that fit into this category are recognized, and they are 1-GC UTCS and FORCAST.

In 1-GC UTCS (MacGowan & Fullerton, 1979-1980), the "matching" mode of signal control plan selection is named the traffic responsive (TRSP), whose update frequency is usually once every 15 minutes. The results from a comprehensive study (FHWA, 1976a) indicated that TRSP generally matched or exceeded the performance of

TOD and well-timed fixed-time signal. To ensure a smooth transition from one signal control plan to another, 1-GC UTCS equips itself with a transition routine. Also, it contains a critical intersection control (CIC) feature, which enables adjustment of green splits at selected signals that saturate frequently. However, for years the efficacy of the CIC algorithm has been doubtful (FHWA, 1976a; Farradyne Systems, Inc., 1990).

It is proven complex and burdensome to maintain a library of signal control plans in 1-GC UTCS (FHWA, 1996), due to the fact that the efforts of great magnitude are required to collect, assemble and input data for an off-line signal optimization model. To overcome this problem, a simple signal control plan generating procedure is necessary. It uses data collected by a signal control system, analyzed off-line, and followed by automated loading of the generated plans into the system. In literature, this concept has been referred to as the One-and-a-Half Generation Control (1.5-GC), and FORCAST (FHWA, 1996) is an example of 1.5-GC.

Second Generation Control Strategies

The Second Generation Control (2-GC) is a real-time, on-line strategy that computes and implements signal control plans based on surveillance data and predicted changes. Three on-line signal control strategies that fit into this category are recognized, and they are 2-GC UTCS, SCOOT and SCATS.

The 2-GC UTCS program, or TANSTP (Traffic Adaptive Network Signal Timing Program) (FHWA, 1976b), includes an optimization algorithm (i.e., SIGOP), a traffic prediction model, subnetwork configuration models, a CIC feature, and a signal transition model. In 2-GC UTCS (MacGowan & Fullerton, 1979-1980), the optimization process is repeated at 5-minute intervals; however, to avoid transition disturbances, new signal

control plans cannot be implemented more often than once every 10 minutes. For the subnetwork configuration models, they are designed to dynamically decompose the network into subnetworks based on prevailing traffic conditions so that the optimum signal control plans are then computed for individual subnetworks and the subnetworks are interfaced to assure smooth traffic progression across the subnetwork boundaries.

The performances of 2-GC UTCS were mixed (FHWA, 1976a). Although demonstrating some small improvements on arterial streets compared to well-timed fixed-time signal, it showed degradation in performance on a network-wide basis. In general, its overall performance was inferior to 1-GC UTCS.

SCOOT (Hunt et al., 1981) was initiated by the British Transport and Road Research Laboratory (TRRL) in the 1970's, with its first commercial system installed in 1980. There are now about 60 implementations within the United Kingdom, and more than a dozen installations around the world (Martin & Hockaday, 1995).

SCOOT is a cyclic, parametric, centralized, traffic-responsive signal control system, which does not take advantage of vehicle-actuated control tactics at local intersections. Its principles are basically threefold (Robertson, 1986):

- On-line measurement of CFPs. A CFP, as discussed earlier, is the foundation of the traffic flow model used in TRANSYT; it is fundamental to SCOOT as well. For this reason, the first idea in SCOOT is to measure CFPs on-line.
- On-line traffic flow model. SCOOT is essentially an on-line TRANSYT. With the measured CFPs, it performs queue estimates, which are needed in real-time by the signal optimizer. Notice that it assumes platoons traveling at

a known cruising speed with some dispersion and queues discharging at a known and constant saturation flow rate.

- Incremental optimization. Given the difficulty of predicting traffic situations in a relatively long period of time, SCOOT makes signal control plan changes in a series of frequent, but small, increments. It uses an "elastic" plan that can be stretched or shrunk to fit into the latest situation suggested by the measured CFPs in the following manner. A few seconds before every phase change, its split optimizer determines whether to advance or retard the scheduled change by up to four seconds, or to leave it unaltered. Then, once a cycle, the offset optimizer assesses whether the PI can be reduced by altering the offset of each intersection by up to four seconds earlier or later. Favorable split and offset alterations are implemented immediately. The cycle time for a group of intersections may, in a similar fashion, be incremented up or down by a few seconds every few minutes.

The benefits of SCOOT have been reported compared to well-timed fixed-time signal (Hunt et al., 1981; Greenough & Kelman, 1998). There are critiques on SCOOT reported as well. For example, it is incapable of optimizing signal phasing and has no mechanism to impose turning restrictions based on traffic demand.

SCATS (Lowrie, 1982) has been developed by the Australian Roads and Traffic Authority of New South Wales since the early 1970's. Similar to SCOOT, it is a cyclic, parametric, traffic-responsive system. However, unlike SCOOT, it uses two levels of control, also known as "strategic" and "tactical," to adjust cycle time, splits and offset.

SCATS controls signals in groups with a critical intersection specified for each group. Cycle time and splits are calculated for each critical intersection. As for other intersections, their signal control plans are selected to match the plans for the critical intersections. Strategic control is carried out by a higher-level system, responsible for a group of up to about 10 signals. Tactical control, on the other hand, is handled by local controllers, which is basically seeking self-optimization within the restrictions imposed by strategic control.

Cycle time optimization takes place each cycle. Splits may vary each cycle based on the average degree of saturation on approaches over the last three cycles. Phases may terminate earlier or, when there is no demand, be omitted altogether. To avoid oscillating situations, the optimum offset is calculated each cycle, but only implemented when at least three out of the previous five cycles have suggested a change to that offset. Overall speaking, the methodology adopted by SCATS has been seen as empirical in nature.

A survey (Luk et al., 1982) indicated that SCATS showed no significant reduction in travel times compared with operation using TRANSYT; however, there was a large reduction in the number of stops. Similar results have also been reported recently by Wolshon and Taylor (1998).

SCATS does not have any facilities to automatically ban turning movements to improve network performance. Although some phase skipping capabilities are available at local controllers, changes to a phase order will require programming the central computer in advance by TOD. The requirement to identify critical intersections within groups of intersections has been reported less satisfactory, especially for new road networks or ones in which conditions are different at various times.

Third Generation Control Strategies

The Third Generation Control (3-GC) is a fully traffic-responsive, on-line signal control strategy, which permits essentially "cycle free" operations and allows the parameters of signal control plans to change continuously in response to real-time measurement of traffic variables. Four on-line signal control strategies that fit into this category are recognized, and they are 3-GC UTCS, OPAC, PRODYN, and COP.

In the traffic-responsive mode of 3-GC UTCS (FHWA, 1976c; MacGowan & Fullerton, 1979-1980), signal control plans are updated by a central computer for all controllers at least once every three minutes. For unsaturated conditions, they are guided by the policy called CYRANO, or the Cycle-Free Responsive Algorithm for Network Optimization, which is the subject of the following discussion. For the sake of completeness, the other policy is called CIC/QMC, which consists of a CIC policy coupled with a QMC (Queue Management Control) policy that is applied along paths of congestion.

For CYRANO, a signal control plan is not defined explicitly in terms of cycle length, splits and offset, but rather as a sequence of signal switching points. The signal offset and splits are determined to minimize vehicle delay and stops along each approach and to provide network-wide coordination. The signal coordination is accomplished by implementing a coarse simulation of traffic flow and then systematically adjusting the signal settings at each controller to minimize a disutility function, which is a linear combination of vehicle stops and delay aggregated over all approaches. As a result, a signal transition routine is essentially unnecessary because the system is constantly in a state of transition as it responds to changing traffic conditions.

3-GC UTCS showed serious degradation in performance under almost all of the conditions for which it was evaluated (FHWA, 1976a). It worked worst when supposed to work best during the arterial off-peak period. Three factors contributing to its poor performance are significant:

- Its entire signal switching sequence was predetermined for every control period, and therefore it was not necessarily responsive to real-time traffic conditions.
- Its variable cycle was not an output of its optimization process.
- Its optimization process required a relatively long period of time for convergence (much longer than 2-GC UTCS for the same size of network), which prevented it from reaching an optimality as a result.

OPAC has been developed in the U.S. since the early 1980's (Gartner, 1983; FHWA, 1989), which is an on-line computational strategy for demand-responsive traffic signal control. It is currently applicable only to isolated intersections; however, its network version is being developed (Farradyne Systems, Inc., 1995) and field-tested (Andrews et al., 1997).

Unlike parametric models that optimize parameters such as cycle time, splits and offsets, OPAC features a dynamic optimization algorithm, which optimizes the control decisions in stages rather than simultaneously. The length of a typical stage is similar to a cycle length for a fixed-time signal and consists of an integral number of fixed-time intervals. For each stage, given the initial queues on each approach and the arrivals for each interval of the stage, the underlined optimization problem is to determine the switching times (in time intervals) which yield the best performance index (i.e., total

delay) to vehicles over the entire stage. During each stage, OPAC requires at least one phase change and up to three phase changes are allowed. For any given switching sequence at stage *n*, the performance index is defined as follows:

$$\Phi_n(t_1,t_2,t_3)=\sum_{i=1}^{k}(Q_0+A_i+D_i) \tag{2.5}$$

where

(t_1,t_2,t_3) = possible switching times during stage *n*;

Q_0 = initial queue;

A_i = arrivals during interval *i*;

D_i = departures during interval *i*; and

k = total number of intervals in stage *n*.

The procedure used for solving the aforementioned optimization problem is an optimal sequential constrained (OSCO) method. It is an exhaustive (i.e., brute-force) search of all possible combinations of valid switching times within the stage to determine the optimum set. Valid switching times are constrained only by minimum and maximum phase durations.

Critiques on OPAC have been reported in literature, which are summarized as follows:

- It is implemented through the rolling-horizon technique, and has been shown to reduce delays in a number of studies (Gartner et al., 1991).
- Although its calculated results approach the theoretical optimum, it does not guarantee a global optimality due to the fact that its optimization process does not lend itself very well to dynamic programming (Sen & Head, 1991).

- Its optimization process is limited due to the tremendous computational efforts involved in the OSCO search and the incapability of handling signal phasing. Although up to three phase changes are permitted for each stage, it allows at most two phase changes to further reduce the search space (FHWA, 1989).
- It employs a simplistic traffic flow model that does not consider real-world traffic scenarios, and consequently its applicability is limited.

PRODYN (Henry et al., 1983; Barriere et al., 1986; Henry & Farges, 1989) is a French real-time traffic control algorithm, which has been developed by the Centre d'Etudes et de Recherches de Toulouse (CERT) over the last decade. It is part of the European DRIVE (Dedicated Road Infrastructure for Vehicle Safety in Europe) program (Khoudour et al., 1991). PRODYN evolves from two stages of development: two-level hierarchical control (PRODYN-H) and then decentralized control (PRODYN-D). The former offers the best result; however, its applicability is restricted due to complex computations involved and network sizes (limited to about 10 intersections). The latter, on the other hand, alleviates those limitations and prevails eventually. Two approaches have been studied: no exchange (PRODYN-D1) versus exchange (PRODYN-D2) of information between the so-called intersection optimization blocks (IOBs). It has been reported that one of the PRODYN-D2 methods offers as good results as PRODYN-H.

Both PRODYN-H and PRODYN-D are built upon a basic IOB, which features forward dynamic programming. Similar to OPAC, the PRODYN optimization algorithm is implemented through the rolling-horizon technique, which features an open-loop optimal feedback. Consider a horizon that is equal to a finite number of fixed-length sampling periods. Given a set of discrete-time, nonlinear state variables, the optimization

problem for each intersection in each sampling period is to determine the switch-over at the end of the current period such that the total delay is minimized over the horizon that starts at the end of the current sampling period. The total delay is approximated by the sum of the vertical queues over each approach to the intersection and each sampling period of the horizon, plus the sum of the terminal criterion for each approach (Robertson & Bretherton, 1974). The constraints introduced to the optimization problem include the binary nature of the control variables, minimum and maximum green times, and limits of queue lengths. As for the state variables used for optimization, they are given as follows:

- Predicted vehicle arrivals of each approach for the current sampling period and for the entire horizon;
- Estimated queue lengths for each approach at both the beginning and the end of the current sampling period;
- Signal status and its time elapsed since the last switch-over; and
- Predicted conflicting non-priority movements.

The procedure (Larson & Casti, 1978 & 1982) used to solve the aforementioned forward dynamic programming problem in PRODYN differs from the classical one in that:

- It performs forward comparisons inside disconnected subsets of the state space having a maximum storage of only one state, instead of using reverse state equation; and
- Memory is not allocated to all possible subsets at all the time steps but only to subsets which correspond to reached states at a given time step.

Although an earlier experiment showed no benefit over an isolated fixed-time signal, PRODYN-D has been reported to reduce travel time in a single intersection case and several small network cases (Farges et al., 1990 & 1991). Major critiques on PRODYN reported are related to its limitation to two-phase operations, restriction for allowing only four approaches to an intersection, incapability of handling signal phasing, and employment of simplistic traffic flow model.

COP was originally developed by Sen (1991) as a module to provide intersection control for a system called Real-time, Hierarchical, Optimized, Distributed and Effective System, or RHODES (Head et al., 1992). Recently, it has been further enhanced (Sen & Head, 1994 & 1997) and included in a project called Real-time, Traffic-Adaptive Control System, or RT-TRACS (Farradyne Systems, Inc., 1993; Dell'Olmo & Mirchandani, 1995 & 1996; Mirchandani & Head, 1998), for the Federal Highway Administration (FHWA).

Like OPAC and PRODYN, COP is a non-parametric model implemented through the rolling-horizon technique. However, as opposed to the exhaustive search procedure adopted by OPAC, it is based entirely on dynamic programming characterized by only one state variable. Consider a desired sequence of n signal phases ($n \geq 1$) and a rolling horizon of T fixed-length time intervals. Given the predicted demand for each phase j in [1,n], the underlined optimization problem is to determine each phase duration x_j (in time intervals) such that a given performance index is optimized over the horizon as well as two safety requirements are satisfied: minimum green for a phase allocated a nonzero green time and an all-red (or clearance) time between phase changes. Notice that COP treats each phase in the initial phase sequence as a stage in dynamic programming.

Let the state variable s_j be the total number of time intervals that have been allocated after stage *j* has been completed. To incorporate the all-red time *r* (in time intervals) for transition between phases, the following relationship is established:

$$s_{j-1} = s_j - h_j(x_j) \tag{2.6}$$

where $h_j(x_j)$ = 0, if x_j = 0; x_j + *r*, otherwise. However, given a value for the state variable s_j, the control variable x_j can assume values that must exceed a prescribed minimum (requirement) whenever it is allocated a nonzero value. Thus, the forward dynamic programming can be written as:

$$V_j(s_j) = \text{Min}\left\{ f_j(s_j, x_j) \circ V_{j-1}(s_{j-1}) \,\middle|\, x_j \in X_j(s_j) \right\} \tag{2.7}$$

where

$V_j(s_j)$ = cumulative value of the performance index at stage *j*;

$f_j(s_j, x_j)$ = incremental value of the performance index associated with making decision x_j at stage *j*;

$X_j(s_j)$ = discrete set of feasible values that the control variable x_j can assume for the phase planned for stage *j*; and

$\circ$ = operator "+" when minimizing total delay or total number of stops, or operator "Max" when minimizing maximum queue lengths.

That is, at each stage *j*, $V_j(s_j)$ is evaluated for each feasible phase duration x_j, associated with the state variable s_j, to determine the optimal control decision x_j^* of the stage. Complete all the calculation for each stage *j*. Then, the optimal phase sequence

with each phase duration can be determined through the backward pass of the forward dynamic programming.

The performance indices (criteria) considered by COP include total delay, total number of stops, maximum queue length and queue spillback. Their calculations are based entirely on the formation and dissipation of queue lengths, which are functions of the state variable and the predicted vehicle arrivals to the intersection. However, it is important to note that queue lengths are calculated for accounting purposes and their traditional role as state variables (for optimization) is abandoned (Sen & Head, 1994).

Limited simulation experiments on COP have been conducted (Sen & Head, 1994 & 1997), and the results indicated reduction in vehicle delay compared to semi-actuated signal. It has been reported that COP tends to allocate longer phases for approaches in which more vehicle arrivals occur and favors the through movements by skipping the left turn phase on several occasions. More critiques on COP are summarized as follows:

- It requires an initial phase sequence to start with and optimal signal phasing is achieved by skipping unwanted phases in the sequence. It can be shown that not all the phasing possibilities are being considered by COP unless enough cycles of phases have been included in the initial phase sequence. However, doing so inevitably increases its computation since the number of dynamic programming stages has been increased consequently. Strictly speaking, COP does not explicitly optimize signal phasing.
- It applies only to single intersections. Given its current structure, there is a computational difficulty of expanding its methodology to cover network-wide situations.

- Like OPAC and PRODYN, it suffers from the fact that it employs a simplistic traffic flow model. Consequently, its applicability is limited.

Other Signal Control Strategies

Four on-line signal control strategies are discussed in this category, and they are CALIFE (French for Computer Based Traffic Control System), MOVA (Microprocessor Optimised Vehicle Actuation), TRUSTS (Traffic Responsive and Uniform Surveillance Timing Systems), and UTOPIA (Urban Traffic Optimization by Integrated Automation).

CALIFE (Gabard et al., 1986; Gabard, 1991) is an on-line method that calculates and implements fixed-time signal control plans for network-wide signal control. It is designed to overcome the problems found in 1-GC strategies, such as employment of obsolete timing plans and lack of integrated method for pattern-matching plan selection. It has a prediction model and a plan optimizer. The latter is actually taken from the TRANSYT/7 model with some modifications. CALIFE has been evaluated by virtue of macroscopic simulation, and the simulation results indicated that its methodology and its congestion prevention method are feasible.

MOVA (Vincent & Peirce, 1988), a traffic-responsive control system for isolated traffic signals, was developed at the British TRRL during the 1980's. It is designed to improve the performance of a simple gap-seeking vehicle-actuated control system, which has been used by Britain for many years, without frequent time-consuming resetting of a signal controller. It is because of this self-optimizing nature that there is no need for the user to calculate good signal timings. Yet, MOVA is not suitable for coordinated signals.

MOVA has two main modes of operations: delay minimization for unsaturated traffic conditions and capacity maximization for congested traffic situations. The former

is based on Miller's algorithm. MOVA can hold up to three signal control plans, which can either be selected by TOD or manually (either by phone or on site). During its development, MOVA has been extensively tested at four sites. The results indicated reduction in delay, compared with well-timed actuated signal.

TRUSTS (Tsay, 1989) was developed in Taiwan in the late 1980's. It is a decentralized system. Similar to 1-GC strategies, it has basic modes of signal control plan generation: TOD, on-line matching, and on-line generation. On-line signal control plans are generated by a modified version of TRANSYT-7F or a bandwidth maximizing program (Tsay et al., 1991). TRUSTS has been installed in several cities in Taiwan.

UTOPIA (Mauro & di Taranto, 1989) was developed by the Italian FIAT Research Center in the early 1980's. It has two levels of control with emphasis on decentralization of optimization in its lower (intersection) level by decomposing a large-scale network problem so that the resulting problems can be solved in a hierarchical manner. The intersection level consists of two components: a microscopic model and an optimization model. It is an open loop feedback control implemented through the rolling horizon framework. UTOPIA has been field tested and implemented since 1984. Its benefits have been reported.

Summary

In this section, a total of 13 on-line signal control strategies are reviewed. They are two 1-GC's, three 2-GC's, four 3-GC's, and four others. Table 2.4 summarizes the characteristics of the different generation control strategies of UTCS. In Table 2.5, a summary of review findings on the on-line signal control strategies is presented.

Table 2.4. Summarized Characteristics of the Different Generation Control Strategies of UTCS.

Feature	1-GC	1.5-GC	2-GC	3-GC
Update Frequency (Control Period)	15 minutes	15 minutes	5-10 minutes	3-5 minutes (Variable)
Signal Control Plan Generation	Off-line optimization; Selection from library by TOD & DOW, traffic-responsive, or manual mode.	On-line optimization after user confirmation; Selection from library by TOD & DOW, traffic-responsive, or manual mode.	On-line optimization	On-line optimization
Traffic Prediction	(None)	(None)	Historically based	Smoothed values
Critical Intersection Control (CIC)	Fine-tuning of splits	Fine-tuning of splits	Fine-tuning of splits & offsets	(Not applicable)
Cycle Length	Fixed within each section	Fixed within each section	Fixed within variable groups of intersections	Variable in time and space; Predetermined for control period.

Table 2.5. Summary of the On-line Signal Control Strategies.

Strategy	Features	Critiques
1-GC UTCS	Signal control plan library developed off-line; Plan selection based on TOD, manual or traffic-responsive.	Difficulty in maintaining library of signal control plans up-to-date; Irresponsive to real-time traffic conditions.
1.5-GC UTCS	Automated loading of signal control plan analyzed off-line.	Irresponsive to real-time traffic conditions.
2-GC UTCS	Signal optimization model (SIGOP) embedded; More frequent update of signal control plans than 1-GC UTCS.	Simplistic traffic flow model employed; Lack of sound methodology suitable for on-line applications.
SCOOT	Cyclic, parametric, centralized, traffic-responsive signal control method; On-line measurement of CFPs; TRANSYT-based on-line traffic flow model; Incremental optimization.	Deviating from trend of having non-parametric, decentralized control; Simplistic traffic flow model employed; A global optimality not guaranteed.
SCATS	Cyclic, parametric, bi-level, traffic-responsive signal control method with some local adaptability; Principles based on equalized degree of saturation.	Deviating from trend of having non-parametric, decentralized control; Empirical methods employed, lack of theoretical base.
3-GC UTCS	Cycle-free, non-parametric, centralized, fully traffic-responsive signal control method; Most frequent update of signal control plans compared to 1- & 2-GC UTCS.	Deviating from trend of having decentralized control; Lack of sound, on-line methodology considering signal phasing, timing & coordination simultaneously; Computationally infeasible.
OPAC	Non-parametric, on-line optimization (DP) model; Implemented through rolling-horizon technique.	Exhaustive search procedure, not computationally feasible; Applicable to single intersections only; Simplistic traffic model employed.
PRODYN	Non-parametric, decentralized, on-line optimization (DP) method; Implemented through rolling-horizon technique.	Complex computation necessary; Network size restriction; Simplistic traffic model employed; Limited applications.
COP	Non-parametric, on-line optimization (DP) method; Implemented through rolling-horizon technique An optimality achieved	Methodology limited to single intersection case; Initial phase sequence required, consequently not all signal phasing possibilities considered unless enough cycles of phases added to initial phase sequence; Simplistic traffic model employed.

Table 2.5. -- continued.

Strategy	Features	Critiques
CLAIFE	On-line method to calculate & implement fixed-time signal control plans for network-wide signal control; Concept similar to 1.5-GC strategy.	Simplistic traffic flow model employed; Lack of local adaptability.
MOVA	On-line traffic-responsive method to improve performance of actuated signal by adjusting signal controller settings in real-time; Self-optimizing.	Applied to single intersections only; Not applicable to coordinated signal; Incapable of estimating fuel consumption, pollutant emissions, queue spillback, etc.; Degree of saturation only estimated under certain conditions; Capacity not specifically calculated.
TRUSTS	Decentralized system in network of personal computers; Concept between 1.5- and 2-GC; On-line optimization (TRANSYT or bandwidth type) model embedded.	Not considered fully traffic-responsive due to few detectors employ.
UTOPIA	Bi-level signal control method with strong emphasis on decentralization of optimization by decomposing a large-scale network problem to be solved in hierarchical manner; Open loop feedback control implemented through rolling-horizon technique.	Simplistic traffic flow model employed; Limited applications found so far after more than a decade of field implementation.

Microscopic Traffic Simulation Models

It is worthwhile to highlight the significance of using a simulation model in a dynamic urban traffic control. In search of literature, it has been observed that a traffic flow model employed in an on- or off-line signal control plan generating method tends to be simulation-based, given the fact that a complex, real-world traffic network with randomness in nature simply cannot be specified completely and accurately by a pure mathematical model that can be evaluated analytically. Examples can be found in off-line signal optimization models such as EVIPAS, SIGOP-III and TRANSYT-7F, and on-line signal control strategies such as 2-GC UTCS, 3-GC UTCS and SCOOT. For a signal control plan generating method that employs a simplistic traffic flow model, which deviates from that fact just stated, it suffers from the fate of having limited applications. To name a few, MAXBAND and PASSER IV are of such a case. In other words, the applicability of a signal control plan generating method highly relies on the capability of how realistically its traffic flow model can address real-world traffic control issues. It is because of this reason that prototypes of advanced control strategies such as DYNAMIT (Dynamic Network Assignment for the Management of Information to Travelers) (Ben-Akiva et al., 1998) and DYNASMART (Dynamic Network Assignment Simulation Model for Advanced Road Telematics) (Mahmassani et al., 1994) are heavily simulation-oriented.

In what follows, a well-known microscopic simulation model named CORSIM (Corridor Microscopic Simulation) is discussed in terms of geometric representation, traffic dynamics, traffic control, driver behavior, MOEs, applications, etc. For the similar discussions on other simulation models, the reader may also refer to INTEGRATION

(van Aerde, 1995), MITSIM (Microscopic Traffic Simulator) (Yang, 1997), PARAMICS (Parallel Microscopic Simulation) (Cameron & Duncan, 1996), VISSIM (German for Traffic in Towns - Simulation) (Fellendorf, 1994), or a comprehensive review recently done by Bernauer et al. (1998).

CORSIM (FHWA, 1999) is a microscopic traffic simulation model, which is part of the TRAF family simulation programs that have been developed by FHWA for more than three decades. It combines two component models, NETSIM (Network Microscopic Simulation) and FRESIM (Freeway Microscopic Simulation), which are capable of simulating urban streets and freeways, respectively, in great detail.

CORSIM simulates a network that represents a traffic environment being studied by dividing it into component subnetworks, and the subnetworks are then interfacing with each other. All the interfacing processes are handled internally by the interface logic; however, the user has total control over partitioning the network being simulated.

CORSIM models each vehicle as a separate entity. The behavior of each vehicle is represented in the model through interaction with surrounding environment by the vehicle, which includes geometry, control devices, events and other vehicles. Each time a vehicle is moved, its position (both lateral and longitudinal) on a link and its relative position to adjacent vehicles are recalculated, as are its speed, acceleration, and status. As a result, each vehicle's behavior can be simulated in a manner reflecting real-world processes such as lane-changing and car-following. The reader may refer to a paper presented by Halati et al. (1997) for further discussion.

CORSIM provides a comprehensive set of features that can be applied to virtually all the day-to-day traffic situations in real-world. Some of them are listed as follows:

- Detailed geometric representation, which includes turn bays, lane restrictions and channelizations, interchanges, freeway mainlines and auxiliary lanes, lane drops (or additions), grades, and radius of curvature and superelevation of a freeway.
- Realistic traffic stream, which includes 10 different types of drivers, 4 fleet components, 9 different performance capabilities of vehicles, and pedestrians.
- Assorted traffic control, which includes sign control, fixed-time and actuated signals, ramp-metering control, and even an interface available for a real-time traffic adaptive control.
- Stochastic simulation, which applies not only to turn movements, free-flow speeds, queue discharge headways and other behavioral attributes, but also to car-following laws and lane-changing decisions.
- Special events, which includes queue spillback, turn pocket overflow, blockages, parking activities, and long- and short-term events.

CORSIM calculates almost all the MOEs commonly used for decision-making purposes, including volumes, speeds, travel times, delays, link occupancies, fuel consumption, pollutant emissions, etc. They can be obtained on network-wide, link-specific or movement-specific bases. Further, they are grouped by types of facilities (e.g., surface streets, freeway sections and special lanes), and types of vehicles (e.g., automobiles, buses, trucks and carpools).

CORSIM is one of the powerful study tool available and has been applied to a wide range of activities, including: basic academia research such as a vehicular delay study, and capacity analyses for signalized and unsignalized intersections and freeway

junctions; practical traffic/transportation engineering activities such as signal re-timing, traffic impact studies, giant parking stadium operations, corridor traffic operations and capacity analyses, and freeway incident detection and management; and evaluations of advanced traffic control strategies such as real-time traffic-adaptive control, real-time traveler information and route guidance, and network-wide dynamic traffic assignment.

Summary

In this chapter, several on- and off-line signal control plan generation methods have been reviewed with a discussion on microscopic traffic simulation models. In what follows, the review findings are summarized and they are critical to the development of the subsequent chapters:

- A signal control is basically seeking optimization within its own capacity;
- An on-line, cycle-free, non-parametric, decentralized methodology suitable for network-wide signal optimization is very much desired;
- Phasing, timing and coordination being considered simultaneously in signal optimization are uncommon due to the computational complexity involved;
- Dynamic optimization processes implemented in rolling-horizon framework are norms in advanced signal control strategies;
- A traffic flow model employed tends to be simulation-based due to the fact that a complex traffic situation cannot be fully specified by analytical models;
- A methodology that employs a simplistic traffic flow model suffers from the fate of having limited applications; and
- Not all the real-time traffic information available has been fully exploited in signal optimization processes.

CHAPTER THREE
A MODELING FRAMEWORK OF NETWORK-WIDE SIGNAL OPTIMIZATION FOR A DISTRIBUTED SIGNAL CONTROL SYSTEM

This chapter develops a modeling framework for a distributed signal control system in dynamic urban traffic control, which is envisioned in this study and described in Chapter One, so that a commonly recognized network-wide signal optimization problem can be modeled and resolved. In what follows, the key operational characteristics of a distributed signal control system are identified first based on the findings in Chapter Two, followed by a list of its functional requirements and the presentation of its modeling framework.

Operational Characteristics

A distributed signal control system is basically seeking optimization within its own capacity based on the geometric aspects given, and the real-time information and specified objectives and constraints provided by its higher-level control. However, it cannot be a successful one without the fully identifiable natures in its operations. In what follows, the key characteristics among those are listed:

- Various geometric aspects. The spatial considerations of different geometric aspects are essential to a distributed signal control system. An urban traffic network is so complex that it may contain different geometric configurations such as isolated intersections, urban interchanges, arterial streets, central business districts, etc. Likely, it involves different geometric designs such as

turn bays, shared lanes, lane drops, lanes specific for carpools and/or buses, uncontrolled driveways, parking lots, traffic circles, etc.

- State-of-the-practice traffic operations. Day-to-day traffic operations are fundamental to any signal control system. They include different left-turn treatments (protected, permitted and protected/permitted), right-turns-on-red (RTORs), shared-lane operations, lane closures, high occupancy vehicle (HOV) lanes, priority signals, pedestrians, buses, trains, etc.
- High variations in real-time traffic data. Real-time traffic measures derived from a surveillance system are subject to a high degree of variability due to the results of interconnected urban signals and the randomness of traffic flows. Consequently, the real-time traffic measures inputted to a distributed signal control system may not be smoothed at all times.
- More traffic predicted information available in real-time. Under the dynamic urban traffic control in conjunction with ATMS and ATIS systems, a lot of accurate traffic predicted information can be made available in real-time to a distributed signal control system.
- Timely updates to signal control plans. Instead of hours, a time interval for a distributed signal control system to update signal control plans is in the matter of seconds, or minutes (the longest), since drivers respond in much smaller time resolution to events such as changing signal phases, moving vehicle traffic, queues at intersections, etc.
- Cycle free operations. In a distributed signal control system, the concept of cycle length does not exist even for signal coordination. On the contrary, the

parameters of signal control plans are allowed to change continuously in response to real-time information and real-time traffic measures. As a result, interconnected signals are coordinated naturally in this manner without a background cycle length.

- Dynamic phase sequencing. For a distributed signal control system, a phase sequence is generated based on real-time information and real-time traffic measures. As a result, it is not fixed over time.
- Time-variant turn control. The severity of turning conflicts at intersections caused by turning movements and their opposing traffic is time-variant. Therefore, a simple time-invariant type of turn control (i.e., protected, permitted, protected/permitted, or no turns at all) may not be appropriate to a distributed signal control system all the times.
- Varying performance indices. Due to the purposes of different TOD and DOW operations, and/or the requirements by different jurisdictions, an ability to consider various performance indices under different circumstances is of greatest interest to a distributed signal control system. Commonly used MOEs are stops, delay, speed, throughput, queue length, travel time, fuel consumption and pollutant emissions, or the combinations of the above. Usually, when traffic is light, stop-based MOEs are of primary concern since delay is expected to be low anyway. When traffic is getting saturated, queue-based MOEs are used to determine if any spillover toward upstream intersections occurs, since delay is expected to be high anyway. In-between, delay-based MOEs are employed.

- Special events. Events such as blockages, incidents and parking activities happen abruptly. However, they need to be taken into consideration as well by a distributed signal control system.

Functional Requirements

Based on the operational characteristics identified in the previous section, a distributed signal control system should have the following functions to:

- Handle various geometric aspects;
- Consider day-to-day traffic operations;
- Respond to assorted real-time traffic predicted information;
- Realistically capture time-spatial traffic evolution in real-time;
- Dynamically generate phase sequencing and time-variant turn control in response to real-time information;
- Timely optimize and update cycle-free, non-parametric signal control plans based on real-time information and instructions provided by its higher-level control;
- Evaluate various performance indices of concern; and
- Assess impacts due to special events such as incidents and blockages.

Modeling Framework

To satisfy the functional requirements specified in the previous section, the modeling framework for a distributed signal control system should consist of two main components: a signal optimization module and a microscopic simulation module. They

are corresponding to the two major components of a signal control plan generating method, on- or off-line, as reviewed in Chapter Two.

Figure 3.1 depicts such a modeling framework that is logically derived from the aforementioned discussions, given that the following pieces of information are already available:

- Systems data, which includes the duration of discrete time interval, total length of time horizon, yellow time and all-red time;
- Static geometric data, which includes intersections, streets, pavement markings and lane designations; and
- A set of signal phases designed for each intersection.

As indicated in Figure 3.1, the output from the distributed signal control system basically contains the signal control plans that will be implemented next at intersections in individual subnetworks. The signal control plans generated will also be fed back to the higher-level control to assure smooth traffic progression across the subnetwork boundaries. On the other hand, the input to the distributed signal control system from its higher-level control contains the following three items:

- Real-time information, which includes: (1) dynamically partitioned subnetwork boundaries, (2) traffic measures derived from the surveillance system, (3) traffic predictions also contributed by ATMS and ATIS systems, (4) prevailing traffic operations, (5) incidents being reported, and (6) transit (bus) operations.
- Specified objectives, which are specified performance indices to be investigated in conjunction with individual subnetworks being optimized.

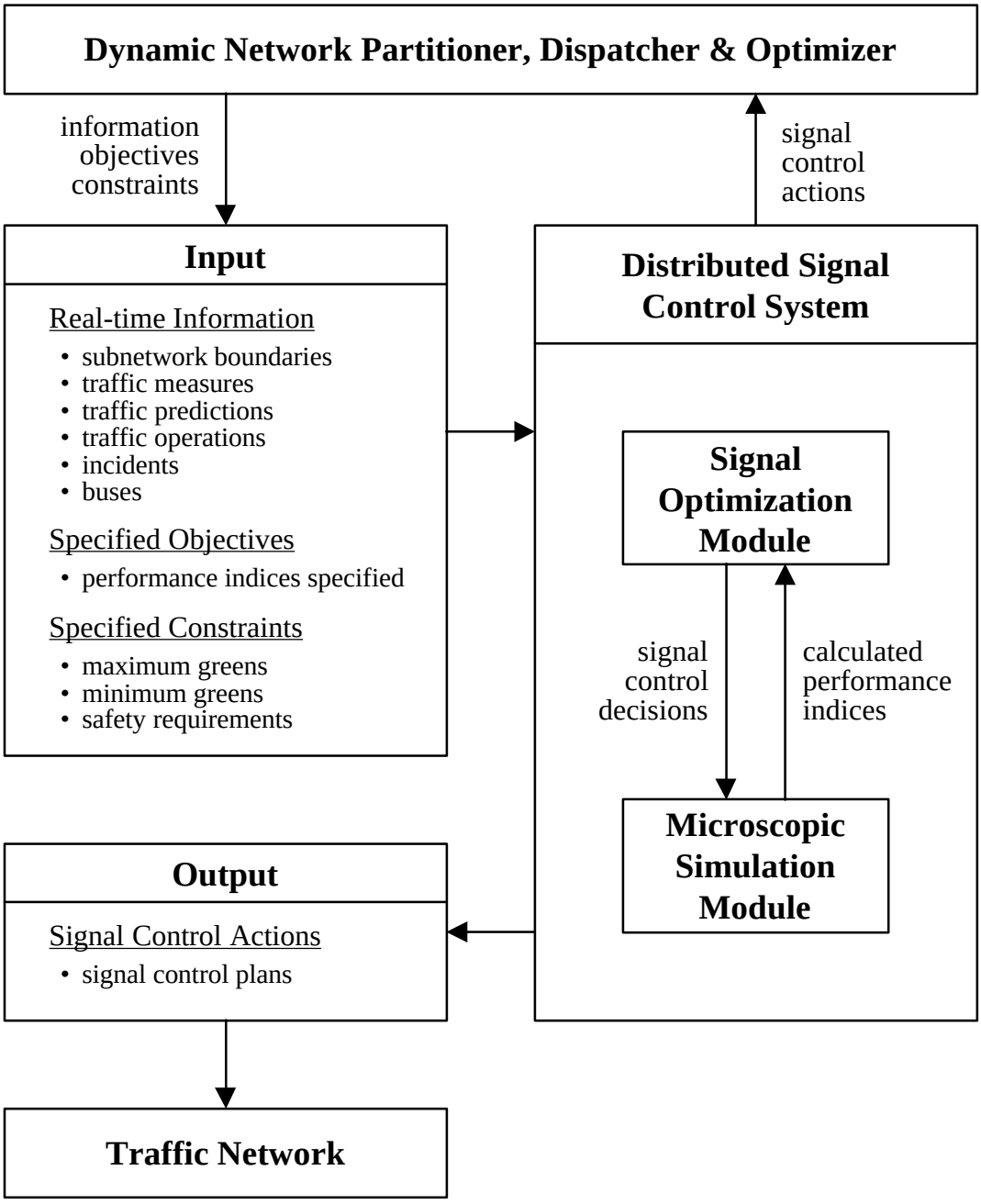


Figure 3.1. A Modeling Framework of Network-wide Signal Optimization for a Distributed Signal Control System in Dynamic Urban Traffic Control.

- Specified constraints, which include: (1) maximum green times allowable for movement-specific phases, (2) minimum green times required for each phase, and (3) some safety requirements such as "yellow must turn red." For the last constraint, there are guidelines discussed by Kell and Fullerton (1991) about amber signal displays, especially for left-turn movements, and they will be adopted in this study.

It is worthwhile to elaborate the usage of the maximum and minimum green time constraints imposed by the higher-level control. For example, the higher-level control may want to guarantee a movement to receive a specified amount of green time. In which case, it may set the constraints for both the maximum and minimum green times to that specified value for that movement-specific phase. To ignore a movement, on the other hand, it may simply zero those two green time settings. In other cases, it may allow the green time to be allocated for a movement within a range by properly setting up the values of the maximum and minimum greens. However, when traffic is getting saturated, one can expect to see that range getting smaller as an obvious result.

The microscopic simulation module inputs the real-time information specified earlier. Together with the signal control decisions provided by the signal optimization module, it realistically captures time-spatial traffic evolution in real-time. In return, it provides the signal optimization module with the calculated performance indices of concern for the purpose of optimizing signals. Notice that in this study no new microscopic simulation model will be reinvented; instead, the existing CORSIM simulation model will be fully exploited to carry out the task given the fact that the CORSIM model has already had most of the capabilities required in this study. For

detail, the reader may refer to the discussions on CORSIM in Chapter Two. However, the CORSIM model will be heavily modified and enhanced so that it can fit into the modeling framework of this study. A detailed discussion on the customized version of CORSIM will follow in Chapter Six.

The signal optimization module is the intelligent part of the distributed signal control system. Based on the real-time information, and specified objectives and constraints provided by its higher-level control, it directs the signal optimization process by solving the mathematical programming being formulated using the solution algorithms being developed. During the optimization process, it properly and timely launches the microscopic simulation module whenever required to assess the needed performance indices of concern. What it does is provide the microscopic simulation module with the necessary signal control decisions being tested so that with the given real-time information the microscopic simulation module can in return provide it with the resulting performance indices; and so forth.

In what follows, a mathematical formulation of the generally recognized network-wide signal optimization problem will be laid out in Chapter Four based on the modeling framework developed in this chapter, while its solution algorithm will be presented in Chapter Five.

CHAPTER FOUR
A MATHEMATICAL MODEL FORMULATION FOR THE NETWORK-WIDE SIGNAL OPTIMIZATION PROBLEM

Introduction

As reviewed in Chapter Two, an on-line signal control model traditionally has been stated as a discrete-time optimal control problem like OPAC (Gartner, 1983), PRODYN (Henry et al., 1983) and COP (Sen & Head, 1997). However, these models, more or less, require a desired phase sequence to start with, and their optimization processes basically attach signal timing to the sequence one way or another. That is, both signal phasing and timing being treated simultaneously are uncommon. Besides, not a single model explicitly addresses real-world issues such as detailed geometric aspects, realistic phasing, permitted movements, buses, incidents, etc.

In this chapter, a general integer programming (IP) model is formulated based on the modeling framework developed in Chapter Three, while its solution algorithm will be presented in the next chapter. The IP model formulation has been motivated by the same need to address many of the same issues that have directed the developments of OPAC, PRODYN and COP. Consequently, it serves as a reasonably faithful representation of the signal optimization problem solved by those models. Besides, it addresses many more real-world, network-wide issues that have not yet been attempted before. In what follows, the notations that will be used hereinafter are introduced first, followed by the presentation of the IP model formulation. Finally, a summary is provided as a result.

Definitions and Assumptions

Let $\Gamma = \{N, A\}$ denote a traffic subnetwork (simply stated as network hereinafter) being considered. It consists of a set of nodes N and a set of links A that are dynamically defined in real-time by the higher-level control, as described in Chapters One and Three. Let $\Omega = \{X\}$ denote a solution space, containing a set of signal control decisions X. Based on the modeling framework specified in Chapter Three, the network-wide signal optimization problem is to find an optimal solution ω^* in solution space Ω over traffic network Γ and time horizon T so that performance indices of concern can be optimized, given the real-time information, and specified objectives and constraints provided by the higher-level control.

To facilitate the presentation of the general IP model formulation, the following notations are necessary and they are grouped in a logical order. Notice that they are all non-negative integers.

Systems Variables:

Δt = discrete time interval (given);

T = total number of discrete time intervals, or total length of time horizon (given);

Y = yellow time (given);

R = all-red time (given);

Sets:

N = set of intersections (given);

A = set of links (given);

M_i = set of phases for intersection *i*, each composed of protected and allowable unprotected movements in the intersection (given);

P = set of joint phases for all intersections in traffic network Γ, each composed of a phase *m* in M_i for each intersection *i* (given); i.e.,

$$\mathsf{P} = \left\{ m \,\middle|\, m \in \mathsf{M}_i, \forall i \right\};$$

S_k = set of *k*-phase sequences, where $k \geq 2$ (given); i.e.,

$$\mathsf{S}_k = \left\{ p_j \,\middle|\, p_j \in \mathsf{P}, \forall j \in [1,k];\ p_j \neq p_{j-1}, \forall j \in [2,k] \right\};$$

X = set of signal control decisions; i.e.,

$$\mathsf{X} = \left\{ \begin{array}{l} (x_j^{ps}, \tau_j^s) \,\middle|\, x_j^{ps} \in \{0,1\},\ p \in \mathsf{P},\ j \in [1,k]; \\ \sum_{j=1}^{k} \sum_{\forall p \in \mathsf{P}} x_j^{ps} = k; \sum_{j=1}^{k} \tau_j^s = T; \forall s \in \mathsf{S}_k, \forall k \geq 2 \end{array} \right\};$$

V = set of vehicle types (given);

$\mathbf{\Psi}$ = entire collection of real-time traffic information (given); i.e.,

$$\mathbf{\Psi} = \left\{ \alpha_i^{mv}(t), \xi_{ia}(t), \upsilon_{ia}(t), \zeta_{ia}(t), \eta_{ia}(t), \delta_{ia}(t) \,\middle|\, \forall m, v, a, i, t \right\};$$

Index Variables:

t = discrete time index;

$i \in \mathsf{N}$ = intersection index;

$a \in \mathsf{A}$ = link index;

$m \in \mathsf{M}_i$ = phase index for each intersection *i*;

$p \in \mathsf{P}$ = joint phase index;

$s \in \mathsf{S}_k$ = phase sequence index;

$v \in \mathsf{V}$ = vehicle type index;

Signal Control Variables:

p_1 = initial phase, $p_1 \in \mathsf{P}$ (given);

τ_0 = time already allocated to initial phase p_1 in the immediately previous horizon, assuming that the minimum green requirement has already been satisfied (given);

τ_j^s = time allocated to the j^{th} phase in phase sequence *s*, which includes the duration of a change interval (yellow time plus all-red time), if applicable;

x_j^{ps} = 1, if joint phase *p* is the j^{th} phase in phase sequence *s*; and 0, otherwise;

ϕ_i^{mp} = 1, if phase *m* of intersection *i* is in joint phase *p*; and 0, otherwise; (given);

θ_{ia}^m = 1, if link *a* is in phase *m* of intersection *i*; and 0, otherwise; (given);

$G_i^{m,max}$ = maximum green time allowed for phase *m* of intersection *i* (given);

$G_i^{m,min}$ = minimum green time required for phase *m* of intersection *i* (given);

G_{ij}^{ms} = green time for phase *m* of intersection *i* elapsed since its onset until the end of the j^{th} phase in phase sequence *s*, if phase *m* is in the j^{th} phase of the phase sequence (i.e., $\phi_i^{mp} \cdot x_j^{ps} = 1$); and 0, otherwise;

$\gamma_i^{m'm}$ = signal clearance interval needed from phase m' to phase *m* of intersection *i*: 0, if $m' = m$; and a value equal to either *Y* or *Y* + *R* depending on phases m' and *m*, otherwise; (given);

Link-Specific Traffic Variables:

$\alpha_{ia}^{mv}(t)$ = number of vehicle type *v* arrivals on link *a* associated with phase *m* of intersection *i* at time *t* (given);

λ_{ia}^{max} = maximum number of vehicles that can be accommodated on link *a* of intersection *i* (given);

$\pi_{ia}(t)$ = number of vehicles on link *a* of intersection *i* at time *t*;

$\pi_{ia}(0)$ = number of vehicles on link *a* of intersection *i* at the very beginning of the current horizon (given);

$\xi_{ia}(t)$ = incident information associated with link *a* of intersection *i* at time *t* (given);

$\upsilon_{ia}(t)$ = transit (bus) operations information associated with link *a* of intersection *i* at time *t* (given);

$\zeta_{ia}(t)$ = parking activity information associated with link *a* of intersection *i* at time *t* (given);

$\eta_{ia}(t)$ = pedestrian traffic information associated with link *a* of intersection *i* at time *t* (given);

$\delta_{ia}(t)$ = other event information associated with link *a* of intersection *i* at time *t*, which results in any link capacity reduction (given);

Functions:

$F(\cdot)$ = objective function (functional form given);

$\beta_{ia}^{m}(t)$ = number of vehicle departures from link *a* associated with phase *m* of intersection *i* at time *t* (functional form given);

$\psi_{ia}^{m}(t)$ = number of vehicle departures from link *a* made by permitted movements or RTOR vehicles in phase *m* of intersection *i* at time *t* (functional form given);

$f_{\beta}(t)$ = queue discharge model based on prevailing traffic conditions associated with function $\beta_{ia}^{m}(t)$ at time *t* (functional form given).

The Model Formulation

An optimization problem can be generally stated as: optimize an objective function $F(\cdot)$ subject to a set of constraints. In this section, the concept of the network-wide signal optimization model formulation will be presented first based on the modeling framework developed in Chapter Three, followed by the discussions of the objective function $F(\cdot)$ and the constraints.

Concept

The main idea of the model formulation is with a simple mechanism to associate complicated and tedious modeling issues involving geometric details, individual signal phases, intersection movements (protected and unprotected), traffic dynamics, and signal control decisions. In this study, that mechanism is called "joint phase." By doing so, a network of many intersections can be treated as if it were a single intersection, as long as the linkage has been properly maintained between a joint phase *p* and the corresponding phases of each intersection so that the minimum and maximum green constraints can be verified. Consequently, from the modeling point of view the size of a network will not matter; however, the complexity of the computation incurred still will. In what follows, the concept of the joint phase is illustrated by virtue of a graphical representation.

Figure 4.1 shows a network of two tight offset T-intersections, and Figure 4.2 shows all possible phases associated with each intersection. Based on the definition, it can be shown that

$$\mathbf{M}_1 = \{ m_{11}, m_{12} \};$$

$$\mathbf{M}_2 = \{ m_{21}, m_{22} \};$$

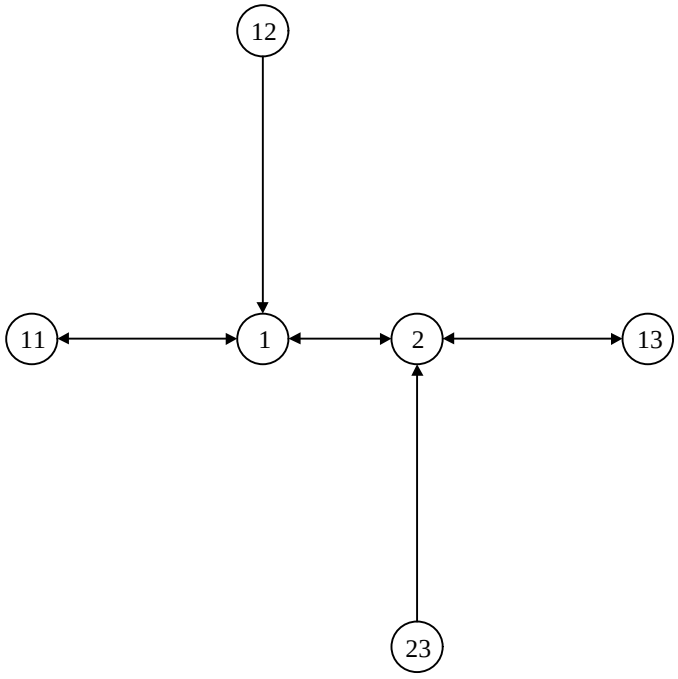
$$\mathbf{P} = \prod_{i=1}^{2} \mathbf{M}_i ;$$

$$= \{ (m_{11}, m_{21}), (m_{11}, m_{22}), (m_{12}, m_{21}), (m_{12}, m_{22}) \}.$$

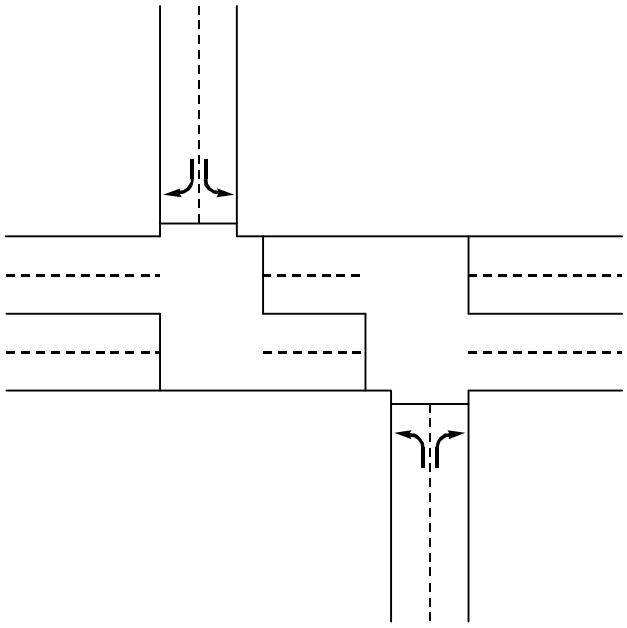
So, a joint phase $p \in \mathbf{P}$ can be a paired-phase of (m_{11} , m_{21}), (m_{11} , m_{22}), (m_{12} , m_{21}), or (m_{12} , m_{22}). Notice that in a single intersection case there will be no difference between a joint phase and an individual phase for the intersection.

Let μ_i and ρ be the cardinal numbers of $\mathbf{M}_i$ and $\mathbf{P}$, respectively. Then, $\mu_1 = 2$, $\mu_2 = 2$, and $\rho = \mu_1 \cdot \mu_2 = 4$. Although this is a network of two intersections each having two unique phases, it can be treated as if it were a single intersection with four unique joint phases. Consequently, the IP model being developed can be formulated and optimized in this manner.

Next, a couple of the so-called "filtering variables" are introduced: ϕ_i^{mp} and θ_{ia}^{m}. That is, whenever a signal control decision is made (i.e., $x_j^{ps} = 1$), it is easy to associate that decision concerning the joint phase *p* that is the j^{th} phase in phase sequence *s* with individual phases of each intersection and with link-specific traffic information by using ϕ_i^{mp} and θ_{ia}^{m}, respectively. For example, if $\phi_i^{mp} \cdot x_j^{ps} = 1$, it indicates that phase *m* of intersection *i* is in joint phase *p* that is the j^{th} phase of phase sequence *s*. Moreover, if



(a)



(b)

Figure 4.1. Two Tight Offset T-Intersections.
(a) Link-Node Diagram; (b) Intersection Layout.

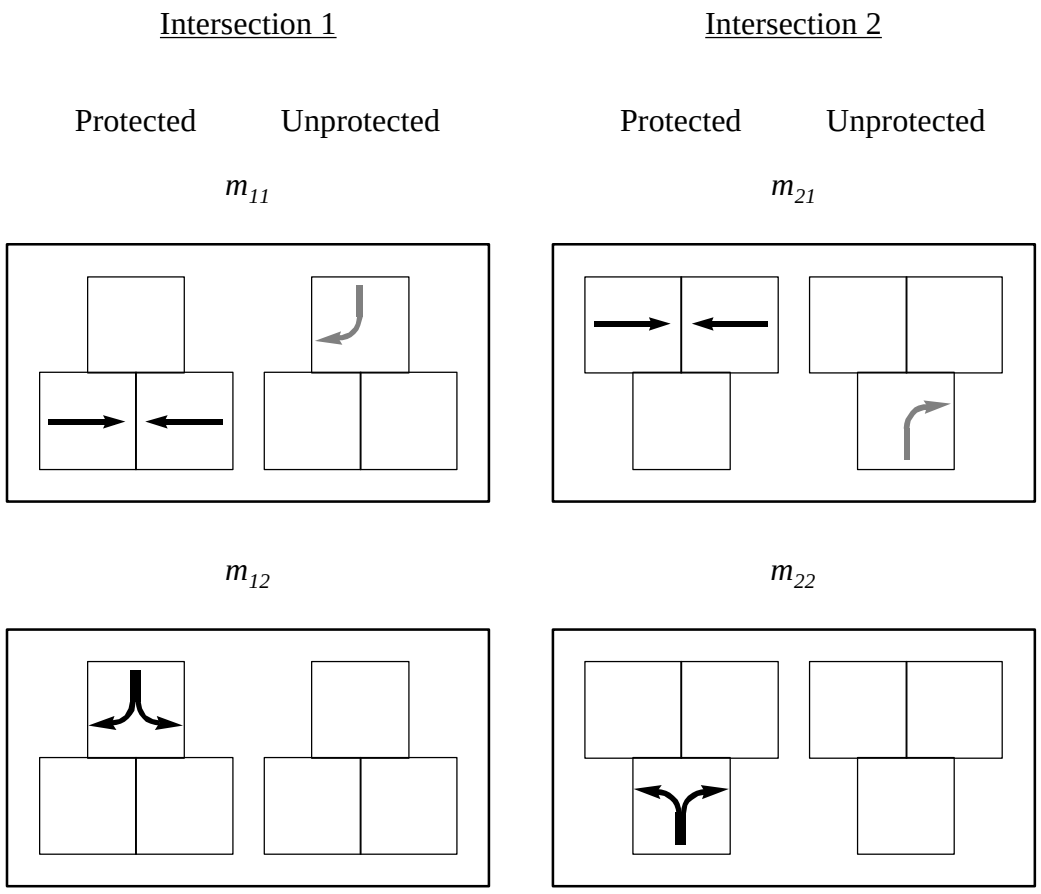


Figure 4.2. Signal Phases for the Case of the Two Tight Offset T-Intersections.

$\theta_{ia}^{m} \cdot \phi_{i}^{mp} \cdot x_{j}^{ps} = 1$, it indicates that link *a* is in phase *m* of intersection *i* and phase *m* is in joint phase *p* that is in the j^{th} phase of phase sequence *s*. From this point on, the reader is expected to familiarize himself or herself with these kinds of expressions. By doing so, it becomes a simple way to model tedious geometric details, intersection movements, signal phases, traffic dynamics, and signal control decisions in a coherent manner.

As far as the computational issue is concerned, ρ can serve as an index to indicate the degree of complexity. For example, a set of *k*-phase sequences with each made up in Φ contains $(\rho\text{-}1)^{k\text{-}1}$ combinations. In the case of the two tight offset T-intersections discussed earlier, the number of combinations is 2187, if $k = 8$. One can soon discover that it can easily grow exponentially as *k* increases. In the next chapter, more discussions on computational issues will continue.

Objective Function

The objective function $F(\cdot)$ inputs geometric data, real-time traffic information, systems information and signal control decisions being tested. In return, it produces calculated performance indices of concern. In this study, the functional form of the objective function is given, which generally assumes the form of $F(\omega, T, \Gamma, \Psi)$ as described earlier.

To produce accurate, realistic and trustworthy performance indices of concern, the object function should be able to model:

- Future traffic arrivals (e.g., vehicles, pedestrians, buses, etc.);
- Different vehicle types (e.g., auto, bus, truck, carpool, etc.);
- Unprotected movements such as permitted left turns and RTORs;

- Vehicular dynamics such as lane-changing, car-following, queue discharge, queue spillover into upstream intersections, start-up, slowing-down, etc.;
- Vehicles reacting to various geometrics, control devices, signal transition, parking activities, incidents, blockages, pedestrians, and other vehicles;
- Signal clearance or transition (i.e., yellow plus all-red, if applicable);
- Transit (bus) operations;
- Parking activities;
- Incidents; and
- Pedestrian traffic.

As reviewed in Chapter Two, the best candidate model that can easily satisfy the aforementioned requirements is a microscopic simulation model. The bottom line is that from the perspective of the optimization problem an objective function is a function. It matters not whether the function has a simple analytic form or is a large simulation model. In this study, the CORSIM microscopic simulation model will be fully exploited to carry out this task. More discussions about how to integrate CORSIM into the optimization model will follow in the next chapter and Chapter Six.

In the concept discussed earlier, it is sufficient to address the signal clearance interval issue on the basis of an individual signal phase for each intersection. However, in the objective function, the signal clearance interval display should be handled on the basis of each movement by considering its involvement from the immediately previous signal phase at the intersection to the current one. In literature, there are general traffic engineering practices regarding amber signal displays, such as the discussion made by Kell and Fullerton (1991), and they should be followed. Table 4.1 summarizes the rules.

Table 4.1. Summary of Movement-Specific Signal Change Interval Displays.

From	To	Change Interval Display
Protected	Prohibited	Yellow + All-Red
Protected	Permitted	Yellow
Protected	Protected	N/A
Permitted	Prohibited	Yellow + All-Red
Permitted (with companioned protected movement being turned off)	Protected	Yellow
Permitted (without companioned protected movement being turned off)	Protected	N/A
Permitted	Permitted	N/A

Constraints

As discussed in Chapter Three, the constraints include the minimum green times required for each phase, maximum green times allowed for each movement-specific phase, and a safety requirement involving signal clearance interval. In addition, there is an obvious constraint concerning the physical storage space of each link, whose capacity cannot be exceeded at any time. Besides, three more housekeeping constraints are recognized as follows:

- Each time there is only one joint phase *p* that can be assigned to a position *j* in a phase sequence *s*;
- The time τ_j^s allocated to each position *j* in a phase sequence *s* should add up to the total length of the time horizon *T*; and
- The decision of a joint phase *p* in the position *j* of a phase sequence *s* is either one (on) or zero (off).

In what follows, the mathematical form of each constraint mentioned above will be presented and discussed in detail.

Due to the idea of using joint phases, some functional manipulations will be necessary to verify the minimum and maximum green constraints on an intersection level. Suppose that joint phase p' be the $(j\text{-}1)^{th}$ phase in phase sequence *s* and that phase m' of intersection *i* be in p', it follows that $\phi_i^{m'p'} \cdot x_{j-1}^{p's} = 1$. Similarly, suppose that joint phase *p* be the j^{th} phase in the phase sequence and phase *m* of intersection *i* be in *p*, it indicates that $\phi_i^{mp} \cdot x_j^{ps} = 1$. Suppose $G_{i,0}^{ms}$ denote the green time for phase *m* of intersection *i* already elapsed since its onset until the end of the immediately previous horizon. Based on the definitions of G_{ij}^{ms} and $\gamma_i^{m'm}$, it is implied that

$$G_{ij}^{ms} = \phi_i^{mp} \cdot x_j^{ps} \cdot \left(G_{i,j-1}^{ms} + \tau_j^s - \phi_i^{m'p'} \cdot x_{j-1}^{p's} \cdot \gamma_i^{m'm} \right); \quad \forall m, m', p, p', s, i, j \tag{4.1}$$

That is, given that phase *m* is in the j^{th} phase, if it is also in the $(j\text{-}1)^{th}$ phase of the phase sequence (i.e., $m' = m$), then $G_{i,j-1}^{ms} \geq G_i^{m,min}$ and $\gamma_i^{m'm} = 0$. In this case, as indicated, $G_{i,j-1}^{ms}$ must already have satisfied the minimum green requirement conceptually. However, this statement can be relaxed for the case of $G_{i,0}^{ms}$ without losing any generality. Otherwise, $G_{i,j-1}^{ms} = 0$. As for $\gamma_i^{m'm}$, it can assume a value of either *Y* or *Y* + *R* depending on phases

m' and m. That is, $\gamma_i^{m'm} = Y$, if the maximum duration of the movement-specific signal change interval is equal to Y considering each movement in phase m' of intersection i transitioning to phase m based on the rules indicated in Table 4.1. Likewise, $\gamma_i^{m'm}$ is equal to the sum of Y and R, if the maximum duration of the signal change interval in such a transition suggests so. Notice that technically phase m can remain green over more than one joint phase in a phase sequence. Therefore, the duration of G_{ij}^{ms} depends on the onset of phase m. On one hand, if phase m starts before the current joint phase, its previous green time will be added to the time that has been allocated to the current joint phase. In which case, no signal transition is necessary. On the other hand, if phase m just starts in the current joint phase, a proper signal clearance interval is required to satisfy the safety constraint. In which case, the duration of the proper signal clearance interval will be subtracted from the time allocated to the current joint phase, if applicable.

By virtue of Equation (4.1), it becomes straightforward to address not only the change interval issue, but also the minimum and maximum green issues. To ensure that G_{ij}^{ms} both satisfies the minimum green time required and is within its maximum green time allowed, immediately it follows that

$$G_i^{m,min} \le G_{ij}^{ms} \le G_i^{m,max}; \quad \phi_i^{mp} \cdot x_j^{ps} = 1; \quad \forall m, p, s, i, j \tag{4.2}$$

To assure that at any time t the number of vehicles on any link a of intersection i does not exceed the physical storage capacity of that link, it makes sense that

$$\pi_{ia}(t) = \pi_{ia}(t\text{-}1) + \sum_{\forall m} \theta_{ia}^{m} \cdot \left[\sum_{\forall v} \alpha_{ia}^{mv}(t) - \beta_{ia}^{m}(t) \right]; \quad \forall a, i, t \tag{4.3}$$

$$\pi_{ia}(t) \le \lambda_{ia}^{max}; \quad \forall a, i, t \tag{4.4}$$

As shown in Equation (4.3), it is obvious that the vehicle number on link *a* at time *t* is equal to the vehicle number at time (*t-1*) plus the difference between the numbers of vehicle arrivals on and departures from the link *a* associated with all phases of the intersection *i* during time *t*. The arrivals and departures are represented by $\alpha_{ia}^{mv}(t)$ and $\beta_{ia}^{m}(t)$, respectively. Notice that θ_{ia}^{m}, as discussed previously, functions like a filter so that only those arrivals and departures associated with the subject link *a* are considered. The information of $\alpha_{ia}^{mv}(t)$ is given, in which different vehicle types are recognized. While for $\beta_{ia}^{m}(t)$, its functional form is given. Generally, it can be stated as follows:

$$\beta_{ia}^{m}(t) = \psi_{ia}^{m}(t) + \phi_{i}^{mp} \cdot x_{j}^{ps} \cdot f_{\beta}(t); \quad t \le \sum_{q=1}^{j} \tau_{q}^{s}; \quad \forall m, p, s, a, i, j, t \tag{4.5}$$

where

$f_{\beta}(t)$ = a queue discharge model based on prevailing traffic conditions associated with function $\beta_{ia}^{m}(t)$ at time *t* (functional form given); and

$\psi_{ia}^{m}(t)$ = a function to estimate the number of vehicle departures from link *a* made by either permitted movements or RTOR vehicles in phase *m* during time *t* (functional form given).

That is, the departures from link *a* associated with phase *m* during time *t* are contributed by either both queue discharges and permitted movements if the phase is green, or RTORs if the phase is not. On one hand, usually the queue discharge model $f_{\beta}(t)$ also addresses issues such as start-up lost times, slow-down lost times due to vehicles reacting to an amber phase, the other vehicles in the queue, the so-called “jumper” and “sneaker” vehicles, and pedestrian interruptions. On the other hand, generally the function $\psi_{ia}^{m}(t)$

represents a gap acceptance model, which considers an unprotected vehicle seeking an acceptable gap in its opposing traffic stream like a permitted left turn and RTOR. While for the inequality of $t \le \sum_{q=1}^{j} \tau_q^s$ presented in Equation (4.5), it is to make sure that the time *t* referenced is associated with the time when the decision x_j^{ps} is made. Finally, Equation (4.4) is a direct result from the above discussion.

To ensure that there is only one joint phase *p* assigned to the j^{th} phase in phase sequence *s*, it follows that

$$\sum_{\forall p} x_j^{ps} = 1; \quad \forall s, j \tag{4.6}$$

Naturally, the times allocated to each phase in a phase sequence should add up to the length of time horizon *T*. That is,

$$\sum_{\forall j} \tau_j^s = T; \quad \forall s \tag{4.7}$$

Finally, to essentially state that $x_j^{ps} \in \{0,1\}$, it follows that

$$x_j^{ps} \cdot \left(1 - x_j^{ps}\right) = 0; \quad \forall p, s, j \tag{4.8}$$

Summary

This chapter first introduces the necessary notations and assumptions, and then presents the idea of using joint phases in the mathematical model formation. With the objective function in form of $F(\omega, T, \Gamma, \Psi)$ and a set of the constraints specified in Equations (4.1)-(4.8), the general IP model formulation for the commonly recognized network-wide signal optimization problem in dynamic urban traffic control can be summarized and stated as shown in Equation (4.9):

$$
\begin{aligned}
&\underset{\omega \in \mathbf{\Omega}}{\text{Min}} && z = F\left(\omega, T, \mathbf{\Gamma}, \mathbf{\Psi}\right) && \\
&\textit{st.} && && \\
& && G_{ij}^{ms} = \phi_i^{mp} \cdot x_j^{ps} \cdot \left(G_{i,j-1}^{ms} + \tau_j^s - \phi_i^{m'p'} \cdot x_{j-1}^{p's} \cdot \gamma_i^{m'm}\right); && \forall m, m', p, p', s, i, j \\
& && G_i^{m,min} \le G_{ij}^{ms} \le G_i^{m,max}; \quad \phi_i^{mp} \cdot x_j^{ps} = 1; && \forall m, p, s, i, j \\
& && \pi_{ia}(t) = \pi_{ia}(t-1) + \sum_{\forall m} \theta_{ia}^m \cdot \left[\sum_{\forall v} \alpha_{ia}^{mv}(t) - \beta_{ia}^m(t)\right]; && \forall a, i, t \\
& && \pi_{ia}(t) \le \lambda_{ia}^{max}; && \forall a, i, t \\
& && \beta_{ia}^m(t) = \psi_{ia}^m(t) + \phi_i^{mp} \cdot x_j^{ps} \cdot f_\beta(t); \quad t \le \sum_{q=1}^{j} \tau_q^s; && \forall m, p, s, a, i, j, t \\
& && \sum_{\forall p} x_j^{ps} = 1; && \forall s, j \\
& && \sum_{\forall j} \tau_j^s = T; && \forall s \\
& && x_j^{ps} \cdot \left(1 - x_j^{ps}\right) = 0; && \forall p, s, j \\
& && G_{ij}^{ms}, \tau_j^s, \pi_{ia}(t), \beta_{ia}^m(t), \psi_{ia}^m(t), f_\beta(t) \ge 0 && \forall m, p, s, a, i, j, t
\end{aligned}
\tag{4.9}
$$

However, there are situations where maximizing the objective function is desired. In which cases, Equation (4.9) can still be used. Note that over any region

$$\text{maximum } z = -\text{ minimum } -z \tag{4.10}$$

That is, a maximization problem can be easily converted into a minimization problem by multiplying the coefficient of the objective function by -1. After the optimization of the new problem is completed, the objective of the old problem is -1 times the optimal objective of the new problem.

CHAPTER FIVE
SOLUTION ALGORITHMS FOR THE NETWORK-WIDE SIGNAL OPTIMIZATION PROBLEM

Introduction

As discussed in Chapter Three, it is suggested that the modeling framework of network-wide signal optimization for the distributed signal control system should consist of two major components: a signal optimization module and a microscopic simulation module. Further, Chapter Four emphasizes the need to realistically calculate performance indices of concern by considering real-world traffic scenarios in the network-wide signal optimization problem. Based on the review in Chapter Two, it is suggested that using a microscopic simulation model be the best way to satisfy that need, since there is no doubt that microscopic simulation is playing an increasingly important role in network-wide signal optimization.

In this chapter, a solution framework is developed to solve the commonly recognized network-wide signal optimization problem that has been formulated in Chapter Four as a general IP problem. The solution framework intimately combines a dynamic programming (DP) optimization procedure and a microscopic simulation model. As mentioned in Chapter Three, a special version of the CORSIM microscopic simulation model is used to carry out the task required for the microscopic simulation module. In what follows, the DP solution framework is introduced first, followed by the presentation of a numerical example to illustrate the DP calculations. A heuristic search

procedure is also developed and it can significantly reduce the DP computation in network optimization.

A DP Solution Framework

In this section, the concept and features of the DP solution framework are introduced first, followed by the DP model formulation. Then, the procedures to calculate the DP optimal value function and optimal policy function are presented, followed by the discussion of the computational efficiency of the DP solution algorithm.

Concept

DP is an optimization procedure that is particularly applicable to problems requiring a sequence of interrelated decisions (Dreyfus & Law, 1977). The essence of DP is Richard Bellman's *Principle of Optimality* (Bellman & Dreyfus, 1962), which states as follows:

> *An optimal policy has the property that whatever the initial state and the initial decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.*

In DP, generally one variable describes how many decisions have thus far been made. Depending on its context, it is either increased or decreased by one after each decision. This particular monotonic variable is called the stage variable. All the remaining variables needed to describe the current situation, given the stage variable, are called state variables. The values of the stage and state variables constitute a description of the situation adequate to allow a DP solution.

Figure 5.1 depicts a graphical representation of the general IP problem formulated in Chapter Four. As defined previously, ρ is the cardinal number of P, which denotes a

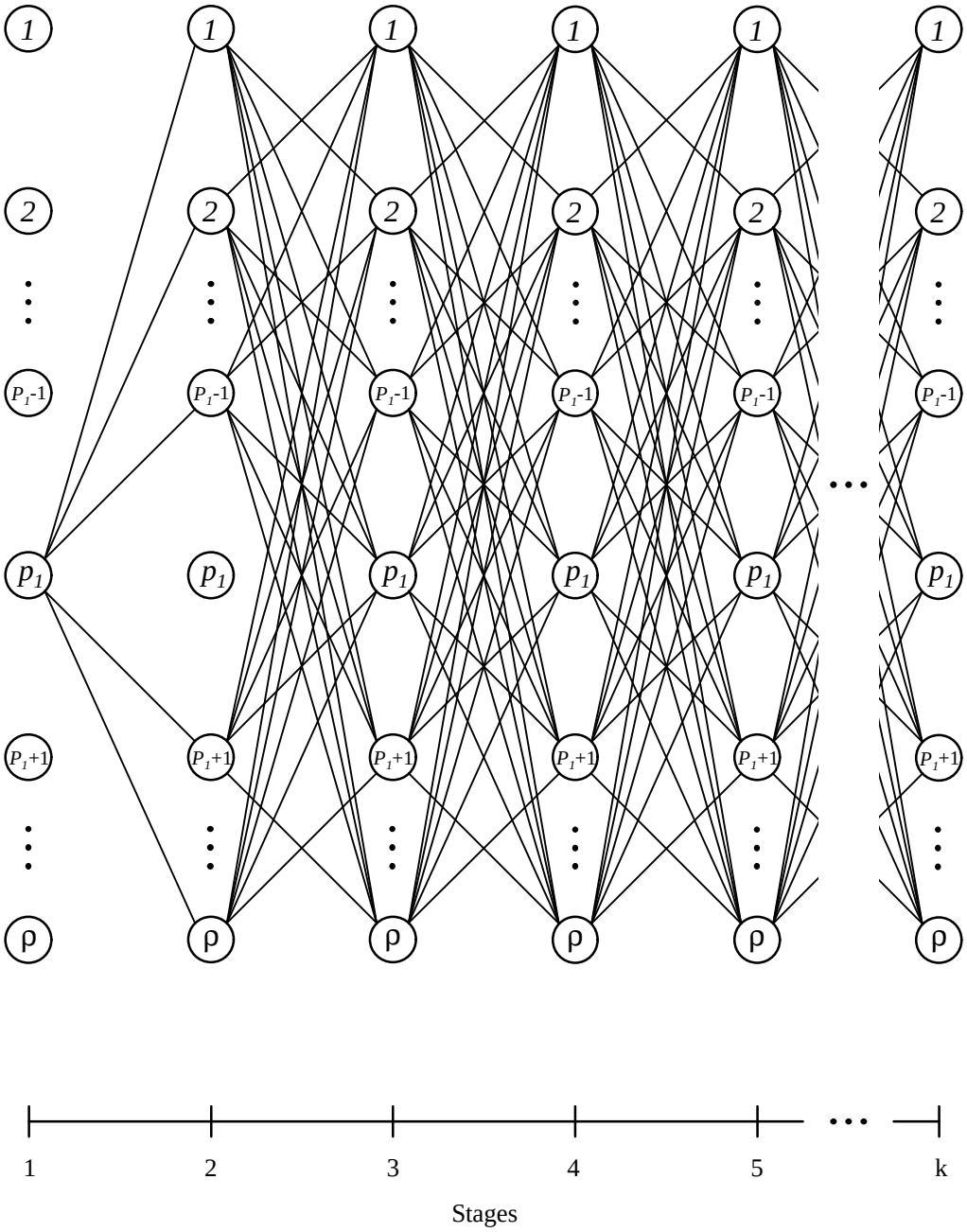


Figure 5.1. A Graphical Representation of the General Network-wide Signal Optimization Problem Formulation Developed in Chapter Four (ρ-Phase Operations over a *k*-Phase Sequence).

set of joint phases for all intersections in the network. Suppose joint phase $p_1 \in \mathsf{P}$ be the initial (i.e., first) phase in a *k*-phase sequence, the subsequent decisions, given the real-time information and specified objectives and constraints, are to select from P enough number of joint phases for the remaining of the *k*-phase sequence so that a global optimality can be achieved. That is, for the second phase in the sequence a joint phase will be chosen from the remaining (ρ-1) joint phases in P (i.e., all except p_1). Similarly, for the third phase in the sequence another joint phase will be selected from all joint phases in P excluding the one already assigned to the second phase; and so forth. In this manner, the optimization process works phase by phase towards the end of time horizon *T* and "smartly" explores all possibilities along the process by eliminating infeasible solutions, as indicated in Figure 5.1. Due to the nature of such an interrelated decision-making process, it perfectly lends itself to the DP solution framework. Using the DP language, naturally joint phases in the *k*-phase sequence are stages in the DP calculation. Also, it will be necessary to employ a set of state variables to keep track of both time that has been allocated to each stage and total time that has already elapsed in the current phase of each intersection associated with the stage variable. By virtue of the latter, for each intersection in the traffic network being considered, not only the maximum and minimum green constraints can be verified, but also a signal clearance interval can be assured being properly implemented between any two distinct phases.

One may soon recognize that in the aforementioned DP optimization process the number *k* itself is also a variable. Its value is so determined as long as the following equality is satisfied:

$$\sum_{j=1}^{k} \tau_j^s = T; \quad \forall s \in S_k \tag{5.1}$$

where τ_j^s, as defined previously, is one of the control variables denoting the time that has been allocated to the j^{th} phase in phase sequence $s \in S_k$. Notice that τ_j^s includes the duration of a change interval (i.e., yellow time plus all-red time) for the corresponding phases of each intersection in the traffic netowrk, if applicable. The other control variable is x_j^{ps}, or p_j^s, denoting the joint phase assigned to the j^{th} phase in phase sequence *s*. As defined previously, x_j^{ps} = 1, if joint phase *p* is the j^{th} phase in phase sequence $s \in S_k$; and 0, otherwise.

With the above concept in mind, it is ready to lay out the DP model formulation. However, before doing so, let's first identify the characteristics of the DP solution algorithm.

Features

Like OPAC, PRODYN and COP, the DP solution algorithm is also implemented based on the rolling-horizon framework. However, unlike its counterparts, it has the following unique features:

- Network-wide signal optimization performed;
- Global optimality achieved;
- More computationally efficient;
- Signal phasing and timing considered simultaneously;
- Signal coordination automatically accomplished based on network-wide objectives;

- An initial phase sequence not required;
- Microscopic simulation model incorporated;
- Real-world traffic scenarios considered; and
- Performance indices realistically calculated.

Model Formulation

In the following discussion, the terminology used by Dreyfus & Law (1977) is adopted. First, more notations, in addition to those that already defined in Chapter Four, are developed to facilitate the DP concept introduced earlier. Notice that the reader may refer to Appendix A for a summary of notations. Next, the range of values that control variable τ_j can assume is discussed, since it is essential to the DP calculation. Then, the recurrence relation and boundary conditions for the forward dynamic programming (FDP) are presented. As far as the optimal policy function for the backward dynamic programming (BDP) is concerned, it will be discussed later.

Notice that for the reason of simplicity hereinafter the superscript *s* may be dropped from those variables that are defined in Chapter Four. For example, control variables x_j^{ps} and τ_j^s may be simply written as x_j^p and τ_j, respectively; variable G_{ij}^{ms} as G_{ij}^m; etc. Based on their context, they may be used interchangeably.

Definitions and Assumptions:

p_j = j^{th} phase, $p_j \in \mathsf{P}$;

τ_j^{min} = minimum value that τ_j can assume;

τ_j^{max} = maximum value that τ_j can assume;

T_j = total number of discrete time intervals that have already been allocated after the completion of the j^{th} phase;

$f_j\left(p_{j-1}, p_j, T_j, \tau_j\right)$ = incremental performance index from joint phase p_{j-1} to p_j associated with state variable T_j and control variable τ_j;

$F_j\left(p_j, T_j\right)$ = optimal value function denoting the optimal performance index after the completion of the j^{th} phase p_j and the total number of discrete time intervals T_j that have been allocated;

F^* = global optimal performance index.

Range of Control Variable τ_j:

Based on the definitions of state variable T_j and control variable τ_j, it follows that

$$T_j = T_{j-1} + \tau_j \tag{5.2}$$

where $T_0 = 0$. Figure 5.2 depicts the above relationship.

Due to the minimum and maximum green constraints, control variable τ_j can only assume values in the range of [τ_j^{min}, τ_j^{max}], where

$$\begin{cases} \tau_j^{min} = \underset{\forall i}{\text{Max}} \left\{ \left(1 - \phi_i^{mp'} \cdot x_{j-1}^{p'}\right) \cdot \left(G_i^{m,min} + \gamma_i^{m'm}\right) \right\} \\ \tau_j^{max} = \underset{\forall i}{\text{Min}} \left\{ \phi_i^{mp'} \cdot x_{j-1}^{p'} \cdot \left(G_i^{m,max} - G_{i,j-1}^{m}\right) + \left(1 - \phi_i^{mp'} \cdot x_{j-1}^{p'}\right) \cdot \left(G_i^{m,max} + \gamma_i^{m'm}\right) \right\} \end{cases} \tag{5.3}$$

where $\phi_i^{mp} \cdot x_j^p = 1$ and $\phi_i^{m'p'} \cdot x_{j-1}^{p'} = 1$ for every *i*, *m*, *m′*, *p* and *p′*, given *j*.

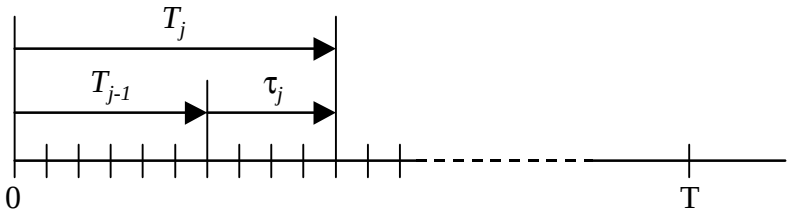


Figure 5.2. Relationship of Control Variable τ_j and State Variables T_j and T_{j-1}.

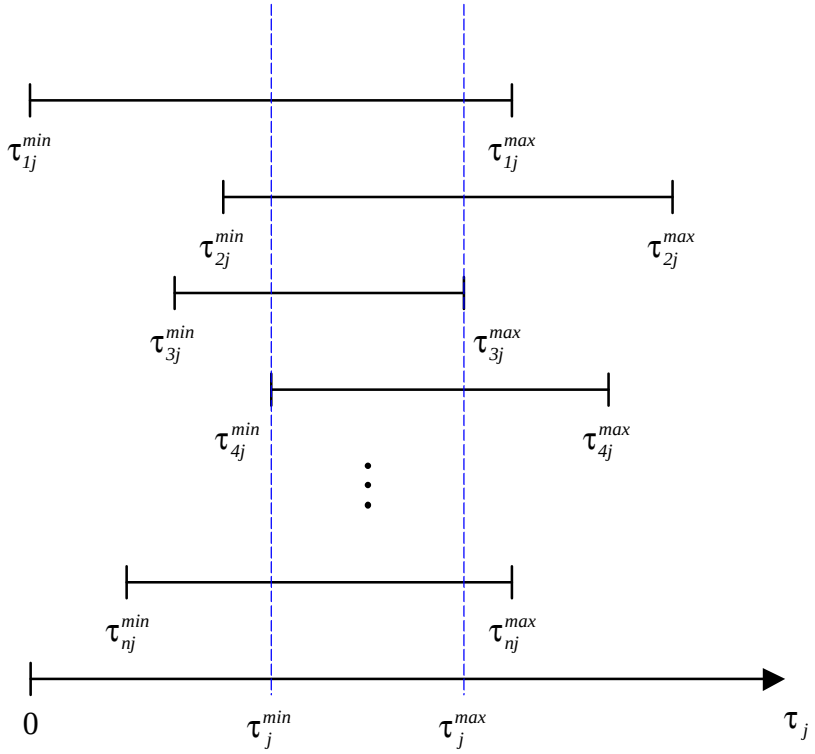


Figure 5.3. A Graphical Determination for the Range of Control Variable τ_j.

Figure 5.3 depicts the idea of Equation (5.3), whose derivation follows. Given *j*, for every *i*, *m*, m', *p* and p' such that $\phi_i^{mp} \cdot x_j^p = 1$ and $\phi_i^{m'p'} \cdot x_{j-1}^{p'} = 1$, Equation (4.1) can be simplified as

$$G_{ij}^m = G_{i,j-1}^m + \tau_j - \gamma_i^{m'm} \tag{5.4}$$

where G_{ij}^m satisfies the minimum and maximum green constraints as follows:

$$G_i^{m,min} \le G_{ij}^m \le G_i^{m,max} \tag{5.5}$$

It is obvious that $p = p'$ for $j = 1$, and $p \ne p'$ for $j > 1$. Since $p_1 = p_0$, therefore either $\phi_i^{m'p_0} \cdot x_0^{p_0} = 1$ or $\phi_i^{m'p_1} \cdot x_0^{p_1} = 1$ indicates the last phase in the immediately previous horizon. Rearranging Equations (5.4) and (5.5) for τ_j, it follows that

$$G_i^{m,min} - G_{i,j-1}^m + \gamma_i^{m'm} \le \tau_j \le G_i^{m,max} - G_{i,j-1}^m + \gamma_i^{m'm} \tag{5.6}$$

For the sake of discussion, let τ_{ij}^{min} and τ_{ij}^{max} be the left- and right-hand-side, respectively, of τ_j in Equation (5.6) for intersection *i*. Then, depending on m', two cases are identified: $m' = m$ or $m' \ne m$.

First, if $m' = m$, it indicates that $G_{i,j-1}^m \ge G_i^{m,min}$ and $\gamma_i^{m'm} = 0$ as discussed in Chapter Four. Since $G_i^{m,min} - G_{i,j-1}^m \le 0$ in this case, therefore $\tau_{ij}^{min} = 0$ due to the non-negative requirement of τ_j. Thus,

$$\begin{cases} \tau_{ij}^{min} = 0 \\ \tau_{ij}^{max} = G_i^{m,max} - G_{i,j-1}^m \end{cases} \tag{5.7}$$

On the other hand, if $m' \ne m$, then $G_{i,j-1}^m = 0$ and $\gamma_i^{m'm}$ can assume a value of either *Y* or *Y* + *R* depending on phases m' and *m* as discussed in Chapter Four. Thus,

$$\begin{cases} \tau_{ij}^{min} = G_i^{m,min} + \gamma_i^{m'm} \\ \tau_{ij}^{max} = G_i^{m,max} + \gamma_i^{m'm} \end{cases} \tag{5.8}$$

By definition, τ_j^{min} is the minimum value that τ_j can assume. It is so determined that the minimum green requirement for each intersection should also be satisfied simultaneously. Thus, the maximum τ_{ij}^{min} over all intersections should be selected for τ_j^{min}. Similarly, τ_j^{max} is the maximum value that τ_j can assume. Therefore, the minimum τ_{ij}^{max} over all intersections should be selected for τ_j^{max} so that the maximum green requirement for each intersection can also be satisfied simultaneously. That is,

$$\begin{cases} \tau_j^{min} = \underset{\forall i}{\text{Max}} \left\{ \tau_{ij}^{min} \right\} \\ \tau_j^{max} = \underset{\forall i}{\text{Min}} \left\{ \tau_{ij}^{max} \right\} \end{cases} \quad \forall j \tag{5.9}$$

Given that $\phi_i^{mp} \cdot x_j^p = 1$ and $\phi_i^{m'p'} \cdot x_{j-1}^{p'} = 1$, the product of $\phi_i^{mp'}$ and $x_{j-1}^{p'}$ can be either zero or one; i.e., $\phi_i^{mp'} \cdot x_{j-1}^{p'} = \{0, 1\}$. If $\phi_i^{mp'} \cdot x_{j-1}^{p'} = 1$, it indicates the first case (i.e., $m' = m$); otherwise, the second case (i.e., $m' \neq m$). Hence, Equation (5.3) can be achieved by directly integrating Equations (5.7) and (5.8) into (5.9).

The beauties of having the range of control variable τ_j are threefold: First, with Equation (5.3) control variable τ_j will have a much smaller number of values to assume, compared to a brute-force search. Consequently, the DP calculation can be reduced. Second, it can be easily determined if a joint phase *p* can be assigned to the j^{th} phase. That is, given *j*, *p* and p' for every *i*, *m* and m' such that $\phi_i^{mp} \cdot x_j^p = 1$ and $\phi_i^{m'p'} \cdot x_{j-1}^{p'} = 1$, a joint phase *p* should be rejected if Equation (5.3) yields a result suggesting that τ_j^{min} is actually greater than τ_j^{max}. By doing so, the DP calculation can be further reduced by

concentrating only on those feasible joint phases. Third, the DP calculation can be even further reduced by eliminating the impossible combinations of state variables T_j and T_{j-1} given control variable $\tau_j \in [\tau_j^{min}, \tau_j^{max}]$ in Equation (5.2).

Recurrence Relation:

With Equations (5.2) and (5.3), the recurrence relation for the FDP can be presented as follows:

$$F_j(p_j, T_j) = \underset{\forall p_{j-1}}{\text{Min}} \left\{ \begin{array}{l} f_j(p_{j-1}, p_j, T_j, \tau_j) + F_{j-1}(p_{j-1}, T_{j-1}): \\ p_{j-1} \neq p_j, \forall j \geq 2; \tau_j \in [\tau_j^{min}, \tau_j^{max}], T_j = T_{j-1} + \tau_j \in [1, T], \forall j \geq 1 \end{array} \right\} \quad (5.10)$$

where $f_j(p_{j-1}, p_j, T_j, \tau_j)$ is a function to calculate an incremental performance index. However, $f_j(p_{j-1}, p_j, T_j, \tau_j) = 0$, if both state variables T_j and T_{j-1} are all equal to the duration of time horizon *T*. Notice that it is obvious that $p_1 = p_0$.

Boundary Conditions:

Naturally, the following value of the optimal value function is obvious, since the optimum performance index at the very beginning of a time horizon (i.e., T_0) is assumed zero:

$$F_0(p_0, T_0) = 0 \quad (5.11)$$

Calculation of the Optimal Value Function

With Equations (5.2), (5.3), (5.10) and (5.11), the FDP to calculate the optimal value function can be written as follows:

1. Given $p_1 = p_0$, $T_0 = 0$ and Equation (5.11), initialize stage $j = 1$.

2. For every $p_j \in \mathsf{P}$ applicable at stage *j*, use Equation (5.3) to calculate the range $[\tau_j^{min}, \tau_j^{max}]$ of control variable τ_j for every $p_{j-1} \in \mathsf{P}$ at stage (*j-1*) switching to the subject p_j, where $p_{j-1} = p_j$ if stage $j = 1$, or $p_{j-1} \neq p_j$ if stage $j > 1$. If no feasible p_j can be found (i.e., $\tau_j^{min} > \tau_j^{max}$ for every p_{j-1} at stage (*j-1*) switching to each applicable p_j at stage *j*, assuming that minimum green times, maximum green times, yellow time and all-red time have been properly set), proceed to Step 6. Otherwise, mark off those infeasible p_j at stage *j* and proceed to the next step with feasible ones.
3. For every feasible p_j at stage *j* identified in Step 2, calculate the range of state variable T_j in $[1, T]$ for every p_{j-1} at stage (*j-1*) switching to the subject p_j based on the relationship specified in Equation (5.2), given the range of control variable τ_j calculated in Step 2 and each value of state variable T_{j-1} calculated at stage (*j-1*) that can generate an optimal value of $F_{j-1}(p_{j-1}, T_{j-1})$ at stage (*j-1*). If no feasible T_j can be found, or the only value that both state variables T_j and T_{j-1} can assume is the time horizon *T* for every p_{j-1} at stage (*j-1*) switching to each feasible p_j identified in Step 2, proceed to Step 6. Otherwise, mark off those infeasible values of T_j for every p_{j-1} switching to each feasible p_j at stage *j* and proceed to the next step with feasible ones.
4. For every feasible combination of p_j and T_j at stage *j* identified in Step 2 and 3, respectively, solve the problem of Equation (5.10) by considering the set of

p_{j-1} at stage (*j-1*) switching to the subject p_j. Record each set of p_{j-1} and τ_j for stage *j* that can generate an optimal value of $F_j(p_j, T_j)$ for the problem just solved with respect to the subject combination of p_j and T_j.

5. Repeat Steps 2, 3 and 4 by increasing stage *j* by one each time until no more problems of Equation (5.10) can be solved.
6. Identify the highest number of stage *j* (i.e., j^*), and calculate the global optimal performance index F^* as follows:

$$F^* = \text{Min}\left\{ F_j(p_j, T) \,\middle|\, \forall p_j \in \mathsf{P},\ j = j^* \right\} \tag{5.12}$$

The usage of Equation (5.3) requires the knowledge of $G^m_{i,j-1}$, which denotes the green time for phase *m* of intersection *i* elapsed since its onset until the end of the (*j-1*)th phase. Thus, in the FDP the state of each $G^m_{i,j-1}$ needs to be kept updating for each stage *j*.

As mentioned earlier, a special version of the CORSIM microscopic simulation model is used to calculate a performance index for every instance raised in the problem of Equation (5.10). However, unlike the incremental fashion presented in Equation (5.10), CORSIM is launched every time starting from the very beginning of a time horizon for two reasons: first, the total duration of a time horizon is usually in a few minutes and CORSIM can finish such a simulation run within a very short period of time; and second, all the optimal DP decisions in earlier stages have already been recorded and they are readily available for the subsequent CORSIM simulation runs. In addition, by doing so the following two obstacles can be avoided:

- Storage space problem. In the DP calculation, there are quite a few stages and states of intermediate information that need to be saved so that they can be

processed in later stages due to the recursive nature of DP. Correspondingly there will be tremendous simulation data that would need to be stored and processed in the meanwhile. The storage space problem that results from evaluating a performance index incrementally in simulation makes such an implementation simply impractical.

- Simulation problem. Technically speaking, it is very difficult to evaluate a performance index incrementally in simulation due to some software implementation issues, particularly for CORSIM, unless one would like to keep track of every detail of simulation and store it for later usage. However, by doing so he or she will soon run into the storage space problem mentioned earlier.

By taking advantage of incorporating the CORSIM simulation model into the DP solution framework, not only all the requirements of an objective function as suggested in Chapter Four, but also the movement-specific change interval issue discussed in the same chapter can be addressed. As a result, a suite of performance indices of concern can be both realistically and trustworthily assessed. More discussions on the detail of how to incorporate CORSIM into the DP solution framework will follow in the next chapter.

Calculation of the Optimal Policy Function

With the FDP completed, the calculation of the BDP for the optimal policy function can proceed as follows:

1. Locate from the FDP the stage j^* and the control variable $p_{j^*}^*$ with the global optimal performance index F^*.

2. Start backward by subtracting stage *j* by one each time after retrieving optimal decisions of control variables p_j^* and τ_j^* determined in the FDP.
3. Repeat Step 2 until stage $j = 0$.
4. Construct all the results in the BDP for the optimal policy.

Computational Efficiency

The computational efficiency of the DP calculation depends on a number of factors. Several key factors are listed below:

- Cardinal number of P, i.e., ρ;
- Cardinal numbers of M_i for each intersection *i*, i.e., μ_i, $\forall i$;
- Total length of time horizon *T*;
- Minimum green times $G_i^{m,min}$;
- Maximum green times $G_i^{m,max}$; and
- Duration of the change interval, i.e., $Y + R$.

Let $G^{min} = G_i^{m,min}$ and $G^{max} = G_i^{m,max}$ for every phase *m* of each intersection *i*. For the sake of simplicity, let $G^{max} = T$ so that it will be unnecessary to consider G^{max} here. One way to demonstrate the computational efficiency of the DP solution algorithm is form a contrast by first showing the results of not using the algorithm developed in this chapter. For this purpose, the so-called brute-force enumeration, or E_{BF}, has been derived, which can be expressed as follows (the reader may refer to Appendix C for its detailed derivation):

$$E_{BF} = \sum_{k=2}^{k^{max}} (\rho - 1)^{k-1} \cdot \begin{pmatrix} T - (k-1) \cdot (G^{min} + Y + R - 1) \\ k - 1 \end{pmatrix} \tag{5.13}$$

where

k^{max} = the integer portion of $\left(\frac{T}{G^{min} + Y + R} + 1 \right)$; and

ρ = $\prod_{i \in N} \mu_i$.

Suppose $T = 20$, $Y + R = 1$, $G^{min} = 2$, and $\rho = 3$, k^{max} can be determined by calculating the integer portion of the value $\left(\frac{20}{2+1} + 1 \right)$, which is 7. Thus, based on Equation (5.13) the total number of the brute-force enumeration in this case is:

$$E_{BF} = 2 \cdot \binom{18}{1} + 2^2 \cdot \binom{16}{2} + 2^3 \cdot \binom{14}{3} + 2^4 \cdot \binom{12}{4} + 2^5 \cdot \binom{10}{5} + 2^6 \cdot \binom{8}{6} = 21{,}204 \quad (5.14)$$

Using the DP solution algorithm developed in this chapter, of course, the number of calculations will be much smaller than the one indicated in Equation (5.14), since a significant number of infeasible solutions will be eliminated along the DP optimization process. To be specific, more computational experiences will be discussed by virtue of a numerical example that will be presented in the next section.

A Numerical Example

To illustrate the DP solution algorithm, let's revisit the numerical example used by Sen & Head (1997). In what follows, the description of the example is presented first, followed by the DP model construction of the example and some DP sample calculations. This section is concluded with a discussion on the computational experiences gained in exercising this example.

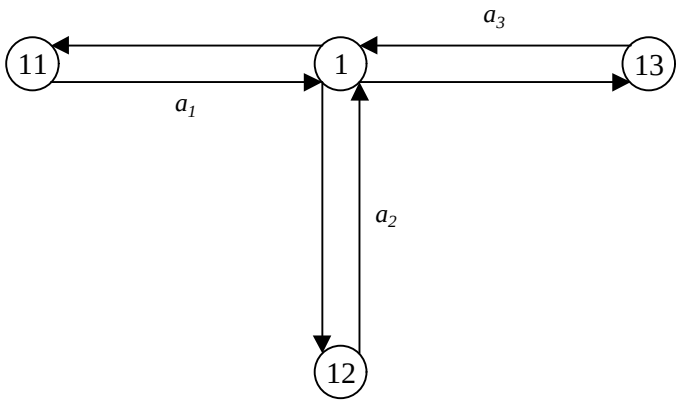
Description

Consider a hypothetical T-intersection, as shown in Figure 5.4. It has three phases, as indicated in Figure 5.5. Figure 5.6 illustrates a graphical representation of its three-phase operations. For the purpose of this example, it is assumed that the initial queue lengths are all zeros and that the initial phase with right-of-way (ROW) is phase m_3. The hypothetical arrival data used in this example is listed in Table 5.1.

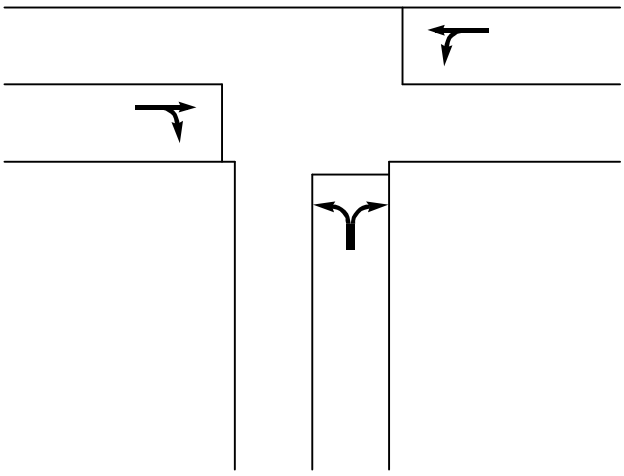
Table 5.1. Arrival Data for the Hypothetical T-Intersection Case.

Time	m_1	m_2	m_3
1	0	0	1
2	0	0	1
3	0	0	0
4	0	1	0
5	0	1	0
6	1	1	0
7	1	0	0
8	1	1	0
9	0	1	0
10	0	0	0

The goal is to minimize total delay over a time horizon T of 10 time units. It is assumed that the signal clearance time is one time unit and the minimum green time for any phase is two time units. For the sake of this example, it is assumed that the initial phase (i.e., m_3) has already satisfied the minimum green requirement. To keep the illustration simple, it is further assumed that vehicles follow instantaneous queue clearance once discharged.



(a)



(b)

Figure 5.4. Hypothetical T-Intersection Case.
(a) Link-Node Diagram; (b) Intersection Layout.

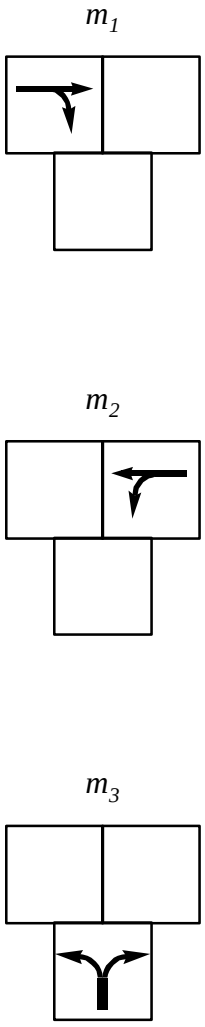


Figure 5.5. Signal Phases Used in the Hypothetical T-Intersection Case.

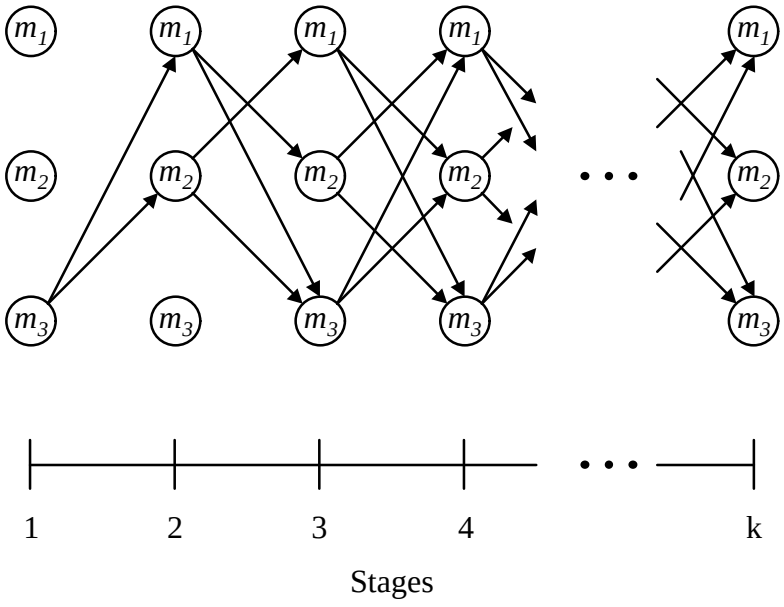


Figure 5.6. A Graphical Illustration of the Three-Phase Operations in the Hypothetical T-Intersection Case.

DP Model Construction

The DP solution framework can be constructed accordingly based on the aforementioned description of the numerical example. Notice that since this is a case of a single intersection for the sake of simplicity the intersection index *i* may be dropped in the DP model construction.

Systems Variables:

$T = 10;$

$Y + R = 1;$

Sets:

$\mathsf{N} = \{ 1 \};$

$\mathsf{A} = \{ a_1, a_2, a_3 \};$

$\mathsf{M} \equiv \mathsf{P} = \{ m_1, m_2, m_3 \};$

$\mathsf{V} = \{ \text{auto} \};$

Signal Control Variables:

$p_1 = m_3 \; (= p_0);$

$\tau_0 \geq G^{min};$

$G^{min} = 2;$

$G^{max} = G^{min} + T;$

$\phi^{mp} : \phi^{m_1 m_1} = \phi^{m_2 m_2} = \phi^{m_3 m_3} = 1;$

$\phi^{m_1 m_2} = \phi^{m_1 m_3} = \phi^{m_2 m_1} = \phi^{m_2 m_3} = \phi^{m_3 m_1} = \phi^{m_3 m_2} = 0;$

$\theta_a^m : \theta_{a_1}^{m_1} = \theta_{a_2}^{m_2} = \theta_{a_3}^{m_3} = 1;$

$$\theta_{a_1}^{m_2} = \theta_{a_1}^{m_3} = \theta_{a_2}^{m_1} = \theta_{a_2}^{m_3} = \theta_{a_3}^{m_1} = \theta_{a_3}^{m_2} = 0;$$

$\gamma^{m'm}$: $\gamma^{m_1m_2} = \gamma^{m_1m_3} = \gamma^{m_2m_1} = \gamma^{m_2m_3} = \gamma^{m_3m_1} = \gamma^{m_3m_2} = Y + R;$

$$\gamma^{m_1m_1} = \gamma^{m_2m_2} = \gamma^{m_3m_3} = 0;$$

Link-Specific Traffic Variables:

$\alpha_a^m(t)$: $\alpha_{a_1}^{m_1}(t) = 1$, if $t = 6$, 7 or 8; 0, otherwise;

$\alpha_{a_2}^{m_2}(t) = 1$, if $t = 4$, 5, 6, 8 or 9; 0, otherwise;

$\alpha_{a_3}^{m_3}(t) = 1$, if $t = 1$ or 2; 0, otherwise;

λ_a^{max} = unlimited, $\forall a \in \mathsf{A}$;

$\pi_a(0) = 0, \forall a \in \mathsf{A}$;

Rest of the real-time information:

$$\xi_a(t) = \upsilon_a(t) = \zeta_a(t) = \eta_a(t) = \delta_a(t) = 0, \forall a \in \mathsf{A}, \forall t;$$

Functions:

$F(\cdot)$ = objective function that estimates total intersection delay.

Based on the assumption for the purpose of this numerical example that vehicles follow instantaneous queue clearance once discharged, a simple accounting procedure is employed to calculate the objective function $F(\cdot)$ by keeping track of vehicle arrivals $\alpha_a^m(t)$ and departures $\beta_a^m(t)$ and total number of vehicles $\pi_a(t)$ in the network for every time *t*. As a result, detailed vehicle dynamics such as car-following, lane-changing, etc., will not be captured, modeled and considered.

DP Sample Calculations

The FDP starts with the initial phase m_3 as stage 1, i.e., $j = 1$. Subsequently, the DP calculations by stages are summarized in Tables 5.2-5.6. The reader may refer to Appendix D for the detailed calculations.

The global optimal performance index F^* of this case is 8 with $j^* = 3$ and the following optimal policy:

$$\begin{cases} p_1 = m_3; \quad \tau_1^* = 3 \\ p_2 = m_2; \quad \tau_2^* = 4 \\ p_3 = m_1; \quad \tau_3^* = 3 \end{cases} \tag{5.15}$$

Table 5.2. Numerical Example of the Hypothetical T-Intersection Case: Stage 1 Calculations [$p_1 = m_3$].

T_1	τ_1^*	F_1	$\pi_{a_1}^{m_1}$	$\pi_{a_2}^{m_2}$	$\pi_{a_3}^{m_3}$	p_0
1	1	1	0	0	1	m_3
2	2	1	0	0	1	m_3
3	3	0	0	0	0	m_3
4	4	1	0	1	0	m_3
5	5	3	0	2	0	m_3
6	6	7	1	3	0	m_3
7	7	12	2	3	0	m_3
8	8	19	3	4	0	m_3
9	9	27	3	5	0	m_3
10	10	35	3	5	0	m_3

Table 5.3. Numerical Example of the Hypothetical T-Intersection Case: Stage 2 Calculations [$p_2 = m_1$].

T_2	τ_2^*	F_2	$\pi_{a_1}^{m_1}$	$\pi_{a_2}^{m_2}$	$\pi_{a_3}^{m_3}$	p_1
7	3	10	1	3	0	m_3
8	3	14	1	4	0	m_3
9	4	18	0	5	0	m_3
10	5	23	0	5	0	m_3

Table 5.4. Numerical Example of the Hypothetical T-Intersection Case: Stage 2 Calculations [$p_2 = m_2$].

T_2	τ_2^*	F_2	$\pi_{a_1}^{m_1}$	$\pi_{a_2}^{m_2}$	$\pi_{a_3}^{m_3}$	p_1
6	3	2	1	1	0	m_3
7	4	3	2	0	0	m_3
8	5	7	3	1	0	m_3
9	6	10	3	1	0	m_3
10	7	12	3	0	0	m_3

Table 5.5. Numerical Example of the Hypothetical T-Intersection Case: Stage 3 Calculations [$p_3 = m_1$].

T_3	τ_3^*	F_3	$\pi_{a_1}^{m_1}$	$\pi_{a_2}^{m_2}$	$\pi_{a_3}^{m_3}$	p_2
9	3	8	0	3	0	m_2
10	3	8	0	2	0	m_2

Table 5.6. Numerical Example of the Hypothetical T-Intersection Case: Stage 3 Calculations [$p_3 = m_2$].

T_3	τ_3^*	F_3	$\pi_{a_1}^{m_1}$	$\pi_{a_2}^{m_2}$	$\pi_{a_3}^{m_3}$	p_2
10	3	16	2	0	0	m_1

Computational Experiences

Several interesting findings have been observed in exercising the numerical example that is just presented. To illustrate a few of them, it is necessary to extend the duration of time horizon *T* from 10 to 20 time units so that more arrival data can be covered and consequently more computations can be shown.

In addition to the hypothetical arrival data listed in Table 5.1, 10 more time units of arrival data have been attached. As a result, Table 5.7 shows the complete arrival data for the hypothetical T-intersection case.

Table 5.7. Complete Arrival Data for the Hypothetical T-Intersection Case.

Time	m_1	m_2	m_3
1	0	0	1
2	0	0	1
3	0	0	0
4	0	1	0
5	0	1	0
6	1	1	0
7	1	0	0
8	1	1	0
9	0	1	0
10	0	0	0
11	0	1	0
12	1	0	0
13	0	1	1
14	1	0	0
15	0	1	0
16	0	1	0
17	1	0	0
18	0	0	0
19	0	0	1
20	0	0	0

Given the rest of the settings fixed as described earlier in this numerical example, the enumeration made by each of the following five methods is compared:

- the brute-force search indicated in Equation (5.13);
- the COP algorithm presented by Sen and Head (1997);
- two OPAC-equivalent methods with a maximum of two and three phase switches, hereinafter referred to as OPAC(2) and OPAC(3), respectively; and
- the DP solution algorithm developed in this chapter, referred to as the Chapter Five DP hereinafter.

To facilitate the comparisons, a computer program has been implemented to perform the calculations based on the COP algorithm and the Chapter Five DP. For the former, enough cycles of phases have been added to the initial phase sequence so that the COP algorithm can guarantee a global optimality.

As reviewed in Chapter Two, OPAC can accommodate a maximum of two or three phase switches in its signal optimization process. Besides, it is indicated that the OSCO method adopted by OPAC is an exhaustive search procedure. Thus, the OSCO method can be approximated by using Equation (5.13). In another word, OPAC(2) and OPAC(3) are equivalent to using Equation (5.13) with the value of k^{max} being constantly set equal to three and four, respectively.

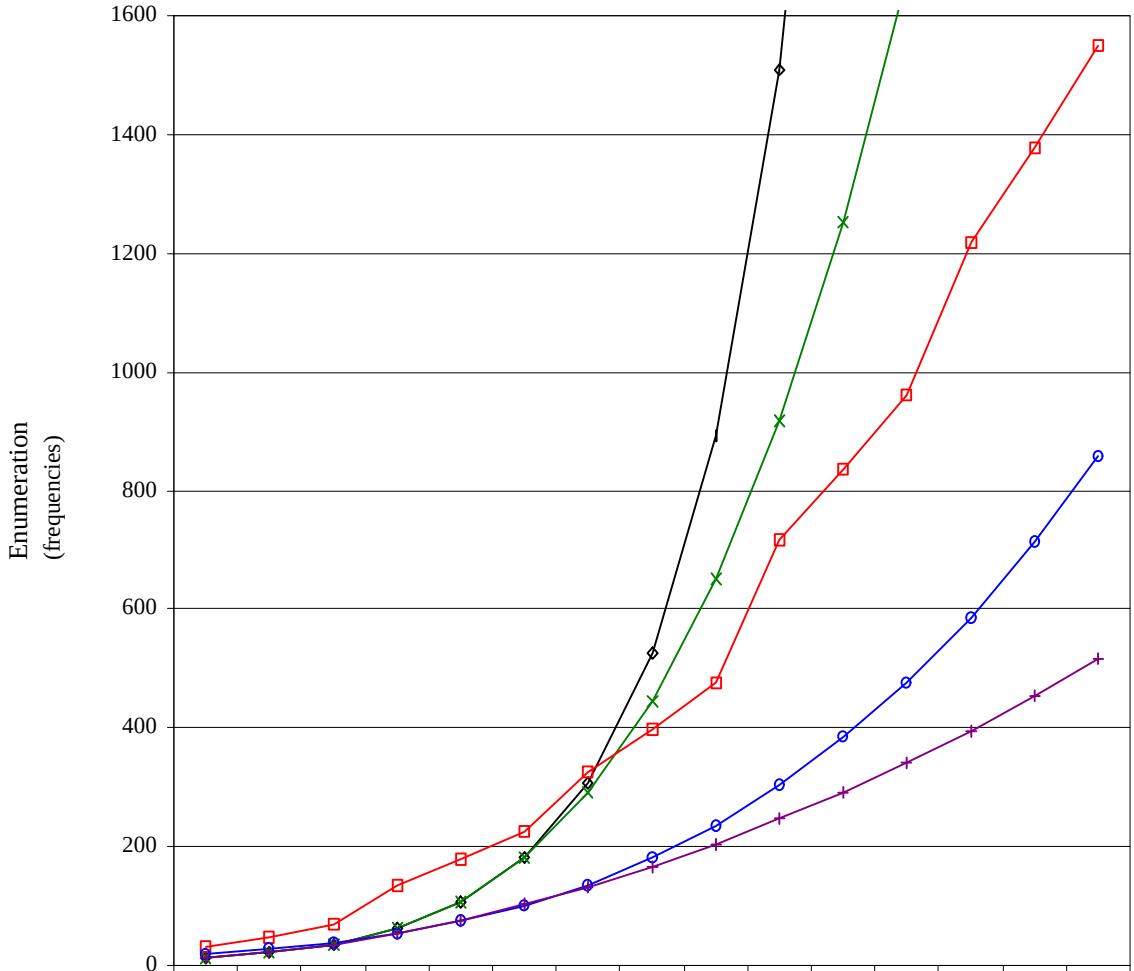
Table 5.8 summarizes the total number of phases being considered by those five search procedures. For each value of time horizon T ranging from 6 to 20 time units, a number next to the COP algorithm or the Chapter Five DP indicates the total number of phases in the optimal phase sequence. However, for the brute-force search, OPAC(2), or OPAC(3), that number indicates the value of k^{max} used in Equation (5.13).

Table 5.8. Maximum Number of Phases in a Phase Sequence Considered by the Five Search Procedures in the Hypothetical T-Intersection Case.

Time Horizon (*T*)	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Brute-Force	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7
OPAC(3)	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
COP Algorithm	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6
Chapter Five DP	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6
OPAC(2)	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3

Figure 5.7 depicts the results from making those comparisons, whose findings together with those observed in Table 5.8 are summarized as follows:

- Unlike the COP algorithm, basically the Chapter Five DP consistently has less enumeration than the brute-force search.
- As the duration of time horizon *T* increases, the Chapter Five DP behaves relatively linearly, while for the brute-force search and OPAC(3) they grow exponentially fairly rapidly. The performance of OPAC(3) agrees with the computational difficulty in OPAC that has been reported in the literature. As to the COP algorithm, its performance is in-between. Roughly speaking, COP generates twice as much enumeration as the Chapter Five DP does.
- Although OPAC(2) generates the least enumeration, neither OPAC(2) nor OPAC(3) guarantees a global optimality. It can be shown in Table 5.8 that there exists cut-off values in time horizon *T*, based on which OPAC(2) and



	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Brute-Force	12	22	36	62	108	182	308	526	892	1510	2564	4350	7372	12502	21204
OPAC(3)	12	22	36	62	108	182	292	446	652	918	1252	1662	2156	2742	3428
COP	30	47	68	135	178	227	327	398	476	717	835	962	1218	1379	1550
Chap5 DP	18	27	38	53	74	101	136	181	236	303	384	477	586	713	858
OPAC(2)	12	22	36	54	76	102	132	166	204	246	292	342	396	454	516

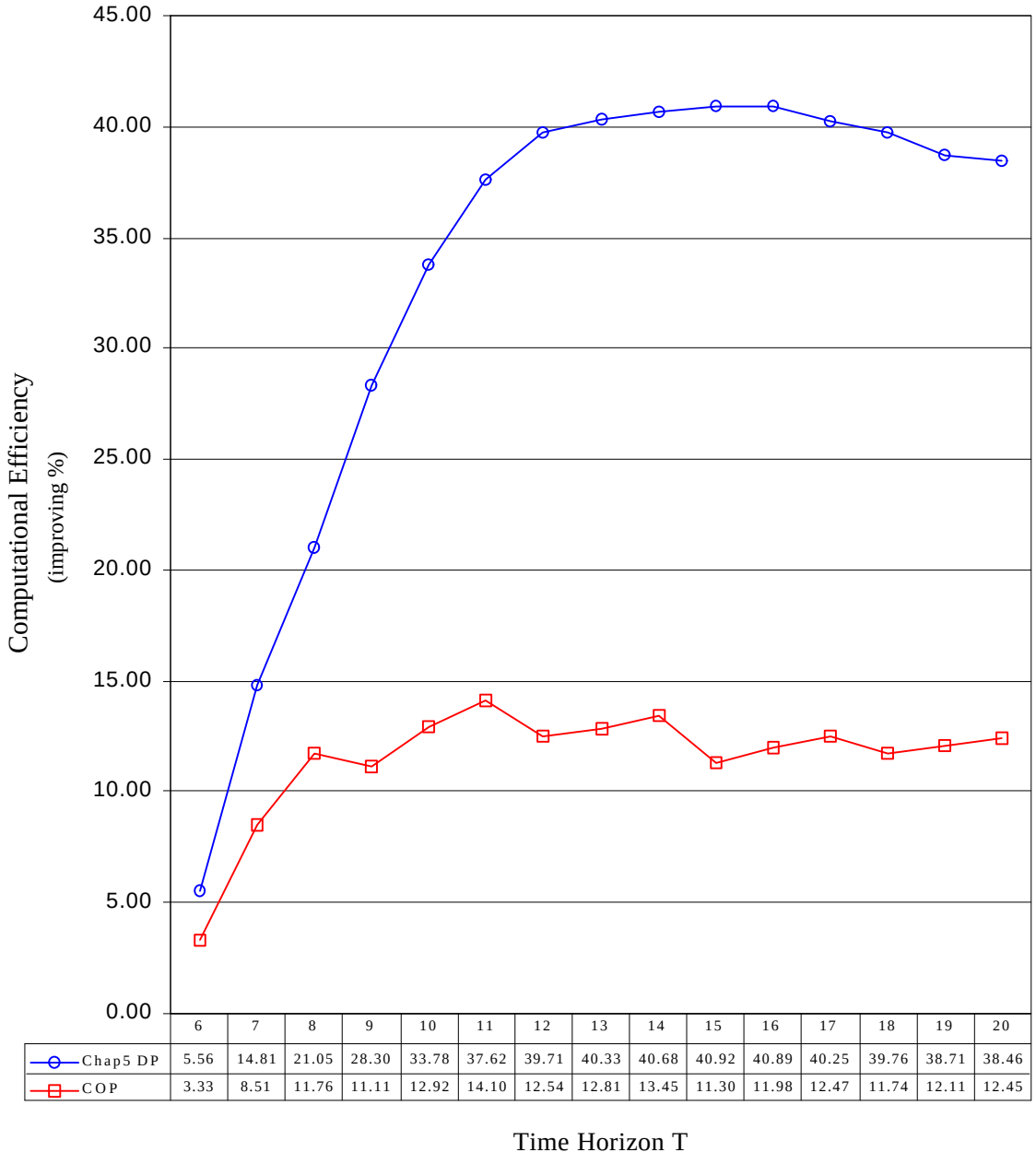
Figure 5.7. Comparisons of Enumeration Made by the Five Search Procedures Considered in the Hypothetical T-Intersection Case.

OPAC(3) can be determined if they can reach a global optimality. The value is 11 for OPAC(2), and 14 for OPAC(3). That is, if $T \leq 11$, OPAC(2) can guarantee a global optimality; otherwise, it cannot because only a portion of the possibilities has been explored. As one can see, that portion is getting smaller as the duration of time horizon T is getting larger, and consequently OPAC(2) is more unlikely to reach a global optimality. The same discussion on OPAC(3) can also be made. The point is that the Chapter Five DP still generates fairly comparable results even given the situations of OPAC(2).

- Compared to the brute-force search, both the COP algorithm and the Chapter Five DP consistently have one phase less in an optimal phase sequence, as indicated in Table 5.8.

Another way to show the computational efficiency of the Chapter Five DP is in terms of the percentages of total enumeration that can actually result in the improvement of the performance index. As a result, a comparison has been performed between the percentages of such an improvement made by the Chapter Five DP and those by the COP algorithm. It is indicated in Figure 5.8 that the former has a comfortable level of about 40% improvement in enumeration, while the latter fluctuates between 10% and 15%. This finding is very significant.

In summary of exercising this numerical example, the DP solution algorithm developed in this chapter consistently generates less enumeration and it has the most efficiency in computation, compared to the brute-force search, the COP algorithm, OPAC(2), and OPAC(3).



	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Chap5 DP	5.56	14.81	21.05	28.30	33.78	37.62	39.71	40.33	40.68	40.92	40.89	40.25	39.76	38.71	38.46
COP	3.33	8.51	11.76	11.11	12.92	14.10	12.54	12.81	13.45	11.30	11.98	12.47	11.74	12.11	12.45

Figure 5.8. Comparison of Computational Efficiency between the COP Algorithm and the Chapter Five DP in the Hypothetical T-Intersection Case.

A Heuristic Search Procedure

Although the DP solution algorithm developed in this chapter can do a much better job than the brute-force enumeration and the COP algorithm as discussed in the previous section, there is a need to address the computational issue when the traffic network being optimized is getting relatively large. In literature, the enlarging state space of the DP calculation is referred to as *curse of dimensionality* by Richard Bellman (1962).

It can be shown that in the DP solution algorithm the cardinal number ρ of $\mathbf{P}$ and the duration of time horizon T are the two major contributing factors to generate such a curse of dimensionality. In virtue of Equation (5.13), of course, one can further reduce the DP calculation by:

- Eliminating unnecessary phases in $\mathbf{P}$ and $\mathbf{M}_i$ for each intersection i, which can consequently result in a smaller number of ρ;
- Decreasing the total duration of time horizon T and maximum green time G^{max}; and
- Increasing minimum green time G^{min}, yellow time Y and all-red time R.

Among those, the most effective way to break the dimensionality is to reduce the duration of time horizon T so that the value of k^{max} in Equation (5.13) can become more manageable. As one can see, the extreme case is to have a value of time horizon T such that k^{max} is equal to two. The point is that when a traffic network being optimized is getting larger it makes sense to consider a smaller duration of time horizon T in the network-wide signal optimization problem. If for some reasons, however, the ways to reduce the DP calculation discussed so far are considered inappropriate, the following

heuristic search procedure can be adopted, which is based on the similar concept just described:

1. Select a reasonable value for k'^{max}, an integer greater than or equal to two, in a sense that the number of DP stages can be manageable.
2. Calculate the time duration T' that is equal to the product of (k'^{max}-1) and (G^{min}+Y+R).
3. Initialize t_{Begin} = 0 and t_{End} = T'.
4. For the time period [t_{Begin}, t_{End}], perform the FDP to calculate the global optimal performance index F^* with the stage j^* and the control variable $p^*_{j^*}$ identified.
5. Perform the BDP to retrieve the optimal policy for the current time period, which will be incorporated into the one cumulated up to the immediately previous time period, if any.
6. If t_{End} is equal to T, stop; otherwise, let t_{Begin} = t_{End} + 1 and go to the next step.
7. If the sum of t_{End} and T' is less than or equal to T, then t_{End} = t_{End} + T'; otherwise, T'' = T - t_{End} and t_{End} = t_{End} + T'' = T. Go to Step 4.

Instead of the entire time horizon T, the heuristic search procedure works on a smaller time horizon T' each time until the total duration of T is covered. Figure 5.9 depicts the idea. By doing so, it can effectively break the curse of dimensionality by transforming to additive from multiplicative computation that generates exponential growth in the DP calculation.

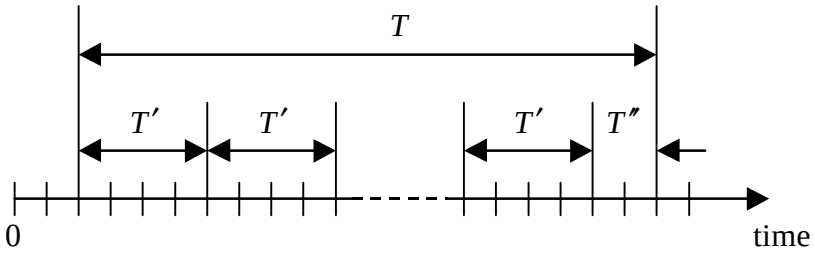


Figure 5.9. Relationship between Time Horizon T and Sub-Time Horizons T' and T''.

CHAPTER SIX
A CASE STUDY

The purpose of this chapter is to assess the efficiency of the methodology developed in this research by virtue of a case study. The case study is conducted in a simulation testing environment, where the two solution algorithms developed in Chapter Five were implemented and evaluated against well-timed fixed-time controls and actuated signals. Although the methodology aims at real-time applications, only off-line evaluations are conducted in this case study. In what follows, the microscopic simulation testbed is introduced first, followed by the presentation of the case study design. The simulation results are presented next, followed by the summary of the findings.

Microscopic Simulation Testbed

The microscopic simulation testbed used in this study includes two customized versions of the CORSIM simulation program: an executable (EXE) and a dynamic link library (DLL). The former simulates the traffic network being studied, while the latter serves as a special function to assess MOEs of concern. Both programs have the capabilities as discussed in Chapter Two. The EXE equips with the logic of the Chapter Five DP and the heuristic search procedure discussed in Chapter Five. Through an interface, the EXE timely launches the DLL in a manner described in Chapter Three.

The interaction between the EXE and the DLL is depicted in Figure 6.1. At time *t* when it is time for the Chapter Five DP or the heuristic search procedure to come up with signal control decisions for the next horizon, the EXE will make a copy of computer

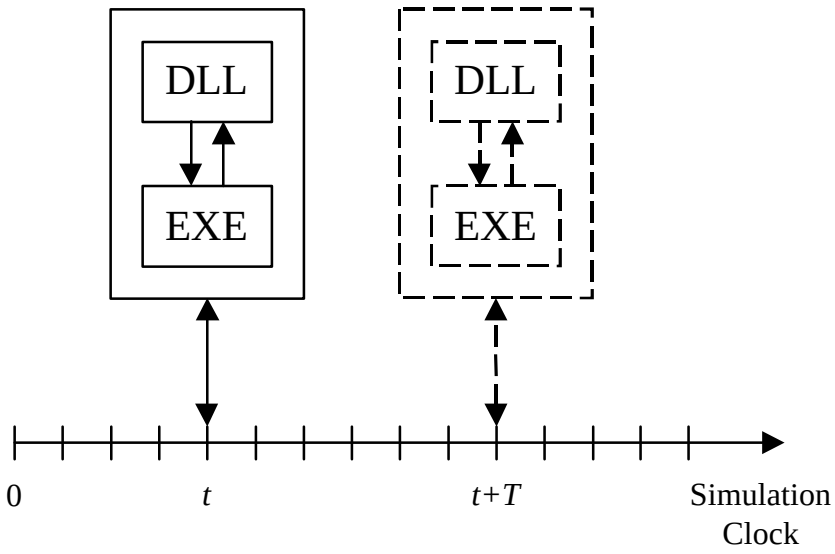


Figure 6.1. The interaction between the EXE and the DLL.

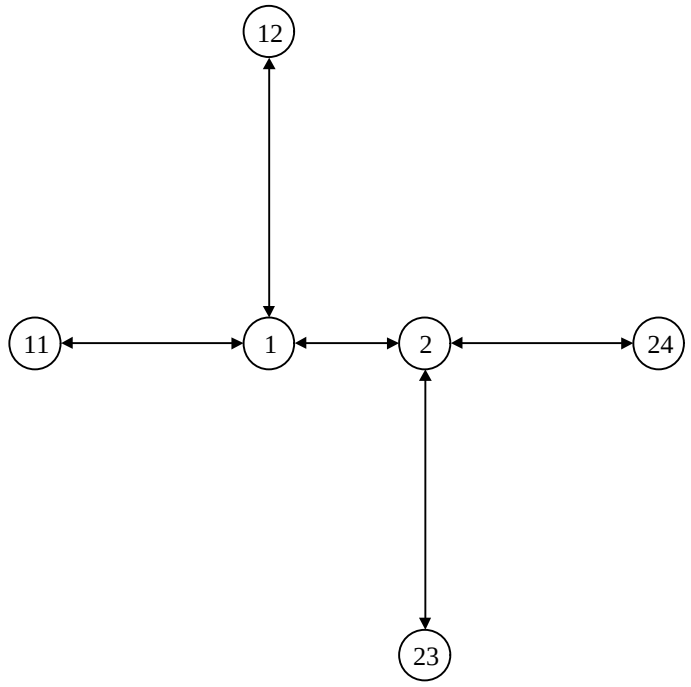
memory containing the current CORSIM status information and launch the DLL many times along the DP optimization process. When a termination condition is reached, the EXE will prepare signal displays for the current horizon based on the decisions made in the optimization process. Then, the simulation clock advances and the EXE performs normal simulation until the next scheduled point in time when the signal control decisions are needed for the next horizon. The procedure repeats itself until the end of the simulation clock time. During the simulation, different statistics of concern can be collected, including the animation display data.

Case Study Design

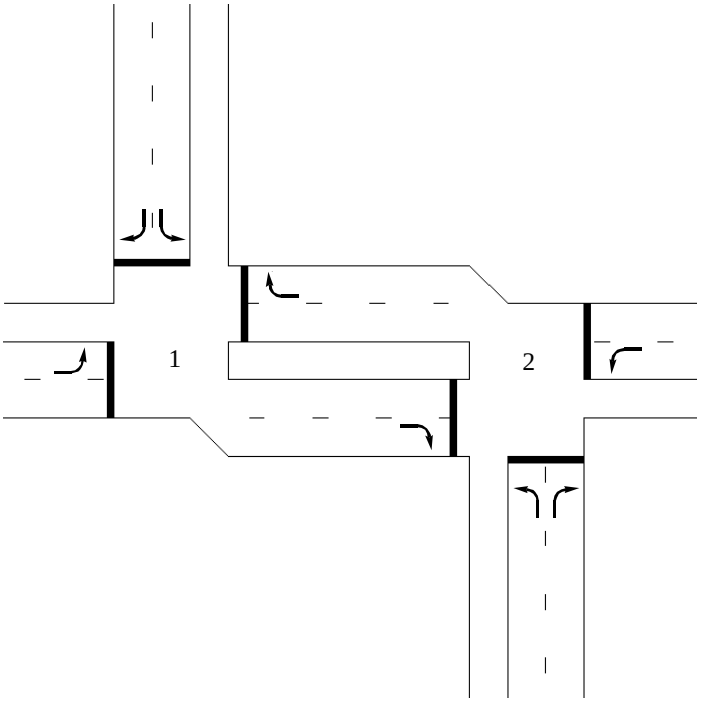
Baseline Simulation Network

The baseline simulation network adopted in this case study involves two tight offset T-intersections. The link-node diagram, intersection layout and detector placement of the baseline simulation network are indicated in Figure 6.2. All possible phase pattern designs available for each intersection in the baseline simulation network are illustrated in Figure 6.3. In addition, the baseline simulation network has the following operational characteristics:

- RTOR allowed,
- No pedestrians,
- Desired free-flow speed of 35 mph,
- Mean queue discharge headway of 2 seconds, and
- Mean start-up lost time of 2.5 seconds.

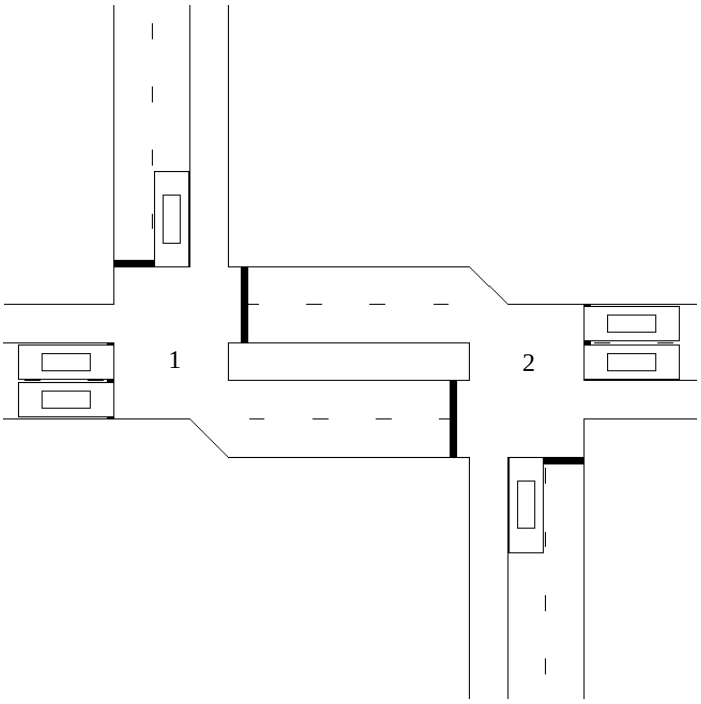


(a)



(b)

Figure 6.2. Two Tight Offset T-Intersections.
(a) Link-Node Diagram; (b) Intersection Layout; (c) Detector Placement.



(c)

Figure 6.2. -- continued.

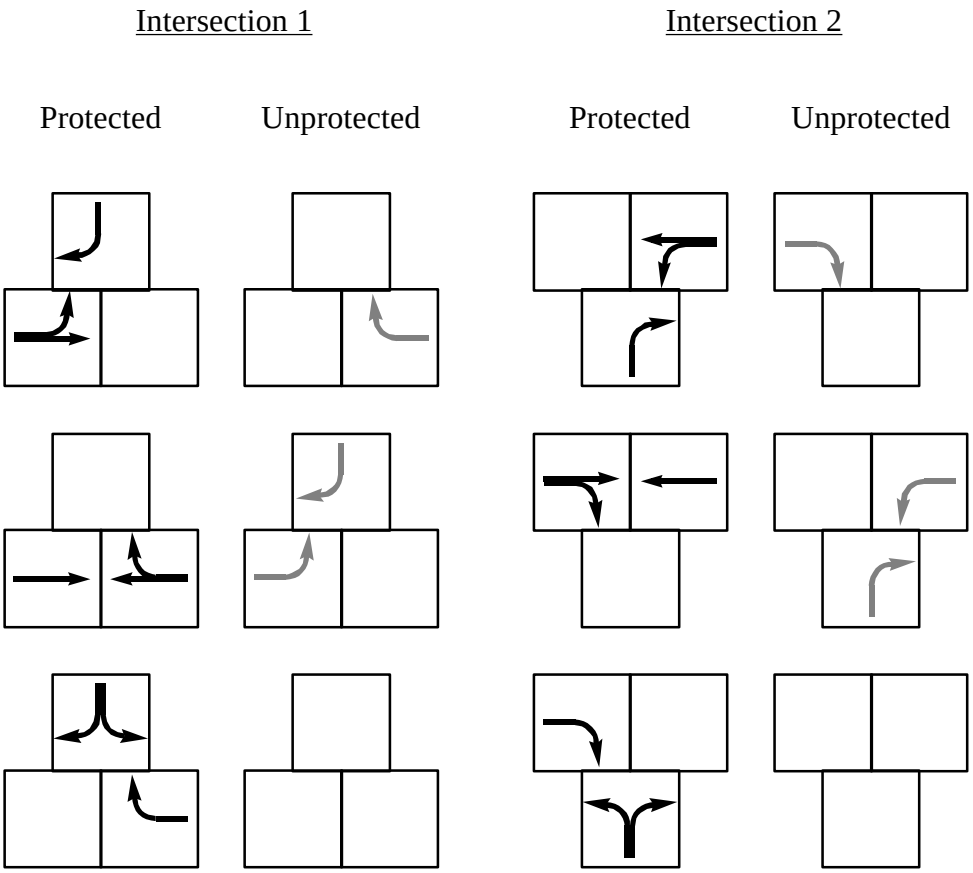


Figure 6.3. All Possible Phase Patterns for the Two Tight Offset T-Intersections.

Alternative Signal Control Strategies

Four signal control alternatives are compared in this case study, and they are: (1) actuated signal, (2) well-timed fixed-time control, (3) the Chapter Five DP, and (4) the heuristic search procedure discussed in Chapter Five.

The actuated signal employed is the one involving utilizing only one actuated controller to operate two intersections. It is a common practice in the field for the situation where the two intersections are so close to each other. For the well-timed fixed-time control, the PASSER IV program is employed to optimize the signal control plans for each testing scenarios. As to the Chapter Five DP and the heuristic search procedure, they are "standalone" in this case study without the presence of a higher-level control to provide necessary information. For this reason, a statically specified objective is used instead. In this case study, an arbitrary objective is selected to serve this purpose and it is to minimize the sum of link-specific queue delay plus one-fourth of queue delay contributed by left turn traffic. Notice that hereinafter the heuristic search procedure will be referred as the Chapter Five DP(h).

The system variables used in the Chapter Five DP and the Chapter Five DP(h) are as follows:

- Discrete time interval of 20 seconds,
- Horizon of 60 seconds,
- Yellow time of 4 seconds,
- All-red time of 1 second,
- Maximum green of 60 seconds, and
- Minimum green of 10 seconds.

Testing Scenarios

Two series of testing scenarios are used in this case study: entry volume variation and left turn traffic variation. For the former, the entry volume for links (11,1), (12,1), (23,2) and (24,2) is increased from 600 vph to 1400 vph with an increment of 100 vph each time. As to the turn percentages for the left turn and the through traffic, they are fixed at 30% and 70%, respectively. For the latter, the left turn percentage for links (11,1) and (24,2) is increased from 10% to 90% with a 10% increment each time. However, all entry volumes are fixed at 1000 vph.

For each testing scenario, 30 replications are generated for the actuated and the fixed-time signal control alternatives. However, only two replications are generated for the Chapter Five DP and the Chapter Five DP(h). All simulation runs are 900 seconds long in time, each with network initialization reaching the state of equilibrium.

Simulation Results and Analysis

The following performance measures are calculated for each signal control alternative for each testing scenario:

- Cumulative number of vehicles discharged,
- Cumulative queue delay,
- Cumulative stop delay, and
- Cumulative travel time.

These measures are all aggregate results by summing over all approach links to the intersections involved; i.e., links (1,2), (2,1), (11,1), (12,1), (23,2) and (24,2). For the Chapter Five DP and the Chapter Five DP(h), the following additional measures are also calculated:

- Total enumeration for each horizon, and
- Frequency of PI improvements for each horizon.

The above performance measures are analyzed to produce the following statistics:

- Average queue delay,
- Average stop delay,
- Average travel time, and
- DP improving ratio.

The results are plotted against different testing scenarios and they are presented in Figures 6.4-6.17, which are organized as follows:

- Figure 6.4 - Number of Vehicles Discharged vs. Entry Volume;
- Figure 6.5 - Average Queue Delay vs. Entry Volume;
- Figure 6.6 - Average Stop Delay vs. Entry Volume;
- Figure 6.7 - Average Travel Time vs. Entry Volume;
- Figure 6.8 - Total Enumeration vs. Entry Volume;
- Figure 6.9 - Frequency of PI Improvements vs. Entry Volume;
- Figure 6.10 - DP Improving Ratio vs. Entry Volume;
- Figure 6.11 - Number of Vehicles Discharged vs. Left Turn Percentage;
- Figure 612 - Average Queue Delay vs. Left Turn Percentage;
- Figure 6.13 - Average Stop Delay vs. Left Turn Percentage;
- Figure 6.14 - Average Travel Time vs. Left Turn Percentage;
- Figure 6.15 - Total Enumeration vs. Left Turn Percentage;
- Figure 6.16 - Frequency of PI Improvements vs. Left Turn Percentage; and
- Figure 6.17 - DP Improving Ratio vs. Left Turn Percentage.

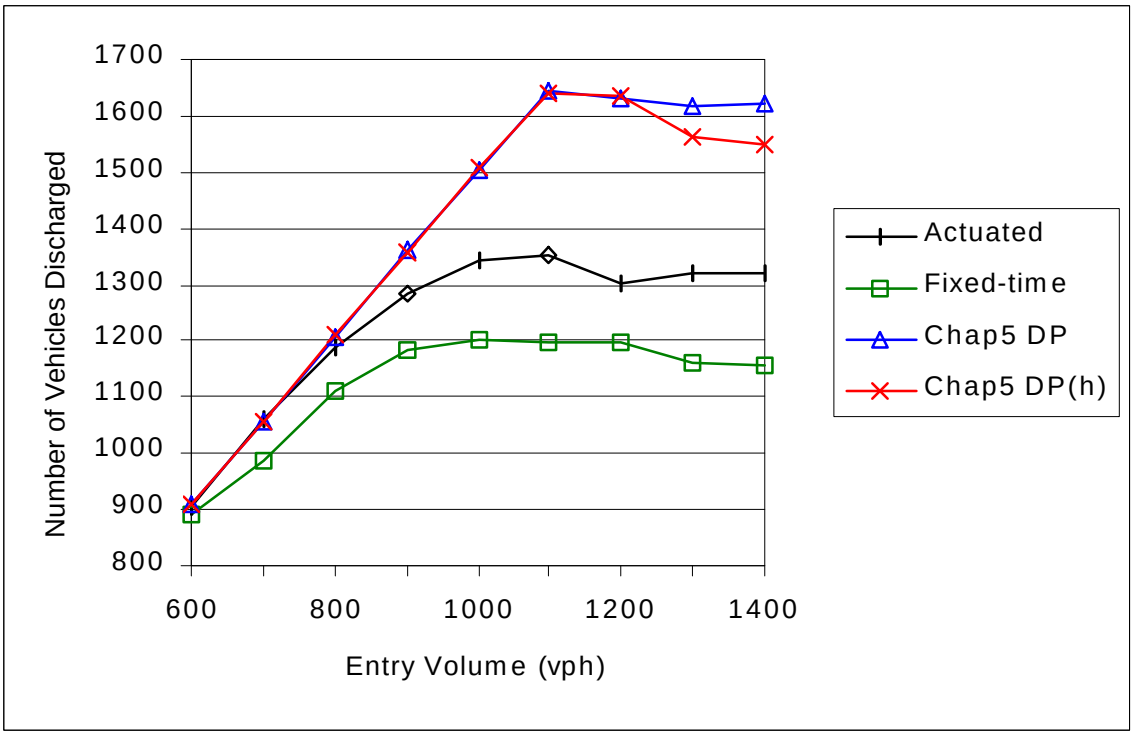


Figure 6.4. Number of Vehicles Discharged vs. Entry Volume.

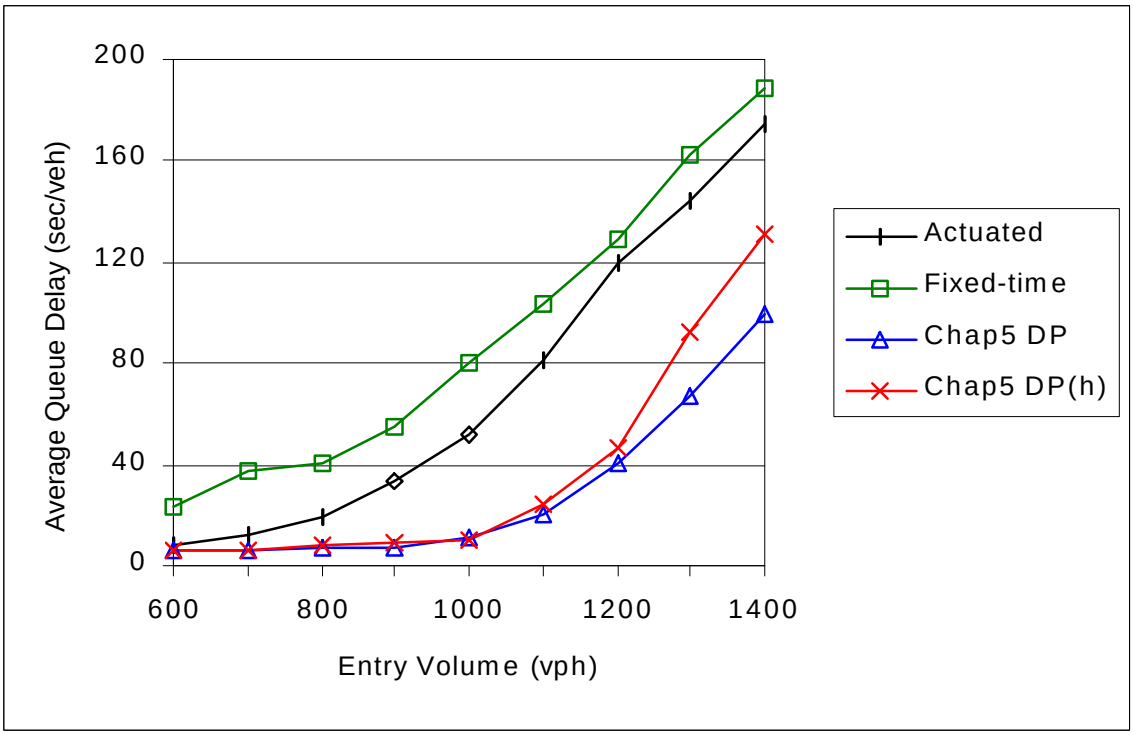


Figure 6.5. Average Queue Delay vs. Entry Volume.

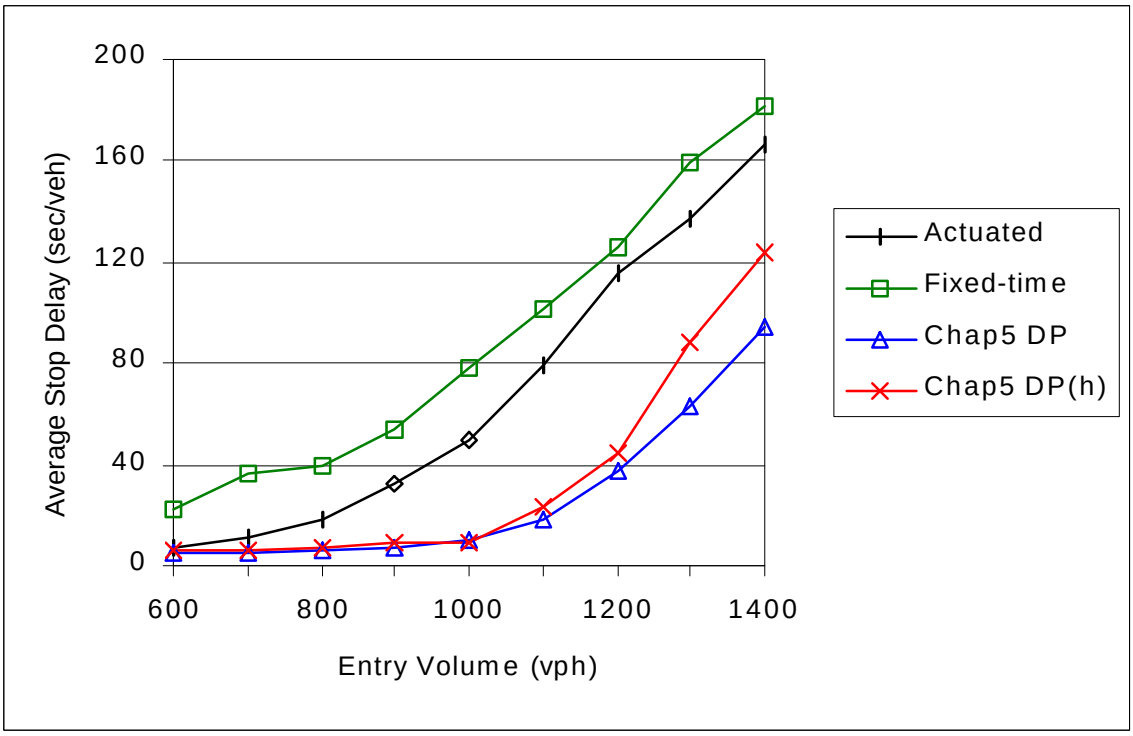


Figure 6.6. Average Stop Delay vs. Entry Volume.

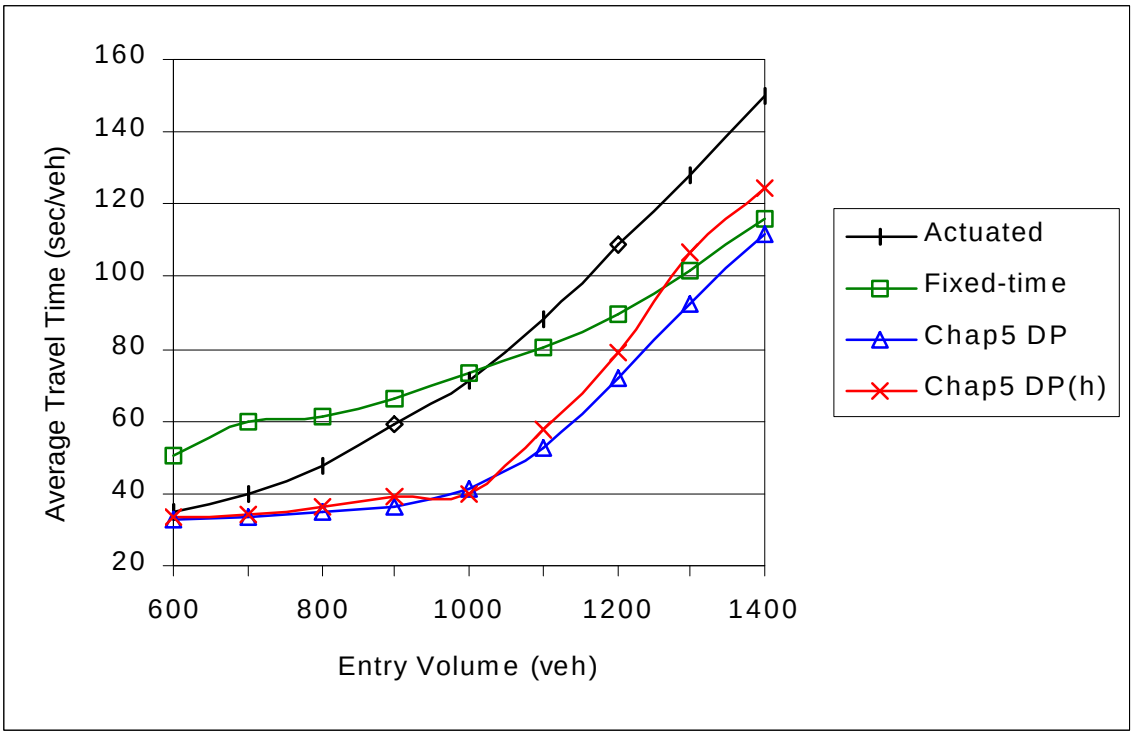


Figure 6.7. Average Travel Time vs. Entry Volume.

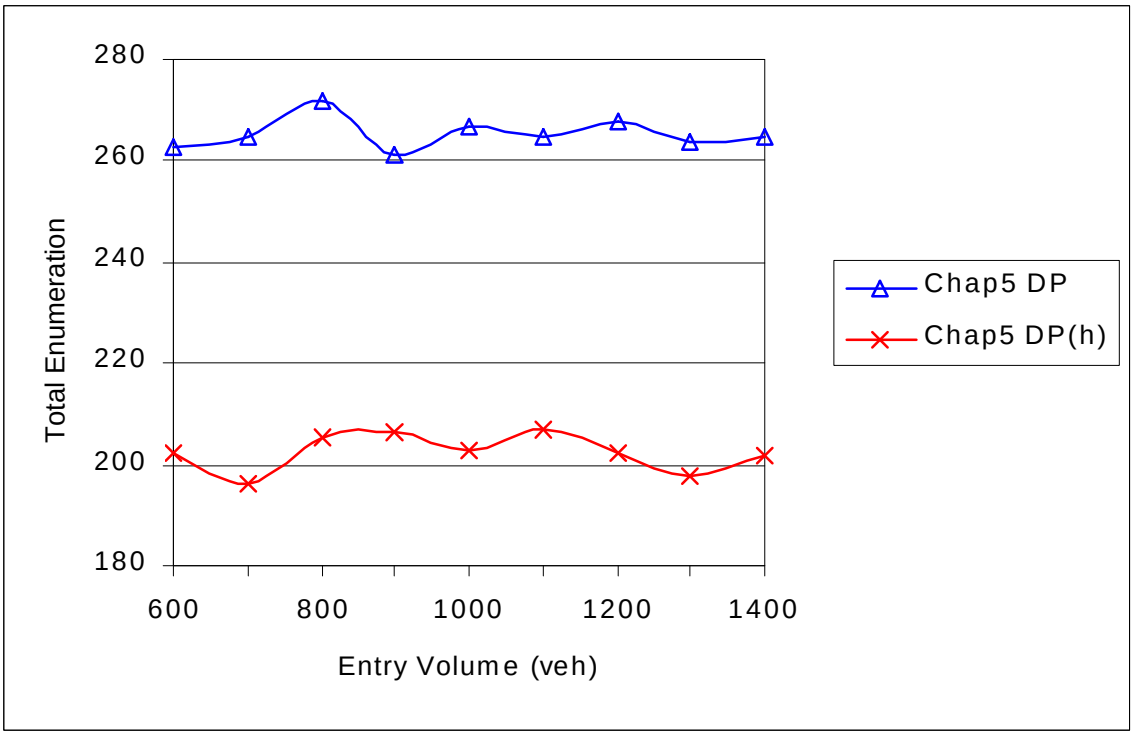


Figure 6.8. Total Enumeration vs. Entry Volume.

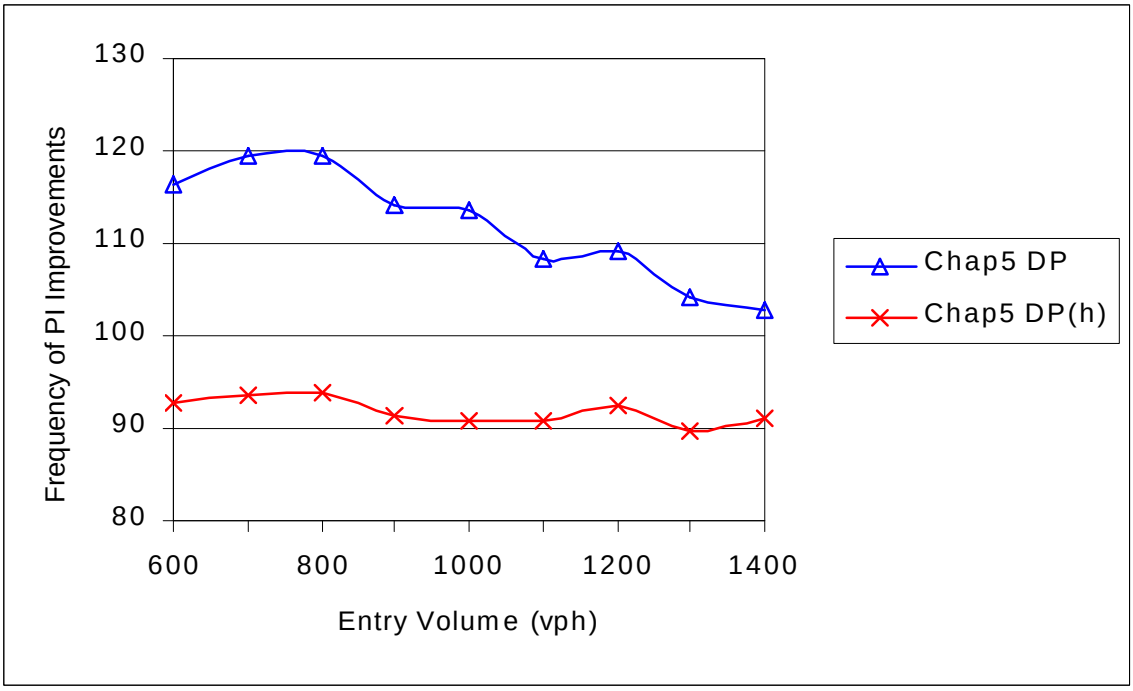


Figure 6.9. Frequency of PI Improvements vs. Entry Volume.

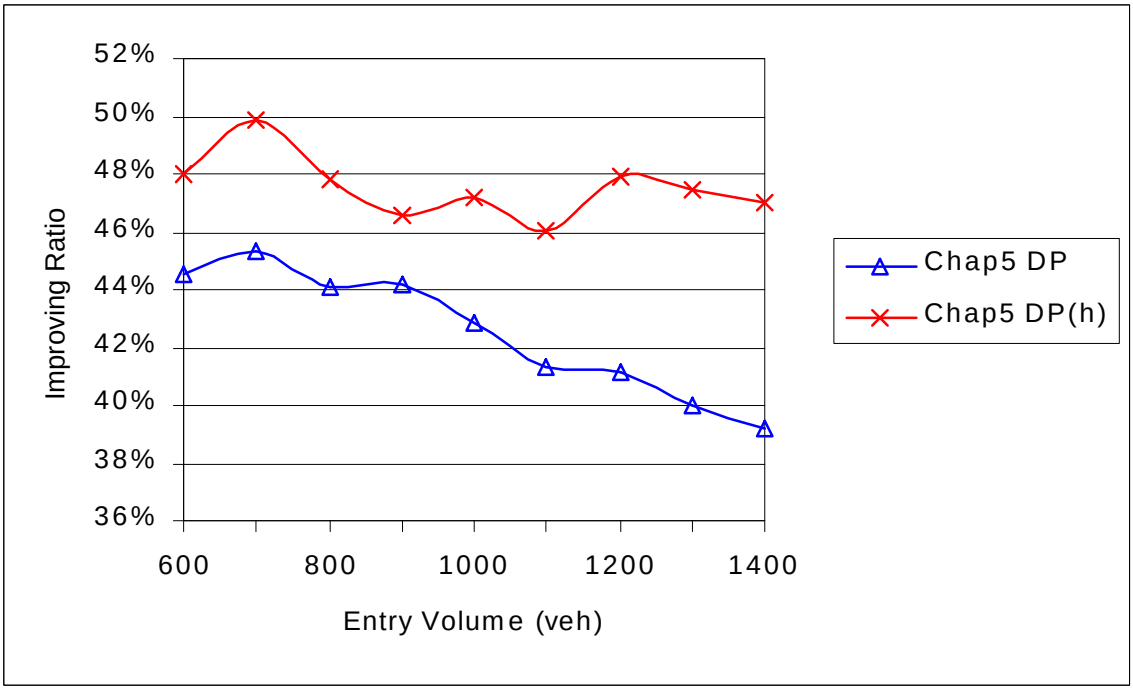


Figure 6.10. DP Improving Ratio vs. Entry Volume.

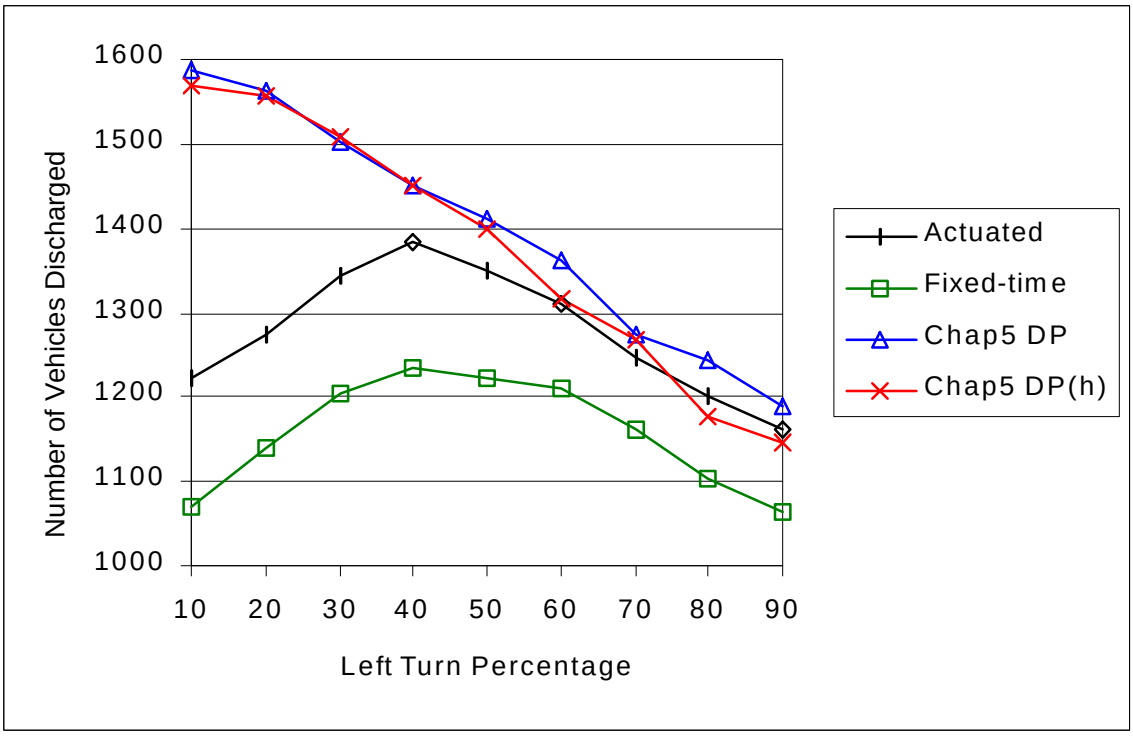


Figure 6.11. Number of Vehicles Discharged vs. Left Turn Percentage.

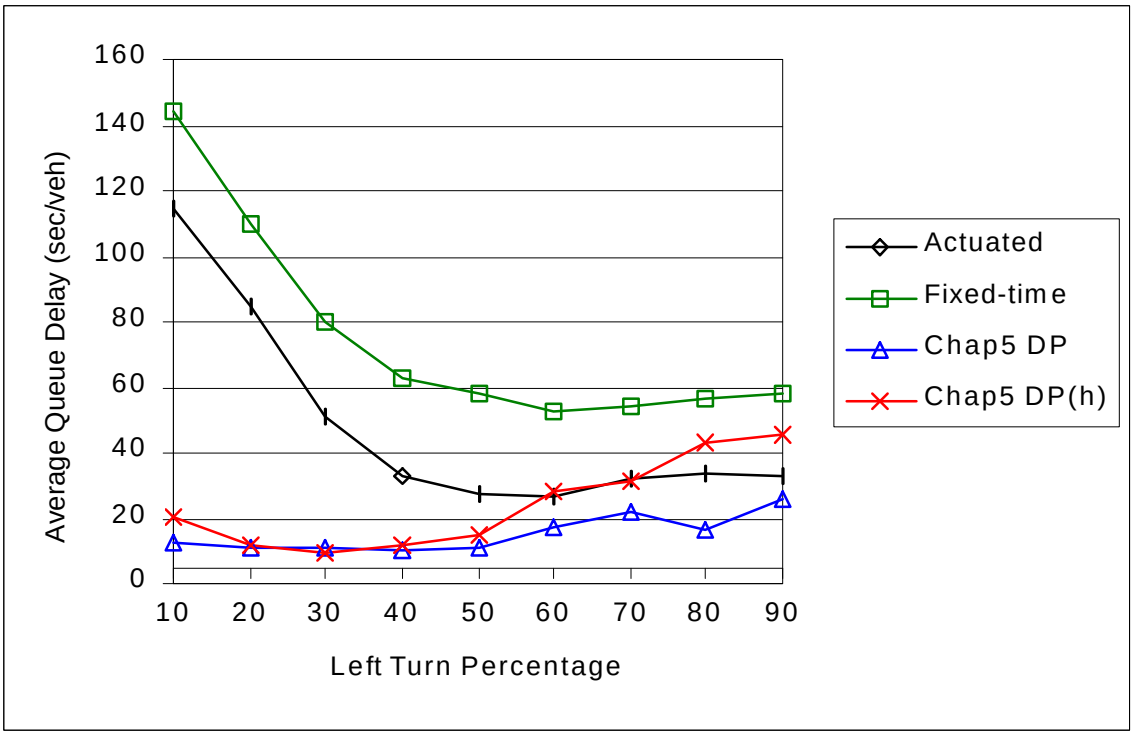


Figure 6.12. Average Queue Delay vs. Left Turn Percentage.

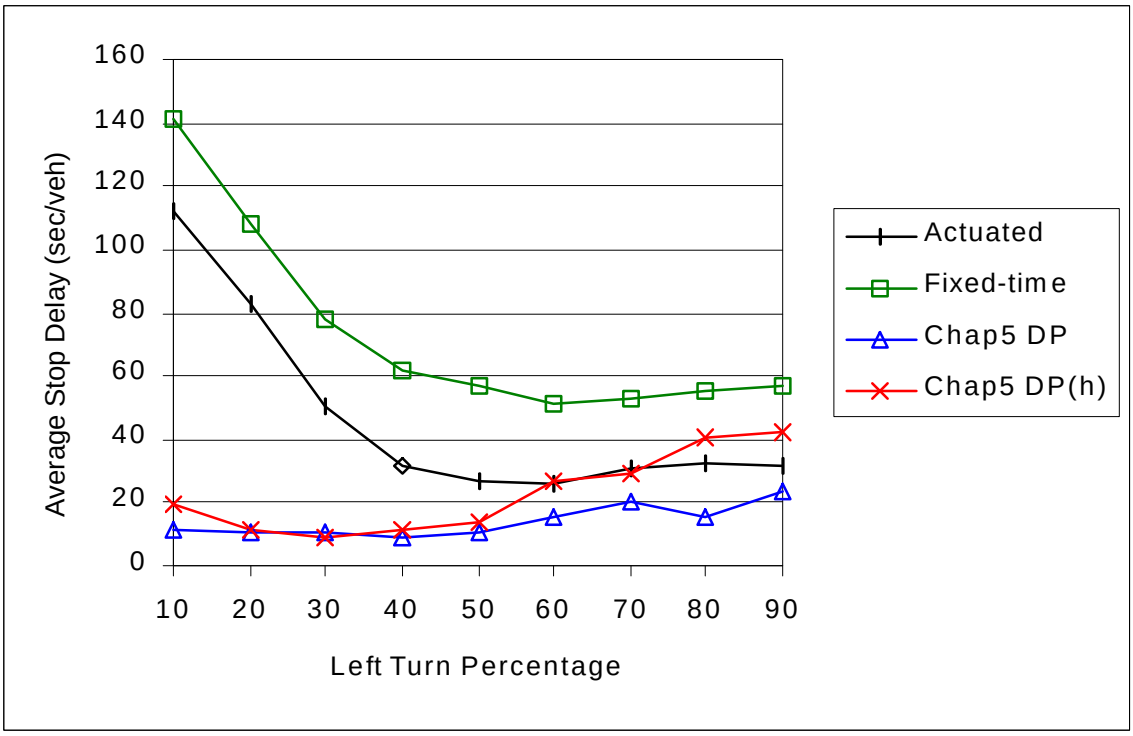


Figure 6.13. Average Stop Delay vs. Left Turn Percentage.

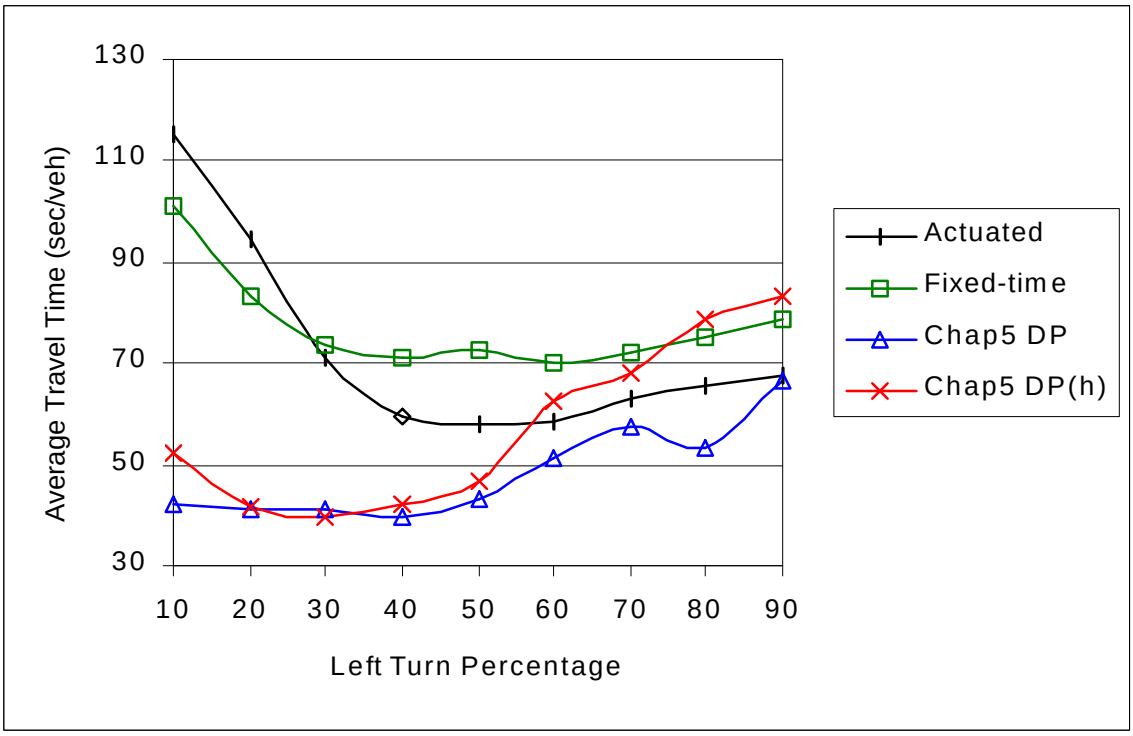


Figure 6.14. Average Travel Time vs. Left Turn Percentage.

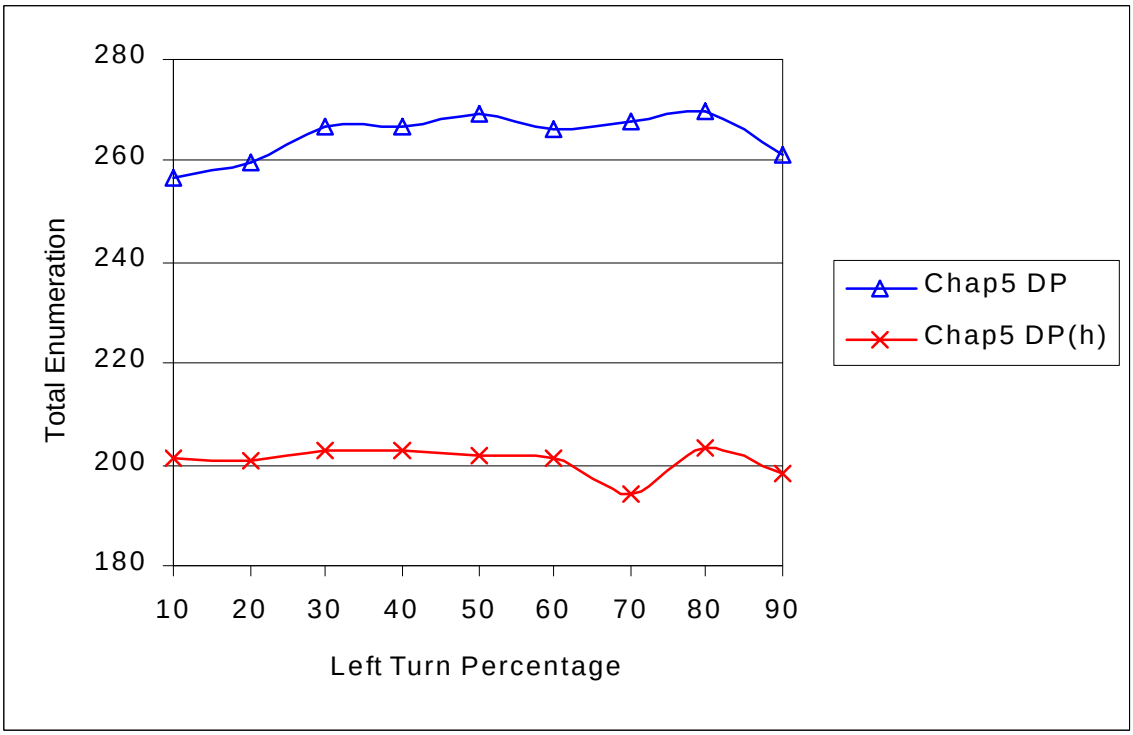


Figure 6.15. Total Enumeration vs. Left Turn Percentage.

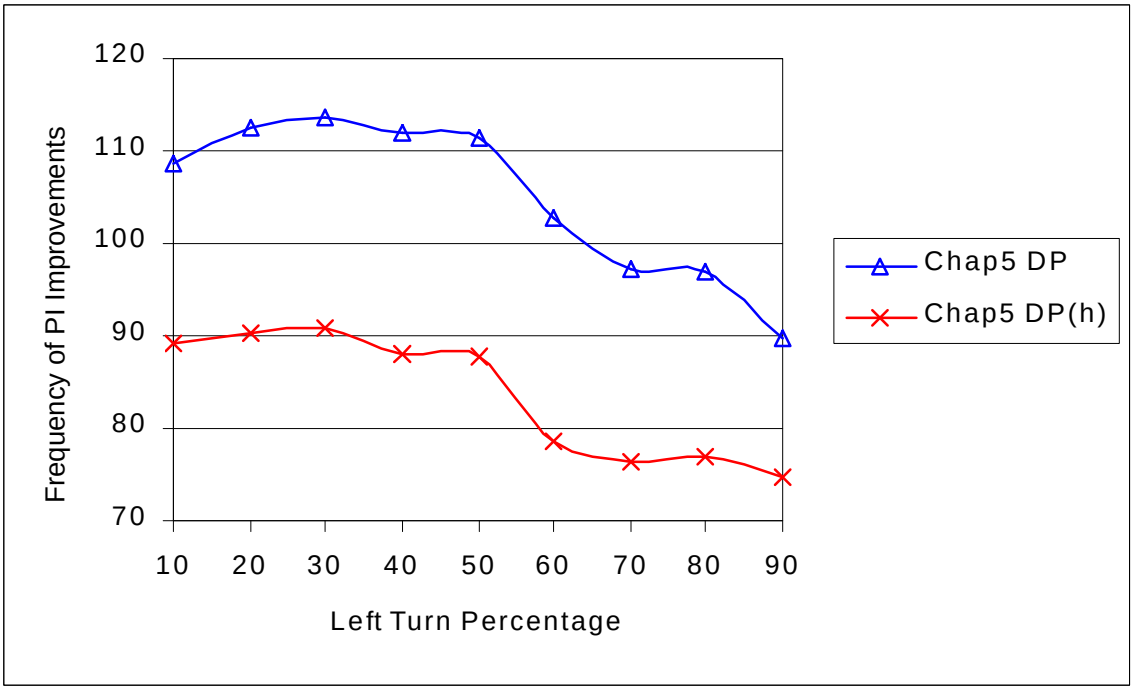


Figure 6.16. Frequency of PI Improvements vs. Left Turn Percentage.

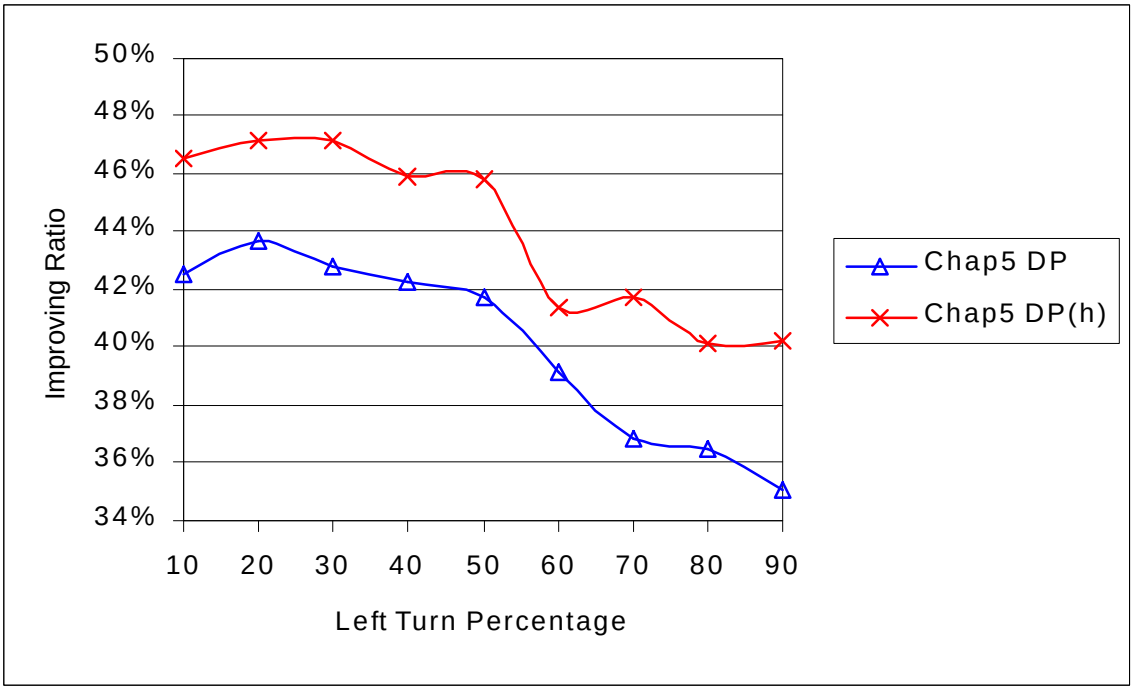


Figure 6.17. DP Improving Ratio vs. Left Turn Percentage.

Summary

The findings from this case study are summarized as follows:

- Number of vehicles discharged. Based on Figures 6.4 and 6.11, both the Chapter Five DP and the Chapter Five DP(h) discharge the largest number of vehicles over all the testing scenarios. The Chapter Five DP(h) can generate the performance results as good as the Chapter Five DP, except for the situations of higher entry volumes and higher left turn percentages. This is due to the high variation in the data used for the Chapter Five DP(h) case since only two replications have been used for the analysis. Moreover, the linear results generated by the Chapter Five DP and the Chapter Five DP(h) in Figures 6.4 and 6.11 suggest the "capacity" of the simulation network. Figure 6.4 indicates the capacity of the simulation network is around 1650 vehicles.
- Average queue delay. Based on Figures 6.5 and 6.12, both the Chapter Five DP and the Chapter Five DP(h) generate the lowest average queue delay. Again, due to the high variation of the data, the performance of the Chapter Five DP(h) was slightly inferior in the situations of higher left turn percentages. Based on Figure 6.12, the performance results generated by both the Chapter Five DP and the Chapter Five DP(h) stay relatively "flat". This matches the expectation of having less variation in the average queue delay since the entry volumes were fixed.
- Average stop delay. Since the stop delay and the queue delay are highly correlated, the findings and the discussion are similar to those of the average queue delay case. The reader may refer to Figures 6.6 and 6.13 for detail.

- Average travel time. Based on Figures 6.7 and 6.14, the Chapter Five DP generates the lowest average travel time. However, due to the high variation of the data used, the performance of the Chapter Five DP(h) was somewhat inferior. Figure 6.14 indicates less variation in the performance results generated by the Chapter Five DP. This actually matches the expectation as explained earlier.
- Total enumeration. Based on Figures 6.8 and 6.15, the total enumeration per horizon generated by the Chapter Five DP(h) is about 75% of the Chapter Five DP over all the testing scenarios.
- Frequency of PI improvements. Based on Figures 6.9 and 6.16, it is indicated that the frequency of improvements per horizon generated by the Chapter Five DP(h) is about 80% of the Chapter Five DP. While the Chapter Five DP(h) stays relatively stable, the Chapter Five DP produces less frequency of PI improvements in situations of higher entry volumes and higher left turn percentages.
- DP improving ratio. Based on Figures 6.10 and 6.17, the overall DP improving ratios of the Chapter Five DP are between 38% and 46% for the entry volume variation testing scenarios; and between 34% and 44% for the left turn percentage testing scenarios. As for the Chapter Five DP(h), those ratios are between 46% and 50%, and between 40% and 48%, respectively. Those statistics of the Chapter Five DP are compatible with the numbers indicated in the numerical example discussed in Chapter Five.

In summary, the Chapter Five DP performs the best over all the testing scenarios conducted in this case study. As to the Chapter Five DP(h), its performances are comparable except for the situations of higher entry volumes and higher left turn percentages.

CHAPTER SEVEN
CONCLUSIONS AND RECOMMENDATIONS

This dissertation aims to develop a methodology suitable for a dynamic urban traffic control environment so that signal phasing, timing and coordination are simultaneously considered, formulated and then optimized based on real-time traffic information. The major activities in this research are summarized as follows:

1. Review the literature in the areas of off-line signal optimization models, on-line signal control strategies, and microscopic traffic simulation models; and summarize the review findings as a result.
2. Identify the functional requirements of a distributed signal control system and develop a modeling framework suitable for the system.
3. Develop a general IP model formulation, which serves as a reasonably faithful representation of the signal optimization problem that is generally recognized by the traffic research community. In addition, it addresses many more real-world issues that have not yet been attempted in other methodologies before.
4. Develop a DP solution framework to solve the general IP model developed.
5. To form a contrast, develop an assessment method to estimate enumeration done by an exhaustive search procedure.
6. Illustrate the DP solution algorithm by virtue of an numerical example, including the discussion on computational experiences learned.

7. Develop a heuristic search procedure that significantly simplifies the DP calculation.
8. Conduct a case study to assess the efficiency of the methodology developed.

Conclusions

The methodology developed in this study is shown effective to formulate and solve the generally recognized network-wide signal optimization problem. The conclusions listed below are offered as a result of this study:

1. The findings from the literature review on off-line signal optimization models and on-line control strategies indicate that few have realistic traffic flow models incorporated in their optimization process. Consequently, performance indices of concern cannot be accurately assessed. The methodology developed in this study resolves this problem.
2. Due to the stochastic nature of traffic signal control, it is shown a feasible concept to incorporate a microscopic simulation model in the network-wide signal optimization process so that that randomness can be easily and realistically captured and modeled.
3. It is shown a feasible concept by using joint phase to associate complicated modeling issues related to geometrics, movements, phases, traffic dynamics, and signal control decisions in network-wide signal optimization.
4. It is shown that the Chapter Five DP solution algorithm considers signal phasing and timing simultaneously. It is capable of performing network-wide signal optimization without compromising global optimality. As for signal

coordination, it can be automatically achieved based on network-wide objectives.

5. It is shown that the Chapter Five DP solution algorithm developed is more computationally efficient than the brute-force search, the OPAC method, and the COP algorithm.
6. It is shown that the heuristic search procedure developed in Chapter Five can significantly reduce the DP calculations and can still generate comparable performance results.

Recommendations

Several recommendations have emerged from this study, and they fall into five areas: calibration of parameters used in the DP solution algorithm, further improvements to the heuristic search procedure, conducting comprehensive laboratory testing, considering driver's expectations, and further research warranted to address issues involving the higher-level control.

Several parameters are used in the DP solution algorithm, and they are very sensitive to the computational capability of the DP solution algorithm. They are the duration of the time interval, the total length of the time horizon, the minimum and maximum green time settings, and the duration of the change interval. There is a need to perform further study so that a general guideline can be established as to how to use those parameters more effectively.

When the size of a network grows, the complexity of the DP calculation grows tremendously. So far, the heuristic search procedure developed provides an intuitive way to break the curse of dimensionality. However, there is a need to perform further study to

quantify the differences between the results generated by the Chapter Five DP solution algorithm and the heuristic search procedure so that enhancements can be made to a more robust heuristic search procedure.

Although the case study shows promising results, it only involves a simple network of two intersections. There is a need to conduct a comprehensive laboratory testing so that more computational experiences can be gained.

The methodology developed in this study performs network-wide signal optimization purely from the perspective of mathematics. As a result, the signal control decisions generated by the methodology may not satisfy driver's expectations. To make the methodology developed in this study more applicable and practical, there is a need to address the issues of the driver's expectations in the methodology.

The success of the DP solution algorithm is heavily dependent upon that of the higher-level control. In this study, it is assumed that all the needed pieces of information are available. Relaxing any of those assumptions offer quite a few research topics that are worthwhile to explore:

1. The methods of how to come up with real-time traffic measures provided by the surveillance system so that the DP solution algorithm can be more productive;
2. The ways to share real-time information with other ATMS and ATIS systems at what rate and in what format;
3. The models to dynamically decompose the entire network in real-time based on prevailing and predicted traffic information;

4. The models to successfully provide real-time traffic predictions to the DP solution algorithm; and
5. The models to integrate all the subnetworks to assure smooth traffic progression across the subnetwork boundaries.

APPENDIX A
SUMMARY OF NOTATIONS

Systems Variables:

Δt = discrete time interval;

T = total number of discrete time intervals, or total length of time horizon;

Y = yellow time;

R = all-red time;

Sets:

N = set of intersections;

A = set of links;

M_i = set of phases for intersection i, each composed of protected and allowable unprotected movements in the intersection;

P = set of joint phases for all intersections in traffic network Γ, each composed of a phase m in M_i for each intersection i;

S_k = set of k-phase sequences, where $k \geq 2$;

X = set of signal control decisions;

V = set of vehicle types;

Γ = traffic network;

Ω = solution space;

Ψ = entire collection of real-time traffic information;

Index Variables:

t = discrete time index;

$i \in \mathsf{N}$ = intersection index;

$a \in \mathsf{A}$ = link index;

$m \in \mathsf{M}_i$ = phase index for each intersection i;

$p \in \mathsf{P}$ = joint phase index;

$s \in \mathsf{S}_k$ = phase sequence index;

$v \in \mathsf{V}$ = vehicle type index;

Signal Control Variables:

p_1 = initial phase, $p_1 \in \mathsf{P}$;

p_j = j^{th} phase, $p_j \in \mathsf{P}$;

τ_0 = time already allocated to initial phase p_1 in the immediately previous horizon, assuming the minimum green requirement has been satisfied;

τ_j^s = time allocated to the j^{th} phase in phase sequence s, which includes the duration of a change interval (yellow time plus all-red time), if applicable;

τ_j^{min} = minimum value that τ_j can assume;

τ_j^{max} = maximum value that τ_j can assume;

T_j = total number of discrete time intervals that have already been allocated after the completion of the j^{th} phase;

x_j^{ps} = 1, if joint phase p is the j^{th} phase in phase sequence s; and 0, otherwise;

ϕ_i^{mp} = 1, if phase *m* of intersection *i* is in joint phase *p*; and 0, otherwise;

θ_{ia}^{m} = 1, if link *a* is in phase *m* of intersection *i*; and 0, otherwise;

$G_i^{m,max}$ = maximum green time allowed for phase *m* of intersection *i*;

$G_i^{m,min}$ = minimum green time required for phase *m* of intersection *i*;

G_{ij}^{ms} = green time for phase *m* of intersection *i* elapsed since its onset until the end of the j^{th} phase in phase sequence *s*, if phase *m* is in the j^{th} phase of the phase sequence (i.e., $\phi_i^{mp} \cdot x_j^{ps} = 1$); and 0, otherwise;

$\gamma_i^{m'm}$ = signal clearance interval needed from phase m' to phase *m* of intersection *i*: 0, if $m' = m$; and a value equal to either *Y* or *Y* + *R* depending on phases m' and *m*, otherwise;

Link-Specific Traffic Variables:

$\alpha_{ia}^{mv}(t)$ = number of vehicle type *v* arrivals on link *a* associated with phase *m* of intersection *i* at time *t*;

λ_{ia}^{max} = maximum number of vehicles that can be accommodated on link *a* of intersection *i*;

$\pi_{ia}(t)$ = number of vehicles on link *a* of intersection *i* at time *t*;

$\pi_{ia}(0)$ = number of vehicles on link *a* of intersection *i* at the very beginning of the current horizon;

$\xi_{ia}(t)$ = incident information associated with link *a* of intersection *i* at time *t*;

$\upsilon_{ia}(t)$ = transit (bus) operations information associated with link *a* of intersection *i* at time *t*;

$\zeta_{ia}(t)$ = parking activity information associated with link *a* of intersection *i* at time *t*;

$\eta_{ia}(t)$ = pedestrian traffic information associated with link *a* of intersection *i* at time *t*;

$\delta_{ia}(t)$ = other event information associated with link *a* of intersection *i* at time *t*, resulting in any link capacity reduction;

Functions:

$F(\cdot)$ = objective function;

$F_j(p_j, T_j)$

= optimal value function denoting the optimal performance index after the completion of the j^{th} phase p_j and the total number of discrete time intervals T_j that have been allocated;

$f_j(p_{j-1}, p_j, T_j, \tau_j)$

= incremental performance index from joint phase p_{j-1} to p_j associated with state variable T_j and control variable τ_j;

$\beta_{ia}^m(t)$ = number of vehicle departures from link *a* associated with phase *m* of intersection *i* at time *t*;

$\psi_{ia}^m(t)$ = number of vehicle departures from link *a* made by permitted movements or RTOR vehicles in phase *m* of intersection *i* at time *t*;

$f_\beta(t)$ = queue discharge model based on prevailing traffic conditions associated with function $\beta_{ia}^m(t)$ at time *t*.

APPENDIX B
DERIVATION OF THE BRUTE-FORCE ENUMERATION

The derivation of the brute-force enumeration can be proceeded by considering each time the product of the combinations of all possible phasing and timing for each set of feasible *k*-phase sequences, where $k \geq 2$. The mathematical induction technique is used to carry out such a derivation.

In what follows, the symbol $\binom{b}{d}$ should be read, "the combination of *b* items taken *d* at a time" , which is equal to $\frac{b!}{d!(b-d)!}$. In addition, the following are given: the initial phase p_1, the cardinal number ρ of P, the total length of time horizon *T*, the yellow time *Y*, the all-red time *R*, and the minimum and maximum green times G^{min} and G^{max} for every phase. Besides, it is assumed that the initial phase p_1 has already satisfied the minimum green requirement. For the sake of simplicity, let $G^{max} = T$ so that it will be unnecessary to consider G^{max} here.

When $k = 2$, the combination of all possible 2-phase sequences is (ρ-1), or $(\rho\text{-}1)^{k-1}$, since p_1 is given. For each 2-phase sequence, the total amount of time left is $[T\text{-}(G^{min}+Y+R)]$, or $[T\text{-}(k\text{-}1)\cdot(G^{min}+Y+R)]$, after both the duration of the change interval $(Y+R)$ has been implemented for the initial phase p_1 and the minimum green time G^{min} required for the second phase p_2 has been satisfied. Notice that there is no need to consider again the minimum green requirement for the initial phase since it has already been satisfied based on the assumption. For the sake of discussion, hereinafter let

$C = G^{min} + Y + R$. It follows that the combination of all possible ways to distribute the amount of time left, i.e., (T-C), over the initial phase and the second phase is [T-(C-1)], or $\binom{T-(k-1)\cdot(C-1)}{k-1}$. That is, when the timing for one phase is determined, the timing for the other is decided automatically. Notice that the reason to subtract one from C is because of the possibility to assign zero (i.e., no extra) time unit to either phase. Hence, the brute-force enumeration for the case where $k = 2$ is:

$$E_k = (\rho - 1)^{k-1} \cdot \binom{T-(k-1)\cdot(C-1)}{k-1}; \quad \text{for } k = 2 \tag{B.1}$$

When $k = 3$, the combination of all possible 3-phase sequences is $(\rho\text{-}1)^2$, or $(\rho\text{-}1)^{k-1}$. For each 3-phase sequence, the total amount of time left is (T-2·C), or [T-(k-1)·C], after the duration of the change intervals has been implemented for the initial phase p_1 and the second phase p_2 and the minimum green times required for the second phase p_2 and the third phase p_3 have been satisfied. Similar to the case where $k = 2$, it follows that the combination of all possible ways to distribute the amount of time left, i.e., (T-2·C), over the three phases is $\binom{T-2\cdot(C-1)}{2}$, or $\binom{T-(k-1)\cdot(C-1)}{k-1}$. That is, when the timing for any two out of the three phases is determined, the timing for the remaining one is decided automatically. Notice that the reason to subtract one from C is, again, because of the possibility to assign zero time unit to any two of the three phases. Hence, the brute-force enumeration for the case where $k = 3$ is:

$$E_k = (\rho - 1)^{k-1} \cdot \binom{T-(k-1)\cdot(C-1)}{k-1}; \quad \text{for } k = 3 \tag{B.2}$$

Suppose that $k = n$ and that the following relation holds [assuming T is big enough so that $\binom{T-(k-1)\cdot(C-1)}{k-1}$ remains valid]:

$$E_k = (\rho - 1)^{k-1} \cdot \begin{pmatrix} T-(k-1)\cdot(C-1) \\ k-1 \end{pmatrix}; \quad \text{for } k = n \tag{B.3}$$

If it can be shown that the above relationship still holds for $k = n+1$, then the relationship is proven.

If $k = n+1$, the combination of all possible k-phase sequences is $(\rho\text{-}1)^n$, or $(\rho\text{-}1)^{k-1}$. For each k-phase sequence, the total amount of time left is $(T\text{-}n\cdot C)$, or $[T\text{-}(k\text{-}1)\cdot C]$, after:

1. The duration of the change intervals has been implemented for the phases starting from the initial phase p_1, the second phase p_2, etc., all the way up to the n^{th} phase; and
2. The minimum green times have been satisfied for the phases starting from the second phase p_2, the third phase p_3, etc., all the way up to the $(n+1)^{th}$ phase.

Similar to the previous discussion, it follows that the combination of all possible ways to distribute the amount of time left, i.e., $(T\text{-}n\cdot C)$, over the $(n+1)$ phases is $\binom{T\text{-}n\cdot(C\text{-}1)}{n}$, or $\binom{T\text{-}(k\text{-}1)\cdot(C\text{-}1)}{k\text{-}1}$. That is, when the timing for any n out of the $(n+1)$ phases is determined, the timing for the remaining one is decided automatically. Notice that the reason to subtract one from C is, again, because of the possibility to assign zero time unit to any n out of the $(n+1)$ phases. Hence, the brute-force enumeration for the case where $k = n+1$ is:

$$E_k = (\rho - 1)^{k-1} \cdot \begin{pmatrix} T-(k-1)\cdot(C-1) \\ k-1 \end{pmatrix}; \quad \text{for } k = n+1 \tag{B.4}$$

Therefore, it is concluded that the brute-force enumeration for any $k \geq 2$ is:

$$E_k = (\rho - 1)^{k-1} \cdot \begin{pmatrix} T-(k-1)\cdot(C-1) \\ k-1 \end{pmatrix}; \quad \forall k \geq 2 \tag{B.5}$$

In Equation (B.5), it indicates that the following relationship has to be satisfied for each value that *k* can assume:

$$k-1 \quad \leq \quad T-(k-1)\cdot(C-1) \tag{B.6}$$

That is,

$$k \quad \leq \quad \frac{T}{C}+1 = \frac{T}{G^{min}+Y+R}+1 \tag{B.7}$$

Hence, k^{max}, the maximum value that *k* can assume, is the integer portion of the right-hand side of Equation (B.7).

In summary, the total brute-force enumeration, E_{BF}, can be written as

$$E_{BF} = \sum_{k=2}^{k^{max}} (\rho-1)^{k-1} \cdot \begin{pmatrix} T-(k-1)\cdot(G^{min}+Y+R-1) \\ k-1 \end{pmatrix} \tag{B.8}$$

where

k^{max} = the integer portion of $\left(\frac{T}{G^{min}+Y+R}+1\right)$; and

ρ = $\prod_{i\in N}\mu_i$.

APPENDIX C
DETAILED CALCULATION FOR THE HYPOTHETICAL T-INTERSECTION CASE

Stage $j = 1$ [$p_1 = m_3$]

T_1	τ_1 †	F_1	$\pi_{a_1}^{m_1}$	$\pi_{a_2}^{m_2}$	$\pi_{a_3}^{m_3}$
1	1^*	1	0	0	1
2	2^*	1	0	0	1
3	3^*	0	0	0	0
4	4^*	1	0	1	0
5	5^*	3	0	2	0
6	6^*	7	1	3	0
7	7^*	12	2	3	0
8	8^*	19	3	4	0
9	9^*	27	3	5	0
10	10^*	35	3	5	0

† Hereinafter, numbers with "*" indicate optimal solutions for each corresponding sub-problem. If no such a number exists for a sub-problem, it indicates that no optimal solution exists for that problem.

Stage $j = 2$ [$p_2 = m_1$, $p_1 = m_3$]

T_2	τ_2	f_2	T_1	F_1	F_2	$\pi_{a_1}^{m_1}$	$\pi_{a_2}^{m_2}$	$\pi_{a_3}^{m_3}$
4	3	7	1	1	8	0	1	2
5	3	6	2	1	7	0	2	1
	4	11	1	1	12	0	2	2
6	3	7	3	0	7	1	3	0
	4	11	2	1	12	1	3	1
	5	17	1	1	18	1	3	2
7	3*	9	4	1	10	1	3	0
	4	10	3	0	10	1	3	0
	5	15	2	1	16	1	3	1
	6	22	1	1	23	1	3	2
8	3*	11	5	3	14	1	4	0
	4	13	4	1	14	1	4	0
	5	14	3	0	14	1	4	0
	6	20	2	1	21	1	4	1
	7	28	1	1	29	1	4	2
9	3	12	6	7	19	0	5	0
	4*	15	5	3	18	0	5	0
	5	17	4	1	18	0	5	0
	6	18	3	0	18	0	5	0
	7	25	2	1	26	0	5	1
	8	34	1	1	35	0	5	2
10	3	14	7	12	26	0	5	0
	4	17	6	7	24	0	5	0
	5*	20	5	3	23	0	5	0
	6	22	4	1	23	0	5	0
	7	23	3	0	23	0	5	0
	8	31	2	1	32	0	5	1
	9	41	1	1	42	0	5	2

Stage $j = 2$ [$p_2 = m_2$, $p_1 = m_3$]

T_2	τ_2	f_2	T_1	F_1	F_2	$\pi_{a_1}^{m_1}$	$\pi_{a_2}^{m_2}$	$\pi_{a_3}^{m_3}$
4	3	7	1	1	8	0	1	2
5	3	4	2	1	5	0	1	1
	4	9	1	1	10	0	1	2
6	3*	2	3	0	2	1	1	0
	4	6	2	1	7	1	1	1
	5	12	1	1	13	1	1	2
7	3	3	4	1	4	2	0	0
	4*	3	3	0	3	2	0	0
	5	8	2	1	9	2	0	1
	6	15	1	1	16	2	0	2
8	3	7	5	3	10	3	1	0
	4	7	4	1	8	3	1	0
	5*	7	3	0	7	3	1	0
	6	13	2	1	14	3	1	1
	7	21	1	1	22	3	1	2
9	3	9	6	7	16	3	1	0
	4	10	5	3	13	3	1	0
	5	10	4	1	11	3	1	0
	6*	10	3	0	10	3	1	0
	7	17	2	1	18	3	1	1
	8	26	1	1	27	3	1	2
10	3	9	7	12	21	3	0	0
	4	11	6	7	18	3	0	0
	5	12	5	3	15	3	0	0
	6	12	4	1	13	3	0	0
	7*	12	3	0	12	3	0	0
	8	20	2	1	21	3	0	1
	9	30	1	1	31	3	0	2

Stage $j = 3$ [$p_3 = m_1$, $p_2 = m_2$]

T_3	τ_3	f_3	T_2	F_2	F_3	$\pi_{a_1}^{m_1}$	$\pi_{a_2}^{m_2}$	$\pi_{a_3}^{m_3}$
9	3^*	6	6	2	8	0	3	0
10	3^*	5	7	3	8	0	2	0
	4	9	6	2	11	0	3	0

Stage $j = 3$ [$p_3 = m_2$, $p_2 = m_1$]

T_3	τ_3	f_3	T_2	F_2	F_3	$\pi_{a_1}^{m_1}$	$\pi_{a_2}^{m_2}$	$\pi_{a_3}^{m_3}$
10	3^*	6	7	10	16	2	0	0

Stage $j = 3$ [$p_3 = m_3$, $p_2 = m_1$]

T_3	τ_3	f_3	T_2	F_2	F_3	$\pi_{a_1}^{m_1}$	$\pi_{a_2}^{m_2}$	$\pi_{a_3}^{m_3}$
10	3	20	7	10	30	2	5	0

Stage $j = 3$ [$p_3 = m_3$, $p_2 = m_2$]

T_3	τ_3	f_3	T_2	F_2	F_3	$\pi_{a_1}^{m_1}$	$\pi_{a_2}^{m_2}$	$\pi_{a_3}^{m_3}$
9	3	14	6	2	16	3	3	0
10	3	14	7	3	17	3	2	0
	4	20	6	2	22	3	3	0

LIST OF REFERENCES

Andrews, C. M., Elahi, S. M. & Clark, J. E. (1997). "Evaluation of New Jersey Route 18 OPAC/MIST Traffic-Control System," Transportation Research Record 1603, Transportation Research Board, National Research Council, Washington, D.C., pp. 150-155.

Barriere, J. F., Farges, J.-L. & Henry, J.-J. (1986). "Decentralization vs Hierarchy in Optimal Traffic Control," Proceedings of the 5th IFAC/IFIP/IFORS Symposium on Control in Transportation Systems, Vienna, Austria, July, pp. 209-214.

Bellman, R. & Dreyfus, S. E. (1962). "Applied Dynamic Programming," Princeton University Press, Princeton, N.J.

Ben-Akiva, M., Bierlaire, M., Koutsopoulos, H. N. & Mishalani, R. (1998). "DYNAMIT: A Simulation-Based System for Traffic Prediction and Guidance Generation," Paper Presented at TRISTAN III, San Juan, Porto Rico, June.

Bernauer, E., Breheret, L., Algers, S., Boero, M., di Taranto, C., Dougherty, M., Fox, K. & Gabard, J.-F. (1998). "A Review of Micro-Simulation Models," SMARTEST Project Deliverable D3, Institute for Transport Studies, University of Leeds, http://www.its.leeds.ac.uk/smartest, March.

Bretherton, R. D. (1989). "SCOOT Urban Traffic Control System: Philosophy and Evaluation," Proceedings of the 6th IFAC/IFIP/IFORS Symposium on Control, Computers, and Communications in Transportation, Paris, France, September, pp. 237-239.

Bullen, A. G. R., Hummon, N., Bryer, T. & Nekmat, R. (1987). "Enhanced Value Iteration Process Actuated Signals (EVIPAS)," Transportation Research Record 1114, Transportation Research Board, National Research Council, Washington, D.C., pp. 103-110.

Cameron, G. D. & Duncan, G. I. (1996). "PARMICS: Parallel Microscopic Simulation of Road Traffic," Journal of Supercomputing 10(1), pp. 25-53.

Chaudhary, N. A. & Messer, C. J. (1993). "PASSER IV: A Program for Optimizing Signal Timings in Grid Networks," Transportation Research Record 1421, Transportation Research Board, National Research Council, Washington, D.C., pp. 82-91.

Chaudhary, N. A. & Messer, C. J. (1996). "PASSER IV-96, Version 2.1, User/Reference Manual," Research Report No. 1477-1, Texas Transportation Institute, Texas A&M University, College Station, TX, October.

Chaudhary, N. A., Pinnoi, A. & Messer, C. J. (1991). "Proposed Enhancements to MAXBAND 86 Program," Transportation Research Record 1324, Transportation Research Board, National Research Council, Washington, D.C., pp. 98-104.

Cohen, S. L. (1983). "Concurrent Use of MAXBAND and TRANSYT Signal Timing Programs for Arterial Signal Optimization," Transportation Research Record 906, Transportation Research Board, National Research Council, Washington, D.C., pp. 81-84.

Dell'Olmo, P. & Mirchandani, P. B. (1995). "REALBAND: An Approach for Real-Time Coordination of Traffic Flows on a Newtowk," Transportation Research Record 1494, Transportation Research Board, National Research Council, Washington, D.C., pp. 106-116.

Dell'Olmo, P. & Mirchandani, P. B. (1996). "A Model for Real-Time Traffic Coordination Using Simulation Based Optimization," Advanced Methods in Transportation Analysis (L. Bianco and P. Toth, Editors), Springer-Verlag, Germany, pp. 525-546.

Dennis, J. E. & Schnabel, R. B. (1983). "Numerical Methods for Unconstrained Optimization and Non-Linear Equations," Prentice-Hall, Englewood Cliffs, N.J.

Dreyfus, S. E. & Law, A. M. (1977). "The Art and Theory of Dynamic Programming," Academic Press, New York, N.Y.

Fambro, D. B., Chaudhary, N. A., Bonneson, J. A., Messer, C. J. & Arabie, L. L. (1991). "PASSER III-90 Users Manual and Application Guide," Texas Transportation Institute, Texas A&M University, College Station, TX, March.

Farges, J.-L., Kamdem, I. & Lesort, J.-B. (1991). "Realization and Test of a Prototype for Real Time Urban Traffic Control," Proceedings of the DRIVE Conference on Advanced Telematics in Road Transport, Brussels, February, Vol. 1, pp. 527-542.

Farges, J.-L., Khoudour, L. & Lesort, J.-B. (1990). "PRODYN: On Site Evaluation," Proceedings of the 3rd IEE Conference on Road Traffic Control, London, England, IEE Conference Publication No. 320, pp. 62-66.

Farradyne Systems, Inc. (1990). "Traffic Adaptive Control, Phase I: Critical Intersection Control Strategies, Final Report," Prepared for National Cooperative Highway Research Program under Project 3-38(3), Rockville, MD, March.

Farradyne Systems, Inc. (1993). "Functional Specifications," Final Task A Interim Report, Submitted to Federal Highway Administration under Contract No. DTFH61-92-C-001, Rockville, MD, April.

Farradyne Systems, Inc. (1995). "RT-TRACS Prototype Control Strategies," Task D Working Paper, Submitted to Federal Highway Administration under Contract No. DTFH61-92-C-001, Rockville, MD, January.

Fellendorf, M. (1994). "VISSIM: Ein Instrument zur Beurteilung verkehrsabhängiger Steuerungen," In: Verkehrsabhängige Steuerung am Knotenpunkt, Forschungsgesellschaft für Strassen- und Verkehrswesen, Köln, Deutschland, pp. 58-68.

FHWA. (1976a). "Evaluation of UTCS Control Strategies, Executive Summary and Technical Report," Publication No. FHWA-RD-76-149 and -150, U.S. Department of Transportation, Federal Highway Administration, Washington, D.C., July.

FHWA. (1976b). "Urban Traffic Control System Traffic Adaptive Network Signal Timing Program, Vol. I, II, and III," Publication No. FHWA-RD-76-125, -126, and -127, U.S. Department of Transportation, Federal Highway Administration, Washington, D.C., August.

FHWA. (1976c). "Third Generation Control Software: Urban Traffic Control System (UTCS) Software Support Project, Vol. I-VI," Publication No. FHWA-RD-76-154/159, U.S. Department of Transportation, Federal Highway Administration, Washington, D.C., May.

FHWA. (1983). "SIGOP-III User's Manual," Publication No. FHWA-IP-82-019, U.S. Department of Transportation, Federal Highway Administration, Washington, D.C., July.

FHWA. (1985a). "SOAP84 User's Manual," Publication No. FHWA-IP-85-007, U.S. Department of Transportation, Federal Highway Administration, Washington, D.C., January.

FHWA. (1985b). "Traffic Control Systems Handbook, Revised Edition," Publication No. FHWA-IP-85-011, U.S. Department of Transportation, Federal Highway Administration, Washington, D.C., April.

FHWA. (1987). "Optimization of Left Turn Phase Sequence in Signalized Networks Using MAXBAND 86, Volume 2: MAXBAND User's Manual," Publication No. FHWA-RD-87-109, U.S. Department of Transportation, Federal Highway Administration, Washington, D.C., August.

FHWA. (1989). "Evaluation of the Optimized Policies for Adaptive Control Strategy," Publication No. FHWA-RD-89-135, U.S. Department of Transportation, Federal Highway Administration, Washington, D.C., May.

FHWA. (1991). "PASSER II-90 Users Guide," Series of Methodology for Optimizing Signal Timing (MOST): Volume 3, U.S. Department of Transportation, Federal Highway Administration, Washington, D.C., December.

FHWA. (1996). "Traffic Control Systems Handbook," Publication No. FHWA-SA-95-032, U.S. Department of Transportation, Federal Highway Administration, Washington, D.C., February.

FHWA. (1998). "TRANSYT-7F Users Guide," Series of Methodology for Optimizing Signal Timing (MOST): Volume 4, U.S. Department of Transportation, Federal Highway Administration, Washington, D.C. March.

FHWA. (1999). "TSIS User's Guide," U.S. Department of Transportation, Federal Highway Administration, Washington, D.C., June.

Gabard, J.-F. (1991). "CALIFE: An Adaptive Strategy for the On-line Calculation of Traffic Signal Plans in Urban Networks," Recherche Transports Sécurité (English Issue), N° 6, pp. 5-10.

Gabard, J.-F., Henry, J.-J., Abours, S. & Lesort, J.-B. (1986). "CALIFE: On-line Calculation of Fixed Time Plans," Proceedings of the 5th IFAC/IFIP/IFORS Symposium on Control in Transportation Systems, Vienna, Austria, July, pp. 233-237.

Gartner, N. H. (1983). "OPAC: A Demand-Responsive Strategy for Traffic Signal Control," Transportation Research Record 906, Transportation Research Board, National Research Council, Washington, D.C., pp. 75-81.

Gartner, N. H. (1985). "Demand-Responsive Traffic Signal Control Research," Transportation Research 19A, pp. 369-373.

Gartner, N. H. (1991). "Road Traffic Control: Progression Methods," Concise Encyclopedia of Traffic and Transportation Systems (M. Papageorgiou, Editor), Pergamon Press, New York, N.Y.

Gartner, N. H., Assmann, S. F., Lasaga, F. & Hou, D. L. (1990). "MULTIBAND: A Variable-Bandwidth Arterial Progression Scheme," Transportation Research Record 1287, Transportation Research Board, National Research Council, Washington, D.C., pp. 212-222.

Gartner, N. H., Stamatiadis, C. & Tarnoff, P. J. (1995). "Development of Advanced Traffic Signal Control Strategies for Intelligent Transportation Systems: Multilevel Design," Transportation Research Record 1494, Transportation Research Board, National Research Council, Washington, D.C., pp. 98-105.

Gartner, N. H., Tarnoff, P. J. & Andrews, C. M. (1991). "Evaluation of Optimized Policies for Adaptive Control Strategy," Transportation Research Record 1324, Transportation Research Board, National Research Council, Washington, D.C., pp. 105-114.

Greenough, J. C. & Kelman, W. L. (1998). "Metro Toronto SCOOT: Traffic Adaptive Control Operation," ITE Journal, http://www.ite.org, May.

Hadi, M. A. & Wallace, C. E. (1994). "Optimization of Signal Phasing and Timing Using Cauchy Simulated Annealing," Transportation Research Record 1456, Transportation Research Board, National Research Council, Washington, D.C., pp. 64-71.

Halati, A., Lieu, H. C. & Walker, S. (1997). "CORSIM: Microscopic Traffic Simulation Model for Integrated Networks," Paper Presented at the Transportation Research Board 76th Annual Meeting, Washington, D.C., January.

Halati, A. & Torres, J. F. (1992). "Enhancement of the Value Iteration Program Actuated Signals, Part 2: Final Report and EVIPAS Users Manual," Publication No. FHWA-PA-91-013+88-12, U.S. Department of Transportation, Federal Highway Administration, Washington, D.C., February.

Head, K. L., Mirchandani, P. B. & Sheppard, D. (1992). "Hierarchical Framework for Real-Time Traffic Control," Transportation Research Record 1360, Transportation Research Board, National Research Council, Washington, D.C., pp. 82-88.

Henry, J.-J. & Farges, J.-L. (1989). "PRODYN," Proceedings of the 6th IFAC/IFIP/IFORS Symposium on Control, Computers, and Communications in Transportation, Paris, France, September, pp. 253-255.

Henry, J.-J., Farges, J.-L. & Tuffal, J. (1983). "The PRODYN Real Time Traffic Algorithm," Proceedings of the 4th IFAC/IFIP/IFORS Symposium on Control in Transportation Systems, Baden-Baden, Germany, April, pp. 305-310.

Hunt, P. B., Robertson, D. I., Bretherton, R. D. & Winton, R. I. (1981). "SCOOT: A Traffic Responsive Method of Coordinating Signals," Laboratory Report 1014, Transport and Road Research Laboratory, Crowthorne, Berkshire, England.

Kell, J. H. & Fullerton, I. J. (1991). "Manual of Traffic Signal Design, 2nd Edition," Prentice-Hall, Englewood Cliffs, N.J.

Khoudour, L., Lesort, J.-B. & Farges, J.-L. (1991). "PRODYN: Three Years of Trials in the ZELT Experimental Zone," Recherche Transports Sécurité (English Issue), N° 6, pp. 89-98.

Land, A. & Powell, S. (1973). "FORTRAN Codes for Mathematical Programming," John Wiley and Sons, London, England.

Larson, R. E. & Casti, J. L. (1978). "Principles of Dynamic Programming, Part I: Basic Analytic and Computational Methods," Dekker, New York, N.Y.

Larson, R. E. & Casti, J. L. (1982). "Principles of Dynamic Programming, Part II: Advanced Theory and Applications," Dekker, New York, N.Y.

Little, J. D. C. (1966). "The Synchronization of Traffic Signals by Mixed-Integer Linear Programming," Operations Research 14(4), pp. 568-594.

Little, J. D. C. (1977). "Maximal Bandwidth for Arterial Traffic Signals: Theory and Interactive Computation," Working Paper WP 970-78, Alfred P. Sloan School of Management, Massachusetts Institute of Technology, September.

Little, J. D. C., Kelson, M. D. & Gartner, N. H. (1981). "MAXBAND: A Program for Setting Signals on Arteries and Triangular Networks," Transportation Research Record 795, Transportation Research Board, National Research Council, Washington, D.C., pp. 40-46.

Lowrie, P. R. (1982). "The Sydney Co-ordinated Adaptive Traffic System: Principles, Methodology, Algorithms," Proceedings of the International Conference on Road Traffic Signalling, London, England, March-April, IEE Conference Publication No. 207, pp. 67-70.

Luk, J. Y. K., Sims, A. G. & Lowrie, P. R. (1982). "SCATS: Application and Field Comparison with a TRANSYT Optimised Fixed Time System," Proceedings of the International Conference on Road Traffic Signalling, London, England, March-April, IEE Conference Publication No. 207, pp. 71-74.

MacGowan, J. & Fullerton, I. J. (1979-1980). "Development and Testing of Advanced Control Strategies in the Urban Traffic Control System," (Three Articles), Public Roads 43(2, 3 & 4).

Mahmassani, H. S., Hu, T.-Y., Peeta, S. & Ziliaskopoulos, A. (1994). "Development and Testing of Dynamic Traffic Assignment and Simulation Procedures for ATIS/ATMS Applications," Technical Report Prepared for U.S. Department of Transportation, Federal Highway Administration under Contract No. DTFH61-90-C-074-FG, June.

Martin, P. T. & Hockaday, S. L. M. (1995). "SCOOT: An Update," ITE Journal 65(1), January, pp. 44-48.

Mauro, V. & di Taranto, C. (1989). "UTOPIA," Proceedings of the 6th IFAC/IFIP/IFORS Symposium on Control, Computers, and Communications in Transportation, Paris, France, September, pp. 245-252.

Messer, C. J., Whitson, R. H., Dudek, C. L. & Romano, E. J. (1973). "A Variable-Sequence Multiphase Progression Optimization Program," Highway Research Record 445, Highway Research Board, National Research Council, Washington, D.C., pp. 24-33.

Mirchandani, P. & Head, L. (1998). "RHODES: A Real-Time Traffic Signal Control System: Architecture, Algorithms, and Analysis," Paper Presented at TRISTAN III, San Juan, Porto Rico, June.

Murtagh, B. A. & Saunders, M. A. (1993). "MINOS 5.4 User's Guide," Technical Report SOL-83-20R, Stanford University, Stanford, CA.

Pillai, R. S., Rathi, A. K. & Cohen, S. L. (1998). "A Restricted Branch-and-Bound Approach for Generating Maximum Bandwidth Signal Timing Plans for Traffic Networks," Transportation Research Part B: Methodological 32(8), pp. 517-529.

Robertson, D. I. (1986). "Research on the TRANSYT and SCOOT Methods of Signal Coordination," ITE Journal 56(1), January, pp. 36-40.

Robertson, D. I. & Bretherton, R. D. (1974). "Optimum Control of an Intersection for Any Known Sequence of Vehicle Arrivals," Proceedings of the 2nd IFAC/IFIP/IFORS Symposium on Traffic Control and Transportation Systems, Monte Carlo, September, pp. 3-17.

Sen, S. (1991). "Coordinated Optimization of Phases," Working Paper, Department of Systems and Industrial Engineering, University of Arizona, Tucson, AZ.

Sen, S. & Head, K. L. (1994). "Controlled Optimization of Phases (COP) at an Intersection," CRET²E Working Paper, University of Arizona, Tucson, AZ, November.

Sen, S. & Head, K. L. (1997). "Controlled Optimization of Phases at an Intersection," Transportation Science 31, pp. 5-17.

Stamatiadis, C. & Gartner, N. H. (1996). "MULTIBAND-95: A Program for Variable-Bandwidth Progression Optimization of Multiarterial Traffic Networks," Paper Presented at the Transportation Research Board 75th Annual Meeting, Washington, D.C., January.

Tarnoff, P. J. & Gartner, N. H. (1993). "Real-Time, Traffic Adaptive Signal Control," Proceedings of the Advanced Traffic Management Conference, St. Petersburg, FL, October, pp. 157-167.

Tsay, H.-S. (1989). "A Microcomputer-Based On-line Traffic Control System," ITE Journal 59(12), December, pp. 29-36.

Tsay, H.-S., Kang, J.-F. & Hsiao, C.-H. (1991). "Algorithm for Estimating Queue Lengths and Stop Delays at Signalized Intersections," Transportation Research Record 1324, Transportation Research Board, National Research Council, Washington, D.C., pp. 123-129.

Tsay, H.-S. & Lin, L.-T. (1988). "New Algorithm for Solving the Maximum Progression Bandwidth," Transportation Research Record 1194, Transportation Research Board, National Research Council, Washington, D.C., pp. 15-30.

Van Aerde, M. (1995). "INTEGRATION Release 2: User's Guide," Transportation Systems Research Group, Queen's University, Kingston, Ontario, Canada, December.

Vincent, R. A. & Peirce, J. R. (1988). "MOVA: Traffic Responsive, Self-Optimising Signal Control for Isolated Intersections," Research Report 170, Transport and Road Research laboratory, Crowthorne, Berkshire, England.

Webster, F. V. (1958). "Traffic Signal Settings," Road Research Technical Paper No. 39, Department of Scientific and Industrial Road Research Laboratory, Her Majesty's Stationery Office, London, England.

Webster, F. V. & Cobbe, B. M. (1966). "Traffic Signals," Road Research Technical Paper No. 56, Ministry of Transport, Her Majesty's Stationery Office, London, England.

Wolshon, B. & Taylor, W. (1998). "Analysis of Intersection Delay under Real-time Adaptive Signal Control," Paper Presented at the Transportation Research Board 77th Annual Meeting, Washington, D.C., January.

Yang, Q. (1997). "A Simulation Laboratory for Evaluation of Dynamic Traffic Management Systems," Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA, June.

BIOGRAPHICAL SKETCH

Shiow-Min Lin was born in Fengyuan, Taiwan, the Republic of China, on July 13, 1960. In 1981, while in college, he was awarded by the Ministry of Education as a member of the 1982 Youth Goodwill Mission of the Republic of China to promote Sino-American cultural exchange. During this capacity, he visited the United States of America for more than two months and performed in more than 20 states. In June, 1983, he received his Bachelor of Science degree from the Department of Industrial Design, the National Cheng Kung University.

Subsequently, he was commissioned a Second Lieutenant in the Republic of China Marine Corps. Upon release from the Marine Corps in 1985, he went back to the National Cheng Kung University, where he served as a teaching assistant in the Department of Industrial Design for the next two years.

In August, 1987, Mr. Lin came to the United States of America and enrolled at the University of Florida, where he studied industrial and systems engineering, specializing in operations research. He was awarded the Master of Science degree in December, 1988.

Following graduation, he accepted a position as a manufacturing coordinator at the Dean Steel Buildings, Inc., where he was responsible for production and inventory control.

In May, 1990, with a profound enthusiasm in transportation and traffic engineering, he came back to the University of Florida to pursue the degree of Doctor of Philosophy in Civil Engineering. During his study at the University, he was also engaged

in several highway work zone projects sponsored by the State of Florida Department of Transportation.

In 1992, Mr. Lin was awarded by the United States Department of Transportation as a recipient of the Dwight David Eisenhower Transportation Fellowship Program. In October, 1992, he relocated to Northern Virginia and started working full-time for the next three years in the Federal Highway Administration Intelligent Vehicle-Highway Systems Office. During this capacity, he was heavily involved in the research and development of the Federal Highway Administration's TRAF-family simulation programs.

In October, 1995, Mr. Lin accepted a position as a traffic engineer at the Viggen Corporation, and worked in the Federal Highway Administration's Traffic Management Laboratory. He continued to provide his expertise in traffic simulation. Also, he actively participated in several nationwide research and development projects involving Advanced Traffic Management Systems.

In February, 1997, Mr. Lin joined the Kaman Sciences Corporation, which later became the ITT Industries, Systems Division. Since then, he has worked in the Federal Highway Administration's Traffic Research Laboratory (formerly, the Traffic Management Laboratory) and continuously conducted the research and development in the areas related to the TRAF-family simulation programs.

Mr. Lin is married to Yiling Tu. He is a registered Professional Engineer in the Commonwealth of Virginia. He is a member of the Institute of Transportation Engineers, the Transportation Research Board, the Intelligent Transportation Society of America, the American Society of Civil Engineers and the Institute of Industrial Engineers.