ITS UCDAVIS
INSTITUTE OF TRANSPORTATION STUDIES

Research Report – UCD-ITS-RR-08-54

An Efficiency-Equity Solution to the Integrated Transportation Corridor Control Design Problem

Spring 2008

Jingtao Ma

Institute of Transportation Studies ◦ University of California, Davis
One Shields Avenue ◦ Davis, California 95616
PHONE (530) 752-6548 ◦ FAX (530) 752-6572
www.its.ucdavis.edu

An Efficiency-Equity Solution To The Integrated Transportation Corridor Control Design Problem

By

JINGTAO MA
B.S. (Tongji University) 1999
M.S. (Tongji University) 2001

DISSERTATION

Submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Civil and Environmental Engineering

in the

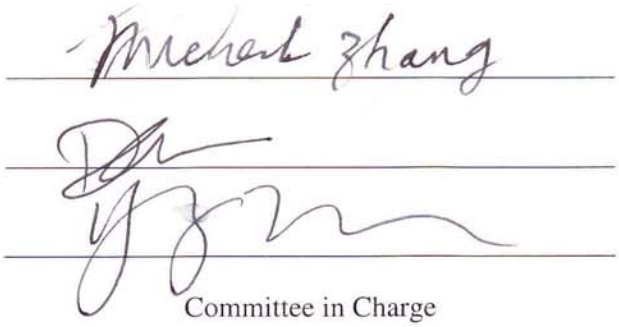
OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:



Committee in Charge

2008

An Efficiency-Equity Solution To The Integrated Transportation Corridor Control Design Problem

by

Jingtao Ma

B.S. (Tongji University) 1999
M.S. (Tongji University) 2001

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Transportation Engineering

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, DAVIS

Committee in charge:
Professor Michael H. Zhang, Chair
Professor Debbie A. Niemeier
Professor Yueyue Fan

Spring 2008

An Efficiency-Equity Solution To The Integrated Transportation Corridor Control Design Problem

Abstract

An Efficiency-Equity Solution To The Integrated Transportation Corridor Control Design Problem

by

Jingtao Ma

Doctor of Philosophy in Transportation Engineering

University of California, Davis

Professor Michael H. Zhang, Chair

This dissertation research proposes the integrated corridor control program based on the bi-criterion of both system efficiency and user equity. For years improving system efficiency, characterized by total travel time or travel delay reduction, has been considered the single most important measure in designing traffic control systems. From the perspective of the traveling public, however, the *fairness*, or how the benefits of the improvement is distributed, becomes more important as the transportation corridor systems get ever more congested. Incorporation of user equity in designing new control systems or updating existing ones will become prominent and urgent as the public is more and more involved in implementing these decisions.

The research began with an extensive review of the control practices and recognized three major deficiencies in the literature: 1) unclear and incomplete user equity measures and their fragmental applications in traffic control, 2) ad hoc and inaccurate modeling of the corridor traffic dynamics and subsequent congestion evolution and, 3) sub-optimal solutions of only isolated or sub-systems within the corridor.

Aiming to overcome the deficiencies, we first developed a general traffic flow dynamics model based on the kinematic wave model. Various control measures including urban signals, ramp metering and priority-rule controls were adapted and embedded into the flow dynamics model. Within this modeling framework, both the system efficiency measures and the user equity measures at aggregate and disaggregate levels were formulated and defined explicitly. One set of rule-based local synchronization corridor control schemes were then developed, and the schemes were compared with other control strategies of their efficiency and equity performances. We then proceeded to formulate the corridor control program that incorporates both efficiency and equity in the control objective. To solve the problem, one heuristic searching algorithm using simultaneous perturbation stochastic approximation (SPSA) was developed. The SPSA algorithm takes advantage of both heuristic searching and mathematical programming algorithms to quickly obtain the solution in much less computational burdens, while maintaining the realistic traffic flow model. Extensive numerical experiments were carried out to demonstrate the effectiveness of the proposed research, and the major findings were summarized in the last chapter.

To my parents

Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

List of Tables

Acknowledgments

This research could never have been finished without the guidance of my advisor and mentor - Professor Michael H. Zhang. A true scholar, he himself strives for the best quality of research and inspires his students with deeds rather than words. I will always feel indebted to his patience with my plodding progress, his responsiveness to my technical inquiries and his counsel. It is such a pleasure to study and research within the environment he created, independent of deadline pressure but only challenging for my best. I certainly think this style will continue to influence my professional career beyond my UC Davis years.

I have been very lucky to be guided by the wonderful faculty members that one can only hope for in his graduate studies: Professor Debbie A. Niemeier, Professor Patricia L. Mokhtarian, Professor Daniel Sperling, Professor John Harvey and Professor Yueyue Fan. Their challenge for the best quality of work as well as their gentle personalities have made the years most memorable and enjoyable.

Special thanks must be given to Professor James C. Spall at Johns Hopkins University, who provided valuable guidance in my development of Chapter Six at the early stage.

The fellow lab mates have always been my inspirations in both academic and daily life, as we spent so much time together discussing and enjoying great fun: Dr. Xiaojian Allen Nie, Dr. Wen-long Jin, Dr. Yu (Marco) Nie, Wei Shen, Haining Du, Sean Zhen Qian, Changmo Kim and Jinhyun Mun.

The friendship that grew with Song Bai, Peng Wu, Pengcheng Fu, Brett Williams, Nick Fontainey, Huawen Wu and Yuanxin Xi has made my life in Davis so colorful.

While finishing my dissertation, I started my professional life in Puget Sound area, where my supervisor Robert Shull, my colleagues Ed Hayes, Vicki Court-

ney and Chetan Joshi, my friends Hu Dong and Xiaoping Zhang, Richard Montague have provided very kind help and support.

I thank Zijing Li, my wife, for her unending support of my seemingly endless PhD pursuit and her constant assuaging comfort to alleviate the anguish and frustration in the journey. She upheld the tradition that UCD Transportation Engineering PhD students *will* graduate with a spouse.

Ultimately, I acknowledge my parents, to whom this work is dedicated, for their unconditional support throughout my life. Their love and comfort has been the genuine source of my strength to face all my challenges.

The research has been partially funded by California Department of Transportation (Caltrans) through the UC Berkeley PATH Program and by the Federal Highway Administration through Sustainable Transportation Center at ITS-Davis.

Chapter 1

Introduction

1.1 Background

Efficiency has been considered as the single most important measure in designing traffic control systems for years. From the system administrators' viewpoint, it is so natural to use measures like the total generalized cost (e.g., travel time/delay, fuel consumption and so on) to compare and select the 'best' control strategy, that nearly all control systems since the classical work of Webster (1958) are designed with this philosophy. However, recent practices begin to question whether it is appropriate to design a control system relying solely on efficiency. Questioned by urban travelers who endured long delays on metered ramps, for instance, the Minnesota DOT was mandated by the Minnesota legislature to conduct an eight-week ramp metering shut-off experiment to compare the system performances with and without metering(Levinson, Zhang, Das & Sheikh 2002). This experiment confirmed that ramp metering would favor long-distance travelers at the expense of short-distance ones. Understandably, those controls are not considered "efficient" from the perspective of those short-distance travelers. To gain public acceptance and support,

therefore, new or updated control systems must not only consider the overall efficiency improvements, but also how the efficiency gains are distributed among the system's user groups who differ from one another in their departure time, origin-destination, trip purposes, and value of time. Equity, or user fairness, begins to emerge as an important issue in traffic control design.

Compared with the overwhelming number of studies in addressing control efficiency, the analysis of equity issues in traffic control is far less, and their findings are mostly qualitative and sometimes even conflicting. For instance, efficiency and equity are usually considered as two competing requirements: the more efficient a control system is, the less equitable it becomes (Kotsialos & Papageorgiou 2004)(Meng & Yang 2002). However, some microscopic simulation (e.g., Yin et al., 2000) finds that the above claim may not be true because a control algorithm could be more equitable than another, yet maintains a similar level of overall efficiency. Apparently, there is a need to resolve these contradictory findings, and develop proper equity measures to be used in the design of control systems that balance efficiency and equity.

The entanglement around the two requirements originates from several threads. Firstly, the requirements themselves, especially the equity requirement, are not clearly identified. The equity measures applied in the literature are inherently deficient because they are borrowed mainly from social welfare studies in economics, where an appropriate measure is not generally agreed upon(Bowman 1945). Secondly, no control design has taken equity explicitly as its objectives and consequently the equity requirements are only marginally taken care of by some practical constraints such as minimum queuing time and so on. Thirdly and most importantly, control systems implemented in isolation hinder a more efficient manipulation of control elements and the achievement of global efficiency and equity requirements, since current control systems only take care of a sub-network of urban signalized intersections or

metered ramps. To disentangle theses complex issues, we propose a comprehensive study to identify the needs of efficient and equitable control, develop unambiguous measures of equity and efficiency, and design control methods that consider explicitly both efficiency and equity requirements.

1.2 Problem Statement and Conceptual Formulation

In this study of traffic control design, we make the following statement concerning the *user equity* [1] aspect of traffic control design:

An equitable control system demonstrates:

(i) *the total difference of trip costs between any two user groups as minimal;*

(ii) *the maximum trip cost experienced by all user groups as minimal.*

Another statement can be made concerning *system efficiency* of a corridor control system, i.e., the total travel time or delay reduction benefits that all network users can experience collectively:

An efficient control system demonstrates the overall delay of all network users as minimal.

The control design that balances both system efficiency and user equity will then be stated as:

An efficient-equitable control system is one in which no control changes can make neither system efficiency nor user equity better off without making the other worse off.

[1] *User fairness* is also used to describe how the delay-reduction gains from improved traffic control are distributed among network users. Both terms, *equity* and *user fairness*, are interchangeable in this study.

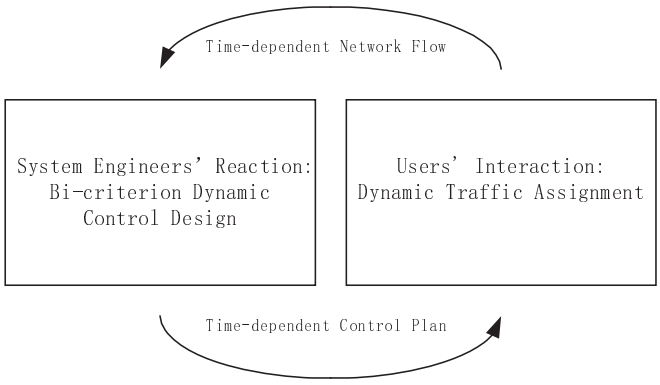


Figure 1.1: Bi-criterion Control Design Conceptual Model

In other words, an efficient-equitable system will reside in a state of Pareto optimal. The arguments for these statements will be fully developed in later chapters.

Towards the goal of achieving both system efficiency and user equity in designing a corridor control system, two interactions must be well studied. The first interaction is the one among individual travelers, and the second is the interaction between the system engineers and the travelers. The first interaction is generally studied in the field of *traffic assignment*. Traffic assignment deals with the problem of mapping the given travel demands onto the network based on certain traveler behavior assumptions. The results from the traffic assignment is the network flow pattern in the form of link flows (or path flows). Link travel cost (or path travel cost) will also be available from traffic assignment. The second interaction is investigated based on the link flows and travel costs from the traffic assignment[2]. The conceptual modeling flow of the two interactions can be illustrated in Figure 1.1.

[2]To note that it is also possible to reveal the second interaction based on observed traffic flow patterns from surveillance systems; but in this work we will study both interactions in an integrated manner.

The two key terms from the conceptual Figure 1.1 are bi-criterion dynamic control design and dynamic traffic assignment. Bi-criterion control refers to explicitly taking both system efficiency and user equity measures into consideration when computing the optimal control plan. dynamic assignment and control refers to the modeling corridor network congestion both spatially and temporally, i.e., tracking the evolution of traffic congestion and computing the control plan in a time-variant or time-dependent manner.

The efficiency-equity bi-criterion corridor control model is conceptualized as follows. A transportation network is given in the node-link representation $G(\mathcal{N}, \mathcal{L})$, where $\mathcal{N}$ and $\mathcal{L}$ are the sets of nodes and links, respectively. The demand is assumed to be known a priori and given in the form of (time-dependent) origin-destination (O-D) matrix $Q(\mathcal{R}, \mathcal{S})$, where $\mathcal{R}$ and $\mathcal{S}$ are the sets of trip origins and destinations, respectively $q^{r,s}$ is the demand between O-D pair (r, s). Let K_{rs} denote the set of paths connecting the O-D pair $r - s$ and the entire path set $K = \bigcup K_{rs}, \forall r, s$. In this way, the network users are differentiated by their origin-destination and path choice characteristics.

Starting with an empty network at $t = 0$, each network user group $q^{r,s}(t)$ will be assigned onto the network, and we shall have

$$q^{r,s} = \int_0^T q^{r,s}(t) d_t$$

Assume all travelers will finish their journey at T' and following certain network study conventions(Nie 2006), we first clarify a few terms in our modeling:

Definition 1.1 (Assignment interval ϕ_a) *A discrete period during which any departure flows will hold constant. The assignment horizon T consists of m_a assignment intervals $T = m_a \times \phi_a$.*

Definition 1.2 (Loading interval ϕ_l) *A discrete period during which the network*

flow conditions are stationary. The loading horizon T' consists of m_l loading intervals $T' = m_l \times \phi_l$.

Further we use $c_k^{r,s}(t)$ to denote the travel cost of the path connecting O-D pair (r, s) departing at t along path k.

Corresponding to the *assignment interval* and *loading interval*, we also have the following time differentiation in relation to control updating:

Definition 1.3 (Control interval ϕ_c) *A discrete period during which all control variables are fixed. The control horizon T_c consists of m_c assignment intervals $T_c = m_c \times \phi_c$.*

Within the network $G(\mathcal{N}, \mathcal{L})$, a subset of the nodes $\mathcal{N}_C$ contains all the nodes with certain kinds of control measures (urban signals, ramp meters or yield/STOP signs and so on). The system designer can have the following instruments *or* control variables at hand to affect the network flow pattern:

- Signal greensplits $g_i^m(t)$ for the phase m signalized intersection i or metering rates $R_i(t)$;
- Phasing and phasing sequence $\eta_i(t)$ for the signal controller at signalized intersection i;
- Cycle length $C_i(t)$ at intersection i;
- Offset $\Delta_i(t)$ of the signal controller at intersection i;

We collectively denote the vector of the control variables

$$(g_i^m(t), R_i(t), \eta_i(t), C_i(t), \Delta_i(t)), i \in \mathcal{N}_C, t \in [0, T_c] \tag{1.1}$$

as vector θ.

By changing the control vector θ, the system engineer can affect the path flow $q_k^{r,s}(t)$ and their corresponding path travel cost $c_k^{r,s}(t)$. Given a fixed control plan θ, the users $q^{r,s}(t)$ will select the path $\kappa_{r,s}(t)$ to minimize their travel cost, and lead to the (time-dependent) path flow solution of dynamic traffic assignment, $\mathbf{f}^*$.

Naturally the system efficiency could be conveniently measured by the total network travel time:

$$TTT = \sum_{(r,s)} \sum_{p} \sum_{t} c_p^{r,s}(t) \times f_p^{r,s}(t)$$

and in a vector form as

$$TTT = \mathbf{c} \bullet \mathbf{f}$$

In contrast to defining system efficiency, quantifying user equity requires examination of the traveler groups $q_p^{r,s}(t)$ at both individual and collective level. Furthermore, measuring user equity also depends on the system engineer's understanding and definition of user equity, as illustrated in our version of the equitable system. Herein we only define the collective and the individual equity measures in abstract form as as $\Psi(\mathbf{c}, \mathbf{q})$ and $\Omega(\mathbf{c}, \mathbf{q})$, respectively.

Let us further use $\mathcal{B}(\theta, \mathbf{f})$ to represent one combined function of the measures of TTT, Ψ, Ω, then the bi-criterion corridor control design problem will be conceptually stated as:

$$\begin{array}{rl} \min_\theta & \mathcal{B}(\theta, q^*) \\ \text{subject to} & \theta \in \Theta \end{array}$$

where $\mathbf{f}^*$ comes from the above dynamic traffic assignment and Θ is the set of feasible control plans.

the Pareto Optimal frontier in the cardinal system of system efficiency and user equity can then be drawn from the definitions of $\mathcal{B}(\theta, \mathbf{q})$. In this study we investigate that under any fixed flow pattern $\mathbf{q}$, the appropriately computed control plan

will improve the efficiency and equity measures along the Pareto Optimal frontier. We will thus focus our study on building the bi-criterion corridor control program based on solid corridor traffic flow modeling.

1.3 Contributions

This dissertation work offers a few unique aspects in traffic control research and practice. Primarily, we recognized the importance of including user fairness besides system efficiency in designing traffic control systems and explicitly modeled user equity in traffic control at both aggregate and disaggregate level. Secondly, the traffic flow dynamics under traffic control within a general corridor network has been modeled based on a finite difference solution scheme of the kinematic wave model and the control measures, including signal control and ramp metering, are embedded coherently into the model. By comparing the system performances under various control strategies, we concluded that user equity measures could be significantly improved, while maintaining the similar system efficiency level. Finally, we established a mathematical program to optimize the integrated control measures so as to balance the system efficiency and user equity measures. Extensive tests were utilized to investigate the interaction between these two dimensions. This research indicates the importance of incorporating user equity explicitly into the corridor control system design and offers one solution.

1.4 Thesis Outline

The main body of the dissertation is organized as follows: Chapter 2 provides a sweeping review of the traffic control literature that includes efficiency and equity measures deployed in control system design, development of integrated signal

control systems and their underlying traffic flow dynamics models as well as solution algorithms. In Chapter 3, a cell transmission model based network flow dynamics study tool is built and urban signal control and ramp metering are modeled coherently and embedded in the tool. This tool serves as the basis for evaluating the performance of various control strategies. To investigate the efficiency and equity performances of different control strategies, an evaluation and comparison study is carried out in Chapter 4. One set of local synchronization control schemes has been developed and compared with prevalent control strategies so that we can understand the interaction between efficiency and equity measures under different controls. In Chapter 5, an integrated corridor control framework is proposed and various forms of the problem are formulated to balance the system efficiency and user equity. In Chapter 6, one special type of heuristic solution method that takes advantage of both the heuristic search and the gradient information is developed to solve the integrated corridor control problem. In Chapter 7, both contrived and real networks were used to illustrate the effectiveness of the proposed solution to the integrated corridor control problem. Chapter 8 concludes the dissertation study and discusses future research directions.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 2

Literature Review

To fully appreciate the merits of previous studies concerning traffic control system design, three aspects are of particular interest: the control objectives or performance evaluation measures, the underlying traffic flow models, and the control design or optimization algorithms.

2.1 Control Objectives: System Efficiency versus User Fairness

As identified in the introduction, reducing the overall travel cost for travelers has been overwhelmingly the single most important objective for most traffic control system design. But a newly designed or an updated control system would have two effects, the "generative" effect and the "distributive" effect (Lakshmanan, Nijkamp, Rietveld & Verhoef 2001). In any society, the transportation sector, due to its complex nature, impacts the society both internally and externally. The generative effect refers to the net social welfare improvements resulting from investment in transportation systems. It characterizes the gains and losses for all network trav-

elers. The latter occur when some of the positive effects are compensated for by the negative ones. Translated into evaluation of the performance of control systems, the generative effect is equivalent to the efficiency performance and the distributive effect is equivalent to the equity performance. In this section, we will examine how both effects have been treated in different control system designs and their effectiveness. The discussion begins with the equity concerns in planning processes.

2.1.1 Equity Concerns in Planning

Used in the Transportation Equity Act (TEA) and the Intermodal Surface Transportation Efficiency Act of 1991 (ISTEA), the term *Equity* conventionally referred to judging the funding allocation policies among jurisdictions and agencies or between travel modes. As a matter of fact, equity or user fairness issue arises when the distributive effect of any transportation policy or project is concerned. For example, in (Garrett & Taylor 1999), the ridership demographics analysis indicated that in the Los Angeles area the public transit funding favors the population groups in low-density suburb areas instead of those in high-density city center areas where public transit service is mostly needed. This dissonance results in a both economically inefficient and socially inequitable public transit system.

A generally accepted taxonomy for evaluation of equity issues in transportation systems was proposed by (Litman 2007). In his work, two general categories of equity are identified:

- Horizontal equity: or *egalitarianism*, is concerned with treating everybody equally, regardless of of factors like race and income. It implies that public policies should avoid favoring certain individual or groups over the others, and the consumers should "get what they pay for and pay for what they get".

- Vertical equity, or *social justice*, is concerned with the distribution of the ben-

efits or losses between individuals or groups that differ in needs and abilities such as income, social class, and in particular, mobility needs and ability. If a policy favors the socially or economically disadvantaged groups, it is considered equitable because it compensates for the overall social inequities. This type of feature policy is called *progressive* ones. If the distribution impact is the oppositive, the policies would be *regressive*.

Equity principles that can reflect the progress toward planning or operational objectives include(Litman 2007):

- treating everyone equally, unless the treatment is justified for special reasons;
- allocating the costs to individuals who impose them;
- being progressive regarding income;
- benefiting transportation disadvantaged people (non-drivers, disabled, children, etc.);
- improving basic access: favors trips considered necessities rather than luxuries.

These equity classification principles and associated measurement indicators have been applied to analyze various types of transportation policies such as public transit subsidies, parking restrictions and congestion pricing (Ma, Sun & Sperling 2005). In the studies, only one or too few indicators are combined to evaluate the equity aspect of the process or the project.

One pioneering work (Vaughan 1985) explored the equity solution to the trip distribution problem. By analyzing the land costs, mostly taxes, associated with homes and workplaces of different locations in Sydney, Vaughan discovered that the disutility of crowding is proportional to a power of the sum of home and

workplaces. He combined the land costs with travel distance as the commute cost (or more formally in trip distribution terms, the friction factor):

$$C_{ij} = k_d|i-j| + k_c(H_i + H_j)^2 \tag{2.1}$$

where C_{ij} is the cost of travel between zone i and j, H_i is the number of houses in zone i and k_d, k_c are the unit costs of travel and crowding respectively. He then formulated the trip distribution problem as a linear program to solve the distributed trips between any zone i and j, T_{ij}. The total cost

$$\sum_i \sum_j C_{ij}T_{ij}$$

is used for evaluating the generative effect, while he further argued that an equitable solution to the trip distribution problem would be to equalize the average trip costs among different zones. The average cost k_j to an individual worker in zone j is

$$k_j = \frac{k_j^*}{W_j} = \frac{\sum_{i=1}^n C_{ij}T_{ij}}{W_j}, \ \ \forall j = 1, ..., n \tag{2.2}$$

where k_j^* is the costs occurring for all workers in zone j. By specifying an additional equity constraint

$$\sum_{i=1}^{n} C_{ij}T_{ij}/W_j - C_{in}T_{in}/W_n = 0, \ \ \forall j = 1, 2, ..., n-1 \tag{2.3}$$

he formulated the equitable trip distribution problem as to minimize the average cost among all zones:

$$\text{Minimize } k = \sum_{i=1}^{n} C_{in}T_{in}/W_n \tag{2.4}$$

with the standard trip distribution constraints of

$$\sum_j T_{ij} = H_i, i = 1, ..., n; \ \ \sum_i T_{ij} = W_i, j = 1, ..., n; \ \ T_{ij} \geq 0, i, j = 1, ...n \tag{2.5}$$

To note that the above mathematical program is a linear one in that C_{ij} is constant and the objective functions and all constraints are all linear equations.

An efficiency index was then simply developed as the following ratio:

$$E = \frac{e - m}{M - m} \tag{2.6}$$

where m, M and e are the minimum, maximum average commute and crowding cost and the cost from the solution, respectively. The results from the linear city case showed, however, the solution resulting from (2.4) did not always reach a truly equitable trip distribution solution, as the standard deviation of commuting and crowding costs are increasing among all commuters, even though the average commute and crowding cost remained the same between different zones. As he concluded, this deficiency would be reduced by addition of measures to confine the dispersion of the average cost in the formulation.

Equity Measures in Network Design Problem (NDP) and Route Guidance System Formulations

One general category of transportation network research deals with identifying the best projects for funding to improve the overall network performance. This category of research is called network design problem (NDP)(Friesz 1985)(LeBlanc & Boyce 1986)(Yang & Bell 1998). Only recently does equity aspect become one concern.

In (Meng & Yang 2002), the authors approached the conventional NDP with an added equity measure that takes the cost ratio of the trips before and after the network improvement projects. Their actual equity index is called *critical trip cost ratio*, defined as the maximum of the trip cost ratios among all trips. In the formulation, the critical trip cost ratio becomes a constraint that must be under a threshold

value to restrict the most disadvantaged group from being penalized too much. The solution to the new NDP problem would be a set of projects that both improve the overall network performance and satisfy the cost ratio constraints. Furthering their deterministic route choice behavior, Chen and Yang (Chen & Yang 2004) considered randomness of both the demand and the users's route choice behavior. The same critical trip cost ratio is still applied as the equity index.

Szeto and Lo (Szeto & Lo 2006) recognized the lack of time dimension in the NDP formulation where the long planning horizon, usually over a few years, is not considered. They brought the planning horizon into the formulation by discounting all cost measures by time. The system efficiency is measured by discounted social surplus that comprises two components: the discounted consumer surplus and discounted profits. The discounted consumer surplus is defined as:

$$CS_{\pi}^{rs} = \left[\int_{0}^{q_{\pi}^{rs}} D_{\pi}^{rs-1}(\nu) d_{\nu} - \varpi_{rs}^{\pi} q_{rs}^{\pi} \right] \tag{2.7}$$

where $D_{\pi}^{rs-1}(\bullet), q_{rs}^{\pi}, \varpi_{rs}^{\pi}$ are, respectively, the inverse of demand function, the travel demand and travel cost for O-D pair (r, s) in period π. At the same time, the discounted profits are simply the difference between the discounted toll revenue and the construction cost of the projects. The user equity takes a variation of mean difference:

$$G = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} |w_i - w_j|^2 \tag{2.8}$$

Besides the same set of constraints i.e., user equilibrium flow pattern and budget limit, the model imposes an upper limit ϵ to the equity measure G and formulated a so-called $\epsilon-$constraint model.

One recent work (Jahn, Öhring, Schulz & Stier-Moses 2005) improves the equity aspects in route recommendations for (static) route guidance system design. The objective is still to minimize the system cost, but the routes must be within

the threshold of the equity index. The equity index is defined similarly to (Meng & Yang 2002), where the ratio becomes the path cost to some nominal cost. The nominal cost can be: 1) the lowest trip cost of users that share the same O-D features (loaded unfairness), 2) the cost of the shortest-distance path (normal unfairness), 3) the cost of the same O-D pair from a user equilibrium flow pattern (user equilibrium unfairness) and, 4) the cost of fastest possible path for the same O-D pair, i.e., the free flow travel cost (free-flow unfairness). Adding the upper bound constraint for the equity index, the formulated program becomes a constrained system optimal (CSO) problem. The solution to the CSO problem offers advantages compared to those from either system optimal or user equilibrium solely, since it eliminates lengthy detours by restricting the path cost from deviating too far from the nominal cost, the travel cost from user equilibrium flow pattern in their case, while the total system cost still remains close to the unconstrained system optimal cost.

The above equity measures all examine the gains and losses of the travelers at their individual level. A more complete formulation that addresses equity for the travelers as a whole was done in (Feng & Wu 2003). Replacing travel time with average travel speed as the measure of travel cost, they defined both horizontal equity and vertical equity(Litman 2007) with regard to jurisdictions. Horizontal equity can be calculated as:

$$HTCM_{ij} = VTCM_{ij} = \frac{L_{ij}}{TCM_{ij}} \tag{2.9}$$

where $HTCM_{ij}, VTCM_{ij}, L_{ij}$ and TCM_{ij} are, respectively, the horizontal equity measure, vertical equity measure, trip length along the shortest path for city i to its center of region j and its associated trip cost measure from city i to region center in region j. The horizontal and vertical equity objectives in the NDP formulation thus

read:

$$\frac{1}{\sum N_i}\sum_i\sum_j\left[\left(HTCM_{ij}-\frac{1}{N_i}\sum_j HTCM_{ij}\right)^2\right]^{1/2} \tag{2.10}$$

and

$$\left[\frac{1}{\sum N}\sum_i\left(\frac{1}{N_i}\sum_j VTCM_{ij}-\frac{1}{N_i}\sum_i\sum_j VTCM_{ij}\right)^2\right]^{1/2} \tag{2.11}$$

By introducing the imbalance measure of both intra-zonal (2.10) and inter-zonal (2.11) trip costs, the NDP solution could avoid the wide dispersion of the improvement benefits, and the service quality in terms of average travel speed is also equalized.

2.1.2 User Fairness in Traffic Control System Design

In design of timing plans for signalized intersections, the conventional methods treat the traffic merely as static volumes of conflicting movements that require right-of-way alternatively. The typical example is Highway Capacity Manual(TRB 2000). Its general design process is to obtain the traffic counts for multiple days, and consequently determines a design hour volume, usually based on a certain percentile of the traffic counts, as the base flow rate for the analysis period. With a given phase sequence and phase groups, the method can determine how much green time within a cycle will be allocated to each phase, or the *green splits*. One fundamental difference of these methods is the design logic to allocate green splits; and these logic will affect how efficient and equitable a timing plan can be. Three major logics have been developed as follows.

Right-of-way Allocation Policies

"Equi-saturation" policy The Equi-saturatoin policy is the canonical form of determining the green splits within a cycle, following Webster's classical work(Webster 1958).

Under the Equi-saturation policy, the green time is determined in such a way that the phase duration will be proportional to its *critical volume-to-capacity (v/c) ratio,* f_a. The critical v/c ratio refers to the v/c ratio of the critical movement in each phase, i.e., the movement with the largest saturation ratio:

$$Eq_A Max_a \frac{f_a}{\lambda S} \tag{2.12}$$

Here the notation Eq is used to denote a process that equates the quantities, through all phases a in the collection A in this case. The resulting degree of saturation is identical for every phase.

Delay minimization (Delmin) policy Allsop (Allsop 1971) proposed a policy that minimizes the total intersection delay:

$$Min \Sigma_a f_a d_a$$

where the delay on each approach is calculated with the Webster's delay formula. This delay minimization policy is actually done through

$$Eq_A \Sigma_a f_a \frac{\partial d_a}{\partial \lambda_a} \tag{2.13}$$

A unique solution can be obtained since Webster's formula (2.15) is a convex function with respect to green time g given a flow pattern q.

The capacity maximization policy P_0 Another policy is the P_0 policy by Smith (Smith 1979*a*) (Smith 1979*b*). He defined *traffic pressure* as the product of the

approach capacity S_i and its average delay d_i for approach link i. This policy maximizes intersection capacity through balancing the traffic "pressures" of conflicting approaches:

$$Eq_i S_i d_i \tag{2.14}$$

This policy favors the road sections with greater capacity S. The P_0 policy has been designed with special consideration of solving combined control-assignment problems(Smith 1981). Understandably, this policy redistributes the flows from the congested links to the links with higher saturation flows, and the links with higher capacity thus receive more green times.

These policies and associated methods are mostly developed to handle an isolated intersection for a fixed flow pattern. When put in a network context of changing flow patterns, e.g., travelers that adjust their route choice behavior through day-to-day driving experiences, these methods were found ineffective to establish an stable system (van Vuren & van Vliet 1992)(Smith 1980) (Smith 1981).

For safety concerns, some measures, e.g., the minimum and maximum green time, have been taken in control design practices to distribute the delays at intersections controlled by vehicle actuated controllers. The same safety concerns are more strongly emphasized in transit signal priority or railway preemption control as well as the signalized intersections with high volumes of pedestrians. Similarly in ramp metering, The Minnesota Stratified Zonal Ramp metering algorithm also explicitly imposes the maximum queue length into the ramp metering (Levinson 20003), (Zhang & Levinson 2004).

2.1.3 Aggregate Equity Measures in Evaluation of Control Systems

Most above studies actually focus on preventing the transportation system from extreme situation where certain individuals or traveler groups are severely dis-

advantaged. By restricting the ratios from an upper limit bound, the researchers hope that the disadvantaged travelers will not sacrifice too much to compensate for others' travel. This is certainly a true claim; but relying on disaggregate measure solely does not fully capture how the system cost are dispersed, as was recognized in (Vaughan 1985). Ideally, the benefits or losses should be more or less evenly dispersed and the dispersion can be best characterized by aggregate measures or statistics.

In social welfare studies, Gini Coefficient (Gini 1936) and the associated Lorenz Curve (Lorenz 1905) have been used to depict the income disparity of a society. These measures were borrowed to capture the dispersion of the control system cost savings. In metering shutdown experiments in the Twin Cities, the Lorenz Curve is plotted with the absolute terms of travel delay (Levinson et al. 2002) for the freeway travelers. In (Yin, Liu & Benouar 2000), the Gini Coefficient is calculated using the trip cost ratios before and after applying ramp metering in a southern California corridor. In this work, an interesting remark was drawn that certain control strategies may reach a similar level of system efficiency yet get a better equity performance measured in aggregate indexes.

However, the aggregate measures themselves, including Gini Coefficient and the associated Lorenz Curve, have had deficiencies themselves in capturing the overall income disparities in economic studies. For example, they may well cover up the lower end of the income spectrum, implying the low-income groups were poorly reflected on the Lorenz Curve. Meanwhile, different measures could draw conflicting conclusions for the same population and income data set, as revealed in analyzing the American society in the 1940s(Bowman 1945). Translated into the control system design, it would mean that the mostly delayed travelers are not reflected well enough in the Gini Coefficient. Therefore, a single aggregate measure may also distort the results.

2.1.4 System Efficiency Measures

To evaluate the performance of a control plan, the most commonly used efficiency measure is *travel delay*, i.e., the extra travel time beyond free flow conditions. When the travelers do not switch routes, the minimization of travel delay is equivalent to the minimization of travel time; otherwise, travel time and travel delay are different measures of the system efficiency in the network context.

Additionally, other efficiency measures have been developed to better evaluate the effectiveness of a control plan. For example, $TRANSYT-7F$(Wallace, Courage, Hadi & Gan 1998) can take the combinations of the following eight measures:

- Disutility index (DI), a linear combination of delay and number of stops, to be minimized;
- Progression opportunity ($PROS$), the number of signals that the vehicle can travel through at its design speed without stopping;
- $PROS$ then DI, ranked dual objectives;
- $PROS/DI$, a combined objective;
- Queue Ratio (QR) $\times DI$, a combined objective; and QR is defined as the average back of the queue on a link divided by the number of vehicles accommodated by the link, to be minimized;
- Throughput($THRU$)$/DI$, a combined objective, throughput is the number of vehicles serviced by all approaches to an intersection within a given time, to be maximized;
- $THRU$ then DI, a ranked dual objective;

- $THRU\ V/C$, the volume capacity ratio of the intersection, to be minimized.

One notices that some of the above efficiency measures or control objectives can be derived from each other, but different aspects of the system efficiency have been characterized. As can be seen from the next section (2.2), the control objectives are generally variations of the above efficiency measures.

2.1.5 A Summary of Control System Objectives

Assessing the control systems has overwhelmingly come from the system administrators' perspective, no matter what concerns they might have (congestion alleviation, safety improvement or emission control). Nonetheless, the equity aspect emerges as an important criterion from control practices, e.g., ramp metering in Minnesota. Both aggregate and disaggregate equity measures have been sporadically studied to understand how the benefits from the control system are distributed, but no holistic picture has ever been drawn to truly reveal the interaction between these two different dimensions in control system design. Explicitly or implicitly, system efficiency and user equity have been considered competing criteria; one can not profit on one dimension without sacrificing the other(Kotsialos & Papageorgiou 2004). However, as implied by some researchers(Yin et al. 2000), balancing the efficiency and equity in one control system may not necessarily be a zero-sum game; one can maintain a similar level of efficiency yet achieving better equity performances. This will require a better design to fully grasp the work mechanism between various control objectives.

2.2 An Overview of Traffic Control Systems

This section provides a brief survey of traffic control systems and methods. By no means is this review intended to fully describe the research and practice in this field as this subject itself would require volumes to complete, but it is to study the advances in traffic modeling and algorithms from representative documentations.

For a retrospect of history, perhaps the earliest written law on traffic control is back to 1652 when the city of New Amsterdam (New York City) banned the driving of wagons or carts at a gallop within the city limits (Kane 1964). Whereas the same safety rationale still plays a major role in today's control practices, the capacity constraint and consequently allocating the right-of-way to conflicting traffic streams eventually emerged as the primary concern in the pre-automotive era as early as 1791. The practice of directing the traffic with hand signals of either "STOP" or "GO" by a police officer began in 1903 in New York City. The "stop and go" hand semaphore appeared to be the most efficient control device until the electronic traffic signal controller was introduced in the 1910s. The earliest documented installation of an electronic controller was seen at the intersection of Euclid Avenue and East 105th Street in Cleveland in 1914(Benesch 1915). Thereafter, the 24-hour-per-day working *automated* electronic signal controllers have been favored. The ideas and analysis tools such as time-space diagram and signal synchronization and coordination among others have been developed and applied in practice. Herein we begin the review with Webster's classical work of designing the control plan for an isolated intersection.

2.2.1 Urban Street Signal Control Systems

This section focuses on the design method and their assumptions on traffic dynamic characteristics. Two aspects of transportation system control will be reviewed in detail, namely traffic responsiveness and control coordination.

Pre-timed Signal Control

F.V. Webster, among the first researchers in the 1950s, used a commercially available computer to analyze complex systems. He simulated the traffic arriving at an intersection with a given average rate (λ of Poisson distribution). One of his great contributions was the delay formula:

$$d = \frac{c(1-g/c)^2}{2[1-(g/c)x]} + \frac{x^2}{2q(1-x)} - 0.65\left(\frac{c}{q^2}\right)^{\frac{1}{3}} x^{2+5(g/c)} \tag{2.15}$$

where d is the average delay per vehicle in seconds, c is the cycle length, g/c is the green ratio and x is the degree of saturation, the ratio of the volume q to the allocated capacity S determined by green time g

$$x = \frac{v}{gS}$$

The first term in (2.15) is the traffic delay of constant arrival and the second term is the "random delay" from the assumption of Poisson arrival. The third term is a correction term to account for the simulation and observation errors that were often omitted in analytical work. Webster formula is so successful and widely used that prevalent HCM design method is simply its extension with adjustments to the flow rates q and the saturation flow rate S.

Under the "Equi-saturation" logic, the green split or effective green time is calculated in such a way that the degree of saturation for critical movement in every phase is balanced to be the same. For phase m that consists of several traffic movements a , define y_m as the critical flow ratio for the phase:

$$y_m = \max_a \frac{f_a}{\lambda s_a} \tag{2.16}$$

Then the cycle length is given by:

$$C = \frac{1.5L + 5}{1 - Y} \tag{2.17}$$

where Y is the summation of all y_m over all phases, and L is the total lost time for a cycle. The allocated green time for phase m is:

$$g_m = (C - L)\frac{y_m}{Y} \tag{2.18}$$

His method is based on a few assumptions as follows:

- Constant arrival rate: the vehicles arrive at a constant rate at the stop line.
- Vertical vehicle queues: the traffic cruises at a constant speed until they join the end of the queue at the stop line. After the light turns green, the traffic will depart the intersection at a constant rate (saturation flow rate) until the queue clears.
- Operation isolation: the controls at one intersection will not affect the operation at adjacent ones. That is, the intersections operate in an isolated way.

All these assumptions were valid in Webster's era, when the traffic is moderate and no wide spread congestion occurred frequently. As the urban congestion becomes increasingly severe, these assumptions became too restrictive to reflect the real traffic situation.

Early attempts to relax these restrictions include (Miller 1963). Miller examined the release of queued vehicles using cumulative-departure curve and thus derived the delay formula for a non-constant arrival curve. Using the formula, he obtained more elaborate timing plan for fixed-cycle signals. Gazis (Gazis 1964) computed the timing plan for over-saturated intersections, and later he extended to solve the tandem over-saturated intersections(D'Ans & Gazis 1976).

Both Miller and Gazis derived the formula for fixed-timing, or pre-timed signal settings and the formulae did not rely on the assumption of constant arrival and Poisson process (2.15). However, both works and others(Newell 1956) still assume

that the arrival patterns are known a priori. In implementation, the arrival patterns are updated through either model prediction or observation in the resolution of one hour, and the time-of-day control library contained the control plans for typical peak hour, non-peak and night traffic patterns.

Traffic Responsive Signal Control

Higher level of traffic responsiveness is also pursued through more responsive control algorithms and methods. Three broad categories of responsiveness can be classified: *passive responsive control, vehicle actuated feedback control and proactive feed-forward control.*

Passive responsive control strategy refers to one category where the signal timing plan is updated when certain traffic conditions are detected. Based on typical traffic state information, various signal timing plans have been computed and stored in a library. These timing plans correspond to certain operational conditions of the intersection. The traffic monitoring system continuously collect and periodically synthesize the traffic state information and then determine the appropriate timing plan to be executed in the next period. This type of responsiveness was suggested and implemented in the Urban Traffic Control System (UTCS) program initiated in 1967 by U.S. DOT(MacGowan & Fullerton 1979-1980). The program consisted of 200 intersections with 512 detectors in a grid network in Washington D.C. Three generations of UTCS have been developed and the major characteristics of each generation are listed in Table 2.1.

One important feature of UTCS is to increase the responsiveness by reducing the updating frequency; but this strategy did not always lead to improved control performance in responding to fluctuating traffic situations(Gartner 1981)(Gartner, Stamatiadis & Tarnoff 1995). Gartner identified a few possible causes: 1) all control

Table 2.1: Characteristics of UTCS control strategies

Feature	First Generation (1-GC)	Second Generation (2-GC)	Third Generation (3-GC)
Update Interval (control period)	15 minutes	5-10 minutes	3-5 minutes (variable)
Control plan generation	Off-line optimization; selection from library by time-of-day, traffic responsive or manual mode	On-line optimization	On-line optimization
Traffic prediction	None	Historically based	Smoothed values
Critical Intersection Control (CIC)	Fine tuning (splits)	Fine tuning (splits and offsets)	N/A
Cycle length	Fixed with each section	Fixed within groups of intersections	Variable in time and space; predetermined for control period

Source: (Gartner 1981)

systems are subject to traffic measurement inaccuracies and more advanced traffic responsive control strategies are more sensitive to these inaccuracies; 2) frequent transition of timing plans comprises the effectiveness of advanced strategies; 3) phasing sequence are pre-determined. Therefore, it was recognized that reducing the control period (fifteen minutes to ten minutes and then to a variable length of three to five minutes) to increase responsiveness of control is not always effective and even harmful at times.

Vehicle actuated feedback control (VAFC) was made possible by the hardware advances during the 1930s, but more standardized controller settings were not in place until two specifications and controllers were developed in the 1970s: 1) the Model 170 controller by California Department of Transportation (Caltrans) and the New York Department of Transportation (NYDOT); 2) the TS1 by National Electrical Manufacturers Association (NEMA). Currently the more popular NEMA TS2 and Model 2070 controllers are updated version of TS1 and Model 170, respectively. Compared to NEMA TS2, Model 2070 is based on a more advanced 32-bit microprocessor with an open-structure, i.e., allowing responsible agencies to adapt different algorithms and even hardware based on a publicly available interface(Henry 2005*b*). Nevertheless, all four types of controllers share four basic circuits during every phase including:

- Initial (minimal) green time: the portion of the green split that is usually set to account for the standing queue between phase detectors and the stop line, or for the safety of the pedestrians;

- Vehicle extension, passage time or gap: the parameter extends the green interval for each vehicle actuation up to the maximum green or the force-off point.

- Maximum: the maximum length of time that a phase can receive;
- Yellow time: after the green interval to indicate the transition to next phase.

These parameters will be specified according to the intersection geometry and traffic situation (e.g., major flow direction). The VAFC is usually applied on an arterial or a network to provide coordination. Nowadays VAFC plays a dominant role in the state-of-practice in the United States (Henry 2005*a*). However, when the traffic load from all approaches is generally symmetric, vehicle actuated controller will function similarly to pre-timed controller(van Zuylen 2002).

Feed-forward responsive traffic control, or *traffic-adaptive control*, refers to the category that control variables such as green time and phases do not depend on any pre-specified parameters but adapts to traffic states flexibly. The control variables are computed with a look-ahead period and determined through comparing the outcomes from various possible timing settings. In this category, dynamic programming (DP) optimization techniques and algorithms are used extensively.

One pioneering work may be credited to DYPIC (Dynamic Programmed Intersection Control) developed at the Transport Research Laboratory (TRL) where Webster worked. More of a theoretical work, DYPIC computes the minimum delay timing trajectories over a 10-minute horizon in an offline manner, i.e., traffic arrivals assumed known ahead of time. Simulation studies (Robertson & Bretherton 1974) showed that DYPIC outperformed Webster optimum fixed timing plan and vehicle actuated controls in a two-approach, two-phase situation with random arrivals. However, the astronomical demand on storage space due to the long horizon made DYPIC hard for any practical application.

The problem was resolved by the introduction of the *rolling horizon* scheme (Gartner 1983). Similar to DYPIC, a traditional DP formulation was derived to minimize the total delay for the future horizon of approximately one cycle length (e.g.,

120 seconds). The formulation was then simplified and the rolling horizon approach is used to make the problem tractable in a real time manner. Field tests(Cohen 1989) helped identify more enhancement requirements and updated features were added, such as extra performance objective of number of stops, call-detectors to skip a phase when no vehicles at the phase are present, and adaptation to dual-ring, eight-phase actuated controller (TS2). More features were added into later developments and evaluated against real traffic (OPAC-RT Version 3.0) in New Jersey(Andrews, Elahi & Clark 1997), including dynamic speed calculations, platoon identification and modeling algorithms for possible coordination mechanisms. Gartner also envisioned combination of traffic adaptive control through better prediction of traffic arrival patterns over the network, but these efforts remained conceptual framework (Gartner et al. 1995). Nonetheless, the development of OPAC is one significant advancement, considering the computational power in the early 1980s. More recently, Fang and Elefteriadou directly extended the OPAC formulation to optimize the traffic flows of closely spaced intersections within a diamond interchange area using NEMA multi-ring structure(Fang & Elefteriadou 1006).

Another important *feed-forward responsive traffic control* algorithm was developed by the researchers from University of Arizona (Mirchandani & Head 2001). Although put in a hierarchical structure as suggested by the name RHODES (Real-time Hierarchical Optimized Distributed Effective System), the control algorithm is essentially based on the responsive control at the intersection level (Sen & Head 1997). Termed Controlled Optimization of Phases (COP), the intersection level control also applied a dynamic programming formulation to minimize the cumulative vehicle delay over the horizon.

The following notations were used in the development of COP(Sen & Head 1997):

Table 2.2: Symbols and notations for the Controlled Optimization of Phases (COP) algorithm

P:	Set of phases;
T:	Total number of discrete time steps, indexed by $t \in [1, T]$;
γ:	Minimum green time (integer number of time steps);
r:	Effective clearance interval (integer number of time steps);
j:	Index of stages in DP;
ℓ:	Index in P that denotes the initial phase;
x_j:	*Control variable* denoting the total number of time steps that have been allocated after stage j has been completed;
$X_j(s_j)$:	Set of feasible control decisions given state s_j;
$f_j(s_j, x_j)$:	Performance measure at stage j, given state s_j and control x_j;
$\upsilon_j(s_j)$:	Value function given state s_j

Given a state s_j, x_j can be selected from the following set:

$$X_j(s_j) = \begin{cases} \{0\} & \text{if } s_j - r < \gamma \\ \{0, \gamma, \gamma+1, ..., s_j-\} & \text{otherwise} \end{cases} \tag{2.19}$$

The relation between successive stages is defined as:

$$s_{j-1} = s_j - h_j(x_j) \tag{2.20}$$

where the difference $h_j(x_j)$ reads:

$$h_j(x_j) = \begin{cases} \{0\} & \text{if } x_j = 0; \\ x_j + r, & \text{otherwise} \end{cases} \tag{2.21}$$

The intersection performance is denoted in the form of:

$$f_1(s_1, x_1) \circ f_2(s_2, x_2) \ldots \circ \ldots \tag{2.22}$$

where the operator $\circ$ is simply additive when the objective is to minimize the total delay. The recursion process of the DP formulation will be as follows.

COP Forward Recursion

0. Initialize $v_0 = 0, j = 1$

1. for $s_j = r, ..., T, \{$

$v_j(s_j) = \min_{x_j}\{f_j(s_j, x_j) \circ v_j(s_{j-1}) | x_j \in X_j(s_j)\}$

Record $x^*(s_j), the optimal solution to the above problem$

2. if $(j < |P|), j \leftarrow j - 1$, and repeat from step 1.

else if $(v_{j-k}(T) = v_j(T_j))$ for all $k \leq |P| - 1$, STOP.

else $j \leftarrow j - 1$, and repeat from step 1.

The retrieval of the optimal control plan is done as follows.

0. $s^*_{j-(|P|-1)} = T$

1. for $j = \{J - (|P| - 1), J - (|P| - 2), ..., 1\}, \{$

read $x^*_j(s^*_j)$ from the forward recursion

if $(j > 1) s^*_{j-1} = s^*_j - h_j(x^*(s^*))$

$\}$

The algorithm starts with the initial intersection state and progresses iteratively in stages corresponding to pre-specified phasing sequences. Since zero-length of stage x_j is allowed, phase skipping is possible.

The traffic arrival pattern is predicted for every movement at the intersection using one set of detectors for each movement: one set at the stop line and one set 325 ft upstream. The stop line detector is to correct the errors in prediction of arrivals and estimation of vehicle queues.

As both COP and OPAC use the rolling horizon scheme to make the algorithms applicable in practice, the scheme was also criticized for it may act myopically

(Newell 1998). As Newell showed, this scheme could actually result in a situation that one movement may never be served when two competing movements are both congested. Realizing this deficiency, the researchers of RHODES improved the formulation by adding *historical delay* to the COP formulation to account for the delay that was passed down from the previous optimization horizon due to an insufficient length(Head, Mirchandani & Shelby 1997). RHODES has been tested in simulation and in field in cities like Tucson, Arizona and Temple, Florida. All tests have shown higher throughput and less delay and number of stops compared with other control systems including PASSER II-90, OPAC and existing control settings(Owen & Stallard 1999).

Besides dynamic programming, integer programming has also been used in feed-forward traffic responsive control. Representative systems include UTOPIA (Urban Traffic Optimization by Integrated Automation) by Italian researchers(Donati, Mauro, Roncolini & Vallauri 1984) and ALLONS-D(Porche 1998). Some unique modeling features of different systems have been developed, e.g., transit priority in UTOPIA and variable length of rolling horizon in ALLONS-D.

The above traffic responsive control algorithms were primarily designed for effective operation at isolated intersections, and most of them later became building blocks for arterial or network wide coordination control systems, which will be reviewed in the following section.

Arterial or Network Coordination Control

At arterial or network level, the control variable of offset, i.e., the start (or end) of the first phase at adjacent intersections in reference to the global clock time, will come into play. The control variables then include:

- phasing, i.e., how the non-conflicting traffic streams are grouped as one phase;

- phase sequencing, the order of the phase sequence;
- green split, the length of each phase, including the amber time and all-red clearance;
- cycle length, the sum of all phases in a cycle;
- offset, the difference between the start (or end) of the first phase of any intersection and the start of the network global clock.

An efficient arterial or network control plan will have to consider all the above control parameters, and the methods to incorporate the control variables are different among the various control programs. Similar to design methods for isolated intersections, these programs can also be classified into either off-line or real-time responsive controls.

Off-line Network Control Programs

Similar to Webster's method, offline network-wide control programs take the design-hour movement volumes as the traffic input, but their algorithms or techniques vary to provide signal coordination. The early offline network signal control optimization program is MAXBAND developed at IBM Laboratory(Morgan & Little 1964). Given a common cycle, phases and associated green splits and travel times between adjacent signalized intersections, MAXBAND computes maximum green bandwidth to either synchronize the signals to obtain equal bandwidth for both directions as large as possible, or produce a favorable bandwidth for one direction. The original program was solved using an enumeration method. Little then reformulated the problem as a mixed-integer linear program (MILP) to model generalized networks and allowed variable cycle lengths(Little 1966). Given a network with n signals each one of which has known red-green splits and the allowable speed

threshold between adjacent signals, the MILP is to find the relative phasing between signals, the speeds between signals and the offsets of the signals. The basic problem is formulated as follows:

One Basic MILP for the MAXBAND Algorithm

Find the bandwidth b, $w_i(\bar{w}_i)$, the time form the start (end) side of signal S_i to the green band, and integers m_i to maximize the bandwidth $b(\bar{b})$ outbound (inbound) from signal S_i to all other signals S_j that

$$\max b \tag{2.23}$$

s.t.

$$\begin{aligned} w_i + b &\leq 1 - r_i, i = 1, ..., n && (2.24)\\ \bar{w}_i + \bar{b} &\leq 1 - r_i, i = 1, ..., n && (2.25)\\ (w_i + \tilde{w}_i) - (w_{i+1} + \bar{w}_{i+1} + (t_i + \bar{t}_i &= m_i - (r_i - r_{i+1}), i = 1, ..., n-1 && (2.26)\\ b, w_i, \bar{w}_i &\geq 0 && (2.27) \end{aligned}$$

where m_i is an integer and the rest of the notations are defined as the following:

r_i:	red time of S_i on the street under study;
$t(h,i)[\bar{t}(h,i)]$:	travel time from S_h to S_i in outbound direction (travel time from S_i to S_h in inbound direction;
$\phi(h,i)[\bar{\phi}(h,i)]$:	time from center of red at S_h to the center of a particular red at S_i, where the two reds must be on the same start (end) side of the green band;

The constraint (2.26) states that for any pairs of adjacent signals, the green starts (end) relative to the green band must be synchronized. The other constraints that complete the above program include:

$$m_i = \phi(h,i) + \bar{\phi}(h,i) \tag{2.28}$$

and supplementing relations that describe any chained signals in a cycled manner:

$$\begin{aligned}
\phi(h,j) &= \phi(h,i) + \phi(i,j) && (2.29)\\
m(h,j) &= m(h,i) + m(i,j) && (2.30)\\
\phi(i,h) &= -\phi(h,i) && (2.31)\\
m(i,h) &= -m(h,i) && (2.32)\\
& && (2.33)
\end{aligned}$$

and the above relations also hold for $\bar{\phi}, \bar{m}$. Note that the same set of constraints were applied in (Gartner, Little & Gabby 1975).

The basic MILP program was extended to allow speed changes and consequently and travel times $t(h,i)[\bar{t}(h,i)]$ between signals. This is considered the first program that optimizes the traffic flows over a general network. A major revision was then made in 1981 and the updated program was finally dubbed MAXBAND(Little, Kelson & Gartner 1981). In later versions of MAXBAND, the left-turn sequences could also be optimized using the branch-and-bound searching technique.

TRANSYT(TRaffic Network StudY Tool) (Robertson 1969*b*)(Vincent, Mitchell & Robertson 1980)(Wallace et al. 1998) is another offline program to optimize green splits, cycle length and offset. Taking the traffic departing from one intersection as the cyclic flow profile(CFP), TRANSYT translates the departing CFP into the arrival pattern at the downstream intersection based on the platoon dispersion model (to be reviewed in Section 2.5.1). TRANSYT-7F, denoting the sponsor of Federal Highway Administration (FHWA) after TRANSYT being adapted in the United States, can take various combinations of efficiency measures as control objectives (Section 2.1.4). As the de facto benchmark network signal control program, TRANSYT has been used to test the effectiveness of other programs.

Real-time Network Control Programs

SCOOT

Among the real-time network control programs, SCOOT(Split, Cycle, Offset Optimization Technique) is recognized as the most widely operated online urban signal control system with its applications in over 170 cities worldwide(Shelby 2001). SCOOT takes the average of the sum of the vehicle queues as the performance index. The vehicle movement on a link essentially uses the traffic dynamics model in TRANSYT; and the cyclic flow profile is generated by the detectors positioned at just the beginning of the upstream exit link. The green split optimization is performed for individual intersections separately; and the decision is made based on whether the switch of the phase can provide better performance index at the current intersection. Another fixed-time timing plan is generated if the decision is to change the current plan, and the process is updated periodically in intervals of several seconds. Therefore, only incremental adjustment of the timing plan is allowed. Offset is optimized based on the traffic arrival from the CFP at each intersection to examine whether alteration of the current offset can provide better progression along the intermediate upstream and downstream intersections. This alteration decision is made only within the pre-specified SCOOT sub-area. The cycle length optimization is also only performed on a sub-area basis. The important criterion of cycle length adjustment is to maintain the maximum degree of saturation under 90 percent for all intersections within a sub-area. As all intersections keep the same cycle length or double cycle, these two alternatives are switchable in the case of the degree of saturation rising above 90 percent. One unique feature from SCOOT is that its later version (Bretherton, Wood & Raha 1998) applies the *gating* technique to prevent the local gridlock once the queue at the downstream link begins to spill back and block

the upstream links.

SCATS

Another famous real-time control program is SCATS (Sydney Coordinated Adaptive Traffic System) that was developed by Australian researchers(Lowrie 1981). SCATS also takes a hierarchical operational structure, with central, regiona and local computers controlling over 1,000 signal controllers in Sydney. Each controller can perform either vehicle actuated control or time-of-day control, and these controllers consist of the *local* level of the hierarchy. The traffic responsiveness of SCATS comes from a few unique features. The *regional* computer maintains up to 120 controllers, where they are grouped into sub-systems of one to ten controllers in each. SCATS is able to regroup or decouple the adjacent subsystems of similar/disimiar cycle lengths, using a vote-casting mechanism. The common cycle length of each subsystem is updated once per cycle with changes of no larger than six seconds. If the difference of the cycle lengths of two adjacent subsystems is within nine seconds, a positive vote is cast. The two subsystems will merge into one if four or more cumulative votes are tallied. If the tallied system has zero or less votes, the system will be de-coupled. The traffic is monitored by only presence detectors, and thus no cyclic flow will be recorded to provide the data for offset adjustment. Rather SCATS uses the ratio of the effectively used green to the available phase green time, abbreviated as DS as the basis for split adjustments. Four pre-specified greensplit are stored for each sub-system, and a vote-casting process is again applied once every cycle to vote for the plan that minimizes the maximum DS within the sub-system. Four consecutive votes are considered enough for the adoption of the plan. Offset options are provided for both within the sub-system and externally between sub-systems. Depending on the time of the day and the traffic situation (balanced or uneven traffic

on both directions of the strategic links), a separate offset plan is tailored to adapt to the situation. Within the coordination constraints and cycle length constraints, the actuation duration is possible for each phase, including phase-skipping or early termination.

Other online network level control programs that applied similar cycle length, green split and offset optimization techniques include MITROP(Gartner, Little & Gabby 1976), PASSER(Messer, Whitson, Dudek & Romano 1973)(Chaudhary & Messer 1993).

Hierarchical Structure in Traffic Responsive Network Control

Network control of traffic responsive control systems with local signal controllers, e.g., OPAC, RHODES or UTOPIA/SPOT, need only come up with coordination modules because essentially these systems do not explicitly maintain the variables of cycle length. In this sense, a hierarchical structure of a sequential optimization of individual intersections and then their coordinations, or vice versa, is generally applied.

In the RHODES system, the network level coordination is determined by the REALBAND module (Olmo & Mirchandani 1992). REALBAND propagated the projected platoons based on an approximate flow model. If two platoons arrived at an intersection simultaneously on conflicting phases, a decision node will be formed to determine which platoon receives the right-of-way. If the platoons arrive at slightly offset, the first arrival may be split to give the other one priority. In such a way, a decision tree will be formed and solved via branch-and-bound algorithm. At the lower level of intersection control, the offset obtained from the optimal roof-to-leaf decision tree will then be the mandated time window for the phases. The Controlled Optimization of Phases (COP) algorithm will use these as the constraints

for successive phase duration optimization.

UTOPIA/SPOT breaks its area level control into *observer* and *controller* components. The detectors are positioned at upstream of each controlled link to track the counts and occupancy for both the departing flows and the arrival flows. The area traffic is monitored in this way and the major routes are predicted on a 3-minute interval basis. The controller optimizes the flow in a rolling horizon scheme on a 30-minute interval basis. The decision variables are the average speed of the major routes over the network and the saturation flow rates constrained by feasible boundaries. The literature said that "suitable" reference rules or reference plans will then be specified.

PRODYN also takes a hierarchical structure in its control program(Henry, Farges & Tuffal 1983). The lower level consisted of all optimization process at each intersection, while the upper level took care of the interactions between all intersections. At the upper level, the program simulated the network flow and generated the departure profile and the sensitivity profile for each controlled link. The lower level would then translate the departure profile into arrival profile for each approach of signal controllers. The sensitivity profile, obtained from the dual variables of the upper level formulation, provided the possible effect from the variation of the arrival profile on its performance on a 5-second interval over a 75-second horizon. In such a way, the offset information can then be calculated and received at the lower level of intersection actions.

Designing OPAC mainly for control at the intersection level, Gartner envisioned that the it could also be the building block for network wide adaptive coordination control. However, the continuing literature has hardly revealed how this could proceed. Nevertheless, later versions of OPAC were enhanced with new features that include adaptation to the traditional dual-ring, eight-phase controller structure

and "virtual fixed cycle (VFC)" concept. These features enabled the application of general offset optimization modules and thus could provide coordination control.

In summary, arterial or area traffic control programs generally take hierarchical structure to optimize the control variables. Off-line programs such as TRANSYT starts computation of optimal offsets at the local intersection level and then adjust the offset between adjacent intersections to provide the progression. This is a bottom-up structure. This structure is also shared by SCOOT and SCATS systems with some variations. Fully responsive programs including RHODES, UTOPIA/SPOT and PRODYN take the top-down structure. By predicting the arrival flows over the network, the programs will first compute the offset and use this information as the constraints for the control optimization at the intersection level. Potentially being able to provide better traffic progression throughout the overall network, the latter structure indeed proved successful in the laboratory simulation. However, this structure was discarded in the real applications such as RHODES because of its high computation burden(Mirchandani & Head 2001).

2.2.2 Ramp Metering

Ramp metering is used to improve the freeway traffic flows. It determines the flow rates of the vehicles that can enter the freeway from on-ramps[1]. The control is enforced by a traffic light installed at the end of the on-ramp and the desired metering rate is imposed by selecting appropriate green and red timings. depending on the driving behavior in the field, the metering rates, usually denoted as number of vehicles per hour, can be set up flexibly according to the local traffic situation.

[1]We also need to note that ramp metering does not necessarily need to operate at *traffic restrictive mode*, i.e., the metering rate is lower then the ramp demand. It can operate at a *traffic spreading mode*(Hegyi 2004), i.e., breaking the vehicle platoons into individual vehicle to avoid creating the disturbances and shockwaves on the freeway mainline. But in general, ramp metering operates at restrictive mode during the peak period to maintain the high flow on the freeway.

For example, 600 vehicles per hour can be enforced by two ways: 1) a 6-second cycle (1 second of green and 5 seconds of red time) and one vehicle per green second or, 2) a 12-second cycle and two vehicles per 3-second green time.

Similar to the intersection control, the ramp metering methods and algorithms can also be functionally classified as pre-timed or traffic responsive ones, and isolated or coordinated ones. Therefore, any ramp metering algorithm can be one of the four type (Zhang 2001):

- Isolated pre-timed;
- Isolated traffic responsive, e.g., ALINEA(Hadj-Salem, Blosseville & Papageorgiou 1991);
- Coordinated pre-timed;
- Coordinated traffic responsive, e.g., SWARM(Paesani, Perovich & Khosravi 1997), BOTTLENECK(Jacobsen, Henry & Mehyar 1989).

Traffic responsive ramp metering methods received the most research interest. The best known responsive metering algorithm is perhaps ALINEA(Hadj-Salem et al. 1991). ALINEA is a simple feedback control law and formulated as:

$$r(t) = r(t-1) + K_R[O_c - O_{out}(t)] \tag{2.34}$$

where $r(t)$ is the metering rate for the current time step and $r(t-1)$ is the metering rate in the previous time step, O_c is the target occupancy to be maintained which is usually slightly lower than the critical density (corresponding to the capacity flow), and $O_{out}(t)$ is the current occupancy. K_R is the only parameter to be adjusted in implementation. Field experiments shows the resultant metering rates are not sensitive to the choice of K_R(Papageorgiou 2000). Both field experiments(Papageorgiou

2000)(Haj-Salem & Papageorgiou 1995) and microscopic simulation studies (Zhang 2001)(Hasan, Jha & Ben-Akiva 2002) reported that ALINEA control can reduce total travel time significantly. Later variations of ALINEA were proposed including flow-based control or ramp queue managements(Smaragdis & Papageorgiou 2003).

Similar to the hierarchical optimization in network signal control, some ramp metering algorithms also take multi-level structure to coordinate the ramp meters. Developed for Interstate 5 north of the CBD area of Seattle, WA, BOTTLENECK used a more simple heuristic approach to coordinate the ramp meters. A two-level structure is taken in BOTTLENECK. At the global level, the method first identifies bottlenecks, takes the difference between the total upstream demand and the downstream capacity as the volume reduction, and then distributes the volume reductions to the ramp meters based on their pre-determined weights. At the local level, the same demand-capacity analysis is conducted for the intermediate freeway section to determine the metering rate for each associated ramp meter. The more restrictive metering rates will be selected. BOTTLENECK is theoretically simple and has less model parameters to adjust. A similar two-level treatment can also be found in SWARM(Paesani et al. 1997), whereby more features, including handling queue spill-back accommodating faulty detectors, as well as predicting congestion evolution, were used to increase the responsiveness and robustness of the algorithm.

Coordinated metering strategies using more sophisticated mathematical programs or "soft computing" techniques have also been explored both in the field and in the laboratory. The local feedback control of ALINEA was extended for coordinated control in (Papageorgiou, Blosseville & Hadi-Salem 1990*a*)(Papageorgiou, Blosseville & Hadi-Salem 1990*b*), using a second-order flow model to describe the traffic evolution on the expressway. The coordinated ramp metering algorithm takes

the form of the classical linear-quadratic integral law:

$$\bar{r}(k) = \bar{r}(k-1) - K_{LQI}^{-1}[\bar{\rho}(k) - \bar{\rho}(k-1)] - K_{LQI}^{2}[\hat{\bar{\rho}}(k) - \hat{\rho}_d] \tag{2.35}$$

where $\bar{r}$ is the vector of controllable ramp volumes, $\bar{\rho}$ is the vector of densities, $\hat{\bar{\rho}}$ is the vector of some selected bottleneck densities, and K_{LQI}^{1}, K_{LQI}^{2} are the gain matrices that denote the desired traffic conditions that the coordinated ramp metering is aimed at. This control algorithm was put into both simulation and practice on the southern part of Boulevard Peripherique in Paris, France and showed significant improvement in total travel time reduction, compared to the cases of no control and local feedback control.

Stephanedes and Chang(Stephanedes & Chang 1993) applied a "center-space" scheme to approximate the kinematic wave model of Lighthill, Whitham (Lighthill & Whitham 1955) and Richards (Richards 1956) (LWR), and solved the formulated nonlinear program by the conjugate gradient algorithm to obtain a set of time-dependent ramp metering rates for a stretch of freeway.

Other representative coordinated ramp metering algorithms include a Fuzzy logic algorithm used in Seattle and the Netherlands(Meldrum & Taylor 1995) and the artificial neural network (ANN) algorithm(Zhang & Richie 1997). These types of metering strategies are more flexible to take complex traffic flow models and capture the nonlinear relationship between dynamic evolution of the traffic and metering strategies. On the other hand, these complex systems need more attention than the above simple rule-based algorithms, e.g., more carefully designed rules (e.g., in Fuzzy logic) or longer processes of reaching acceptable model parameters for further controls(e.g., ANN). Mathematical program based coordinated ramp metering includes either time-of-day control (Yoshino, Sasaki & Haegawa 1995) or real-time control(Gomez & Horowitz 2004*a*), where efficient linear programming (LP) solvers are readily available.

2.3 Integrated Freeway-Arterial Corridor Control Systems

Individual control systems of urban signal control (Section 2.2.1) or ramp metering (Section 2.2) are aimed at the efficiency of the concerned subnetwork within the corridor, but they may not be able to improve the overall efficiency of the entire corridor. In this section, we shall look at the past integrated control studies to have a better understanding of the traffic operations within a general corridor network.

A transportation corridor is operationally rather than geographically or organizationally defined as "a combination of discrete parallel surface transportation networks (e.g., freeway, arterial, transit networks) that link the same major origins and destinations"(FHWA 2005). A corridor usually includes various types of facilities (e.g., freeway sections, ramps and urban streets), which are typically managed by different agencies and jurisdictions. As revealed from the above review, the subnetworks of freeway system or urban arterial system within most corridors are operated separately with little consideration to their coordination. Gradually it becomes more clear that integrating the control measures can improve the operational performance of the entire corridor(van Zuylen & Taale 2003). Even though receiving less attention than signal control or ramp metering, a variety of integrated freeway corridor control studies have been performed in the last two decades. These studies can be classified into two classes: optimization based, time-dependent, and feed-forward system-wide control and locally traffic-responsive feedback control.

2.3.1 Local Integration and Rule-based Control Methods

This class of integrated control makes use of the local traffic information and performs traffic control in a distributed manner. In the early 1990s a study

sponsored by FHWA identified four types of operational strategies to integrate arterial control and ramp control(Pooran & Lieu 1994). Based on the severity of the corridor congestion, the strategies can be:

(1) local coordinated strategy to locally adjust the ramp metering rate and/or signal timing parameters when the freeway demands do not require integrated control,

(2) area wide integrated control to optimize the corridor flow in a responsive manner by frequently adjusting both signal timing and ramp metering to cope with short term freeway flow fluctuations,

(3) diversion strategy to handle incidents on the freeway,

(4) network wide coordination to manage the spread of the congestion when the traffic demand exceeds the capacity within a critical sub-area of a corridor.

For each strategy, a list of sixteen operational tactics have also been specified, which can be selected to deal with possible congestion scenarios. For example, One tactic is to discharge freeway on or off ramp traffic at severe conditions. A simulation study network of Interstate 5 through Seattle CBD area were selected to investigate various combinations of tactics and the results showed operational improvements at both freeway and arterial. However, because the simulation software could not model the close interaction between signal timing and ramp metering, the full potential of the proposed integration strategies were not evaluated in the simulation study.

A more specific integration plan was designed in (Tian, Balke, Engelbrecht & Rilett 2002). This scheme considered both the on-ramp meter and the adjacent traffic signals, setting the ramp metering rates to one of two pre-selected values to deal with various levels of congestion on these ramp segments. A VISSIM simulation study showed that this simple scheme is quite effective in reducing the congestion on the freeway.

A similar local coordination structure was adopted recent in field in Minnesota (Kwon, Ambadipudi & Bieniek 2005), which can shed light on the essence of the local integration control schemes based on predefined rules. Their ramp metering rates are updated with a feedback control as follows:

$$R^{t+1} = R^t \times \alpha^t \tag{2.36}$$

where R^{t+1}, R^t are the metering rates for successive control intervals $t+1$ and t. The value of the adaptive parameter α^t is determined by a set of rules:

$$\alpha^t = 1 + (1 - C_F^t/C_{T,F}^t)/(R_{max}/R^t - 1) \quad \text{if} \quad C_{T,F}^t - C_F^t > 0 \tag{2.37}$$

$$\alpha^t = 1.0 \quad \text{if} \quad C_{T,F}^t - C_F^t = 0 \tag{2.38}$$

$$\alpha^t = 1.0 - (C_{T,F}^t - C_F^t)/(1 - R_{min}/R^t/(C_F^t - 1) \quad \text{if} \quad C_{T,F}^t - C_F^t < 0 \tag{2.39}$$

where $C_F^t, C_{T,F}^t$ denotes the measured and target *congestion index* of the freeway segments under control. The congestion index (CI) is defined as:

$$C_{j,k} = \beta_j(P_{j,k} + V_{j,k})/(1 + V_{j,k}) \tag{2.40}$$

where the notations are:

$C_{j,k}$, CI for detector j in a link at the end of detection time interval k;
$V_{j,k}$, number of vehicles passed detector j during interval k;
$P_{j,k}$, 1.0 if detector j is occupied at the end of the time interval;
β_j, weight parameter for detector j and $\sum \beta_j = 0$

and the link level CI is a combination of all detectors on the same link over the control interval:

$$C_{i,k} = \sum_j \beta_j(P_{j,k} + V_{j,k})/(1 + V_{j,k}) \tag{2.41}$$

The signal timing parameter changes were also incorporated into the above system by the congestion index comparison. Using congestion index allows for capturing the traffic states at both freeway and intersection level and adjusting the ramp metering and signal timing coherently.

An even more myopic adjustment of the ramp metering within an interchange area can be found in (Han & Reiss 1994). The study observed that ramp demand from the cyclic traffic surges from an adjacent signalized intersection could lead to temporary delay on the ramp at sometime while insufficient use in the remainder. Even though ramp metering could be enforced, the traffic tides at the ramp lead to inefficient use of the ramp capacity. A linear program is then formulated to optimally split the ramp metering period into shorter intervals according to the cycle of the signal and calculate the metering rates for each of these interval.

While the local feedback control strategies mentioned above are less rigorously formulated than the system-wide control strategies, and may not steer the system to its most efficient operating state, they are nevertheless quite appealing because

- they are distributed and less data-intensive,
- they may be more robust with respect to component failures, and
- they require less resources to implement and operate.

2.3.2 Integrated Corridor Control Programs

In this category, the integrated control problems are often formulated as optimal control or mathematical programming problems, with traffic flow being modeled as kinematic waves or high-order continuum "fluids". Traffic control measures that have been studied include ramp metering, intersection signal control, and route diversions, or combination of them. These problems were routinely solved by linear or constrained nonlinear programming techniques.

Papageorgiou(Papageorgiou 1995) formulated the integrated control problem as a linear program using a *store-and-forward* approach(Gazis 1974). Optimal

signal timing plan and optimal route diversion routes can be computed efficiently with specially designed linear programming solvers(Banos & Papageorgiou 1995). With the same framework, an integrated traffic-responsive urban corridor control (IN-TUC) system has been developed and field tested in Glasgow in UK(Diakaki & Papageorgiou 1997)(Diakaki, Papageorgiou & McLean 2000). ALINEA is used for the ramp metering within the corridor in IN-TUC, whereas the objective function for the urban signal part takes a quadratic form:

$$\mathfrak{J} = \frac{1}{2}\sum_{k=0}^{\infty}\left(||\mathbf{x}(k)||_Q^2 + \mathbf{\Delta g}(k)||_R^2\right) \tag{2.42}$$

where $\mathbf{x}$ is the vector of number of vehicles x_j (c.f. Table 2.5.1), and $\mathbf{Q}$ and $\mathbf{R}$ are diagonal weighting matrices. The vector of green time increment $\Delta\mathbf{g}(k)$ is calculated as follows.

Refer to the store-and-forward formulations in (2.48, 2.49 and 2.53), substitute the latter three into (2.48) and assume there exists a nominal state $x_j^n = 0$, there will be a steady-state version of (2.48) that reads:

$$0 = T\left[(1-t_{j,0})\sum_i t_{i,j}\frac{S_w\sum_i g_{M,i}^n}{C} + d_j - \frac{S_j\sum_i g_{N,i}^n}{C}\right] \tag{2.43}$$

Subtracting (2.43) from (2.48) result in the following state equation:

$$x_j(k+1) = x_j(k) + T\left[(1-t_{j,0})\sum_{w\in I_M} t_{w,j}\frac{S_w\sum_i \Delta g_{M,i}^n}{C} + d_j - \frac{S_j\sum_i \Delta g_{N,i}^n}{C}\right] \tag{2.44}$$

where $\Delta g = g - g^n$ and $\Delta\mathbf{g}$ is the vector of Δg. In addition to the feasibility constraint (2.51), The following common cycle constraint

$$\sum_j(\sum i, m g_{i,j,m}(k) + L_j) = C \tag{2.45}$$

completes the formulation. The simulation and field evaluation of IN-TUC showed great improvements at the specific freeway sites and of the overall system. As will

be reviewed in Section 2.5.1, The store-and-forward approach requires that the flow updating interval must be no less than the signal cycle lengths. This feature rules out the possibility of modeling offset, making the model more suitable as a dynamic capacity allocation module.

Based on the second-order flow model by Payne(Messmer & Papageorgiou 1990), an optimal control formulation of integrating ramp metering and variable message signs (VMS) has been proposed(Kotsialos & Papageorgiou 1999) and solved using different algorithms including feasible-direction algorithm (Kotsialos, Papageorgiou, Mangeals & Haj-Salem 2002) and conjugate gradient algorithm(Kotsialos & Papageorgiou 2004).

The above system-wide, optimal control based corridor control strategies can be quite effective in reducing traffic congestion. On the other hand, they often present serious challenges to implementation. Firstly, they require extensive data inputs, such as time-dependent origin-destination data, traffic measurements (flow, density, etc) at sufficient number of locations, and driver behavior parameters (e.g. diversion propensity). Secondly, they need a significant support infrastructure such as a central traffic management center (TMC), high performance computing devices, a reliable detection and communication system, and supportive policies and the availability of experienced operating personnel (McLean et al 1998). Compared to the rule-based local coordination control methods(2.3.1), their centralized structure makes them less robust in the presence of component failures, which are very likely to occur in a large control system with hundreds or thousands of field elements.

2.4 Integration of Traffic Control and Traffic Assignment

Both traffic control and traffic assignment are well-researched fields, and it is interesting to notice that the study of one is based on the assumption of the other as fixed input. When traffic control is concerned, it is assumed that the travel demand and consequently the flow pattern is relatively stable and known. On the other hand, when traffic assignment is concerned, the control policies and plan is generally assumed fixed and unchanged. This is particularly true in the practice of travel demand forecast, where static traffic assignment prevails. However, these two processes interact with and respond to each other: when the signals on one street are better coordinated, the travel time using this path will be reduced and more travelers will switch to this street. Consequently, the changed traffic demand upon the route will cause the signal control plan to be obsolete and call for updating of the controls. Such interaction is best captured through integrating traffic assignment and traffic control.

In this work, our objective is to come up with integrated control plan for a transportation corridor; and traffic assignment, the process of loading the travel demand onto the network according to certain users' route choice behavior, will have to be adequately modeled. Therefore, it is necessary to examine how the integrated traffic control and traffic assignment problem has been attacked in the past studies.

2.4.1 Static Traffic Control-Traffic Assignment Studies

Allsop(Allsop 1974) was perhaps the first to formally state the interaction between signal control and traffic equilibrium assignment. He used TRANSYT to optimize a simple six-signal urban network while performing equilibrium traffic as-

signment trivially on the network. He commented that the link travel cost is directly dependent on the flows and signal settings rather than the rough estimation from link performance functions. The problem originated the equilibrium network design problem (ENDP) research (Dickson 1981) that aims to allocate the right-of-way to conflicting movements based on the assumed user equilibrium flow pattern formed collectively by the travelers. The solution to the problem would satisfy:

$$\min_{\lambda} \sum f_a \cdot t_a(f_a^*, \lambda) \tag{2.46}$$

where a is the link index, f_a is the link flow, t_a is the link travel cost that depends on both the flow and the control setting λ. The optimal control setting would correspond to the equilibrium flow pattern f_a^*, which will be solved from the equilibrium traffic assignment(Sheffi 1985).

Various formulations under different assumptions were investigated, and perhaps the most commonly seen will be bi-level programming that was extended from the game theory framework. Fisk (Fisk 1984) showed that equivalence between Wardropian's first principle (Wardrop 1952) and a Nash equilibrium among non-cooperative travelers(Nash 1951). She also formulated the ENDP problem as a leader-follower game, *or* a Stackelberg game, where the system administrator is one player (leader) and the travelers collectively are the other player (follower). While the leader gains some knowledge about how the follower will play, the follower does not have the power to affect the leader's decision, i.e., choosing the signal settings. Exactly in this framework, LeBlac (LeBlanc & Boyce 1986) introduced the bi-level programming structure to solve the ENDP problem. The problem is split into two levels of optimization: at the upper level, the system cost is minimized by changing the green time, subject to the flow coming from the lower level optimization. At the lower level, both deterministic user equilibrium (Yang & Yagar 1995) and stochastic equilibrium (Maher, Zhang & van Vliet 2001) have been modeled to represent the

users route choice behavior.

Two issues were found critical to obtaining a reliable solution of the static ENDP problem: the travel cost assumptions and solution methods. In(van Vuren & van Vliet 1992), the researchers investigated the solutions to the static ENDP problem using both the BPR link performance function(BPR 1964) and Webster's control delay function(Webster 1958), and tested the solutions with both simple contrived networks and real networks in UK. They found that the solutions were far from each other under various cost functions. In lieu of this limitation of rough estimation of travel cost, they developed a new model SATURN (van Vliet 1982)(Hall, van Vliet & Willumsen 1980) to use the cyclic flow profile (CFP) at signalized intersections to get a finer estimation of the delay at controlled nodes.

The second issue is related to solution algorithms. Allsop(Allsop 1974) suggested an iterative procedure to solve the traffic assignment and control optimization alternatively until two processes converge to a stable point, the so-called *mutually consistent point (MCP)*. This procedure was also used in (van Vuren 1990) and the solution methods in bi-level programming also generally fall in this category (Yang & Yagar 1995)(Maher et al. 2001). However, researchers (Tan, Gershwin & Athans 1979) have criticized its ability of only converging to local optima if convergence at all. By observing that the UE flow can be expressed as a set of constraints, they formulated the problem as a hybrid optimization problem (HOP) and developed an Augmented-Lagrangian algorithm to solve the converted unconstrained optimization problem. However, solving the re-formulated HOP would require enumeration of all paths connecting each O-D pair, which had been known as an N-P hard problem for its notorious increase of computation burden associated with the problem scale. Al-Malik(Al-Malik 1991) in his thesis proposed a method to modify the Wardropian principles to attack better MCP. His method divided the trip demand into "fixed"

part such as the external-external demand and the variable part. Only the variable part would participate in the control-route choice interaction. With two very simple networks, he showed that only one single MCP was obtained. Seemingly nice when dealing with *ad hoc* simple networks, the method did not prove applicable for its application in general ones. Meanwhile no specific guidelines were provided to determine the split of the demand into both the fixed and variable part.

In general, the studies of static ENDP suffer from the following deficiencies: 1) because of the formulation of the problem, the links generally interact with one another and the non-separability of the cost function renders the problem an ill-conditioning one. No unique solutions could be guaranteed unless strong restrictions or assumptions be made on the cost function; 2) When strong assumptions on travel cost functions and network structure to make the problem tractable, the solution would lose part of its modeling power; 3) The travel cost function calculated in traffic assignment is usually inconsistent with the control optimization. Nevertheless, the ENDP formulation has provided valuable insight into formulating and solving the integrated corridor control problem.

2.4.2 Dynamic Traffic Control-Dynamic Traffic Assignment Studies

The limitations of static network modeling that traffic flow patterns were constructed without reference to its variation over time has led researchers to model the temporal traffic evolution, initiated by the work in (Merchant & Nemhauser 1978*a*)(Merchant & Nemhauser 1978*b*).

In terms of dynamic control-assignment problem, Chang *et al*(Chang 1993) provided a discrete time optimal control modeling framework to integrate the measures of urban freeway ramp metering and surface street signal control. The framework uses a rolling horizon scheme to monitor and correct the traffic state prediction

from reactive traffic assignment, where the route diversion from freeway ramps is modeled via a logit model based solely on the local trip delay at the diversion point. The link performance function is simplified by a piecewise linear function to ease the computational burden so that the system-optimal control problem is solved through available linear programming solvers.

Chen's thesis (Chen 1998) may be the most complete work on dynamic ENDP problem. Based on different assumptions of the interaction between system administrators and network travelers, he formulated the dynamic traffic control-assignment problem into either Cournot, Nash or Stackelberg games. Within the sub-problem of dynamic traffic assignment, he introduced multiple user classes including habitual drivers, unguided drivers, imperfectly guided drivers and perfectly guided drivers, each following different route choice rules. A C-logit model (Cascetta, Nuzzolo, Russo & Vitetta 1996) was applied to model the route choice behavior. He proposed both an off-line and an online solution frameworks to update all four control variables including green split, phasing, cycle length and offset, but only the green split was modeled in more detail and solved. A whole link model (Chabini & He 1998) was applied in his *dynamic network loading* module, where the traffic evolution on the link is only determined by the entry and exit flows at certain time. Simplified Webster formula was used to compute the link travel time because of its differentiability and consistency between both assignment and control.

The dynamic traffic control-assignment problem advances the integrated control study in several aspects: 1) more realistic network traffic evolution is possible than static modeling; 2) network coordination can be taken into account; 3) the control plan could be computed in a real time manner while taking the users' reaction into account. However, the above studies all depend on link performance functions to compute the link travel cost under various traffic loads and thus perform

the dynamic network loading (DNL) accordingly. These functions were developed mainly from on experimental observations and statistical assumptions that capture the average causal effect between vehicular volumes and travel delay. As far as real time traffic management is concerned, this type of delay-volume relation does not hold its realisticity, e.g., the unlimited delay if the volume exceeds the capacity. This deficiency significantly degrades the credibility of the solution obtained from using this type of cost function. This problem calls for more elaborate and more robust modeling of the traffic dynamics through a general corridor network, which will be discussed in the next section.

2.5 Traffic Flow Dynamics and Optimization Algorithms

After the review in the above two sections, we recognize that two components are fundamental to modeling an integrated corridor control system as by other researchers, e.g., (Stephanedes & Chang 1993). The first is the traffic flow model that realistically represents traffic evolution, and the other is the optimization method to generate optimal control plans.

2.5.1 Traffic Flow Models

Three major categories of traffic flow models have been developed and applied in traffic control studies: the point-queue (P-Q) or vertical queue model, the spatial queue (S-Q) or horizontal queue model and the Lighthill-Whitham-Richards (LWR) model.

Vertical Queue

Vertical queue or point queue model assumes that the vehicles travel at the design speed uniformly on the road section and arrive at the stop line at a constant rate. The vehicles behind the stop line take no physical space and will be discharged at the saturation flow rate during the effective green time. Some variations to vertical queuing have been proposed in various studies.

Relaxing the constant travel speed on the link, the platoon dispersion model in TRANSYT (Robertson 1969*b*) uses an empirical formula termed as the cyclic flow profiles (CFP) to describe the vehicle movements on roads. The arrival profile from the upstream departure flow q_k will be dispersed as the following profile:

$$q^l_{k+t} = f q_k p + (1 - f) q_{k+t-1} \tag{2.47}$$

where $f = \frac{1}{1+0.35t}$ is the smoothing factor and t is 0.8 times the mean cruise time (measured in steps) over the distance for which the dispersion is calculated. However, it can easily be seen that the dispersion model does not hold the flow conservation; that is, some of the flows will never arrive at the downstream stop line along the dispersion process. Despite the relaxation of constant travel speed on the link, platoon dispersion model still lets vehicles queued at the stop line. So essentially the traffic flow model in early versions of TRANSYT is still considered vertical queue.

Most studies and control programs used vertical queue model to model the traffic flow dynamics. These studies include the classical study by Webster and later HCM methods, various derivation programs of TRANSYT such as SCOOT and SCATS; OPAC and earlier versions of RHODES.

Horizontal Queue

Extending vertical queuing assumptions, horizontal queue or spatial queue (S-Q) model assumes that vehicles occupy (uniform) physical spaces and new arrivals can only join the end of the stopped queue. Different from vertical queuing, S-Q model could replicate the spillback of traffic from occupying one road section to another. This feature poses the *non-separability* characteristic to link interactions and thus more difficulties arise in analytical models.

It is interesting to examine the difference between S-Q model and the classical *store-and-forward* modeling approach (Gazis 1974). This approach was early seen in (Gazis 1964)(Gazis 1974) and adapted in recent studies in Europe(Papageorgiou 1995)(Diakaki & Papageorgiou 1997)(Diakaki et al. 2000). We herein review one of them to examine the difference. With the following notation

Table 2.3: Notations of Store-and-Forward Flow Dynamics

j:	link index
i:	node (intersection) index
k:	time step index
τ:	time step duration
$x_j(k)$:	queue size at the end of the link j at time k
$q_j(k)$:	incoming flow rate into link j at time k
$u_j(k)$:	outgoing flow rate from link j at time k
κ_j:	free flow travel time of link j, must be in multiples of T
$d_j(k)$	demand generated within link j at time k
$t_{j,0}(k)$	exit flow ratio within link j at time k
$g_{i,j,m}$:	green split of phase m of link j at intersection i
ν_i:	Number of phases at node i
C_i:	cycle length at intersection i, assumed to be known and C as the common cycle
s_j:	saturation flow rate on link j
M, N:	the tail and head node of link j
w:	$w \in I_M$
I_M:	set of links entering junction M towards link j

we will have the following equations to describe the model. The link conservation equation on an urban street section is:

$$x_j(k+1) = x_j(k) + \tau[(1 - t_{j,0}(k))q_j(k - \kappa_j) + d_j(k) - u_j(k)] \tag{2.48}$$

The inflow to link j will then be:

$$q_j = \sum it_{i,j}u_i(k) \tag{2.49}$$

The freeway link flow conservation equation reads:

$$x_j(k+1) = x_j(k) + \tau[q_j(k - \kappa_j) - u_j(k)] \tag{2.50}$$

The storage capacity of any road section is enforced as:

$$0 \leq x_j(k+1) \leq x_{max} \tag{2.51}$$

The flow updating will then read:

$$u_j(k) = \frac{g_j(k)}{C} s_j \tag{2.52}$$

where $g_j(k)$ is the sum of the green split allocated to link j as

$$g_j(k) = \sum g_{i,j,m}(k) \tag{2.53}$$

Store-and-forward approach implicitly requires that the flow updating interval, indexed by k, should be no shorter than the cycle length of any signal group, as can be seen from equation(2.52). Therefore, the traffic is always "flowing" through the intersections and the feature of intermittent servicing of the signals cannot be modeled adequately. Because of this characteristic, store-and-forward approach cannot model offset and is thus more suitable for network wide capacity allocation(Papageorgiou 1995). Because S-Q does not impose such a restriction,

S-Q can be used for the actual network control studies. In fact, S-Q model is seeing more and more applications in many control programs such as TRANSYT-8 (Vincent et al. 1980), RHODES(Head et al. 1997), (Fang & Elefteriadou 1006). One interesting study (Helbing 2003) firstly applied the LWR model at link level using the triangular fundamental diagram to derive the queuing delay formula and travel time with relaxations; but the relaxation essentially degraded LWR model back to a spatial spatial queue model again.

Both the P-Q and S-Q models can provide good estimates of the queue size, i.e., the number of stopped vehicles, under low to medium traffic loads in particular. But when the traffic load is high and the intersection is near or over saturation, the traffic densities behind the stop line will be in frequent transitions because of the varying arrival rates and intermittent signal services (Stephanopoulos, Michalopoulos & Stephanopoulos 1979). Shock and acceleration waves, interfaces between two differing traffic states, will be generated in such a complicated way that neither the P-Q nor the S-Q model could capture the spatial extent of queue formations and dissipations. Consequently, queue lengths cannot be estimated accurately. Queue length can be stated as "the length of the roadway section behind the stop line where traffic conditions are in the congested region of the flow-density curve, i.e., they range from capacity to jammed." (Stephanopoulos et al. 1979). As they further showed that under these circumstances the control action would be dictated by minimizing queue length instead of delay, more sophisticated models to capture the queuing dynamics more accurately are needed.

LWR Model

In the theory of "kinematic waves", the Lighthill-Whitham-Richards (LWR) continuum flow model is well accepted. the LWR model states the following two

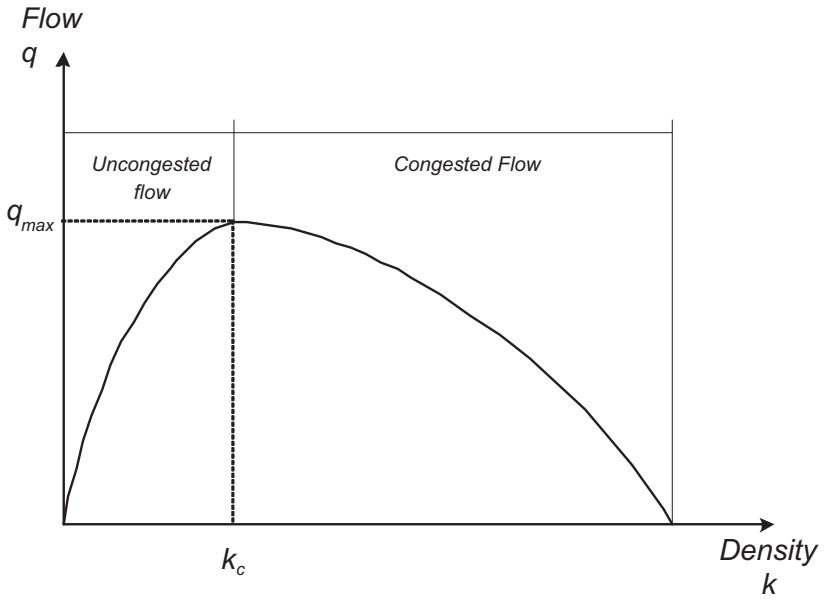


Figure 2.1: The Fundamental Diagram (Flow-Density Relation)

equations to model the traffic dynamics on a road section:

$$\frac{\partial q}{\partial x} + \frac{\partial \rho}{\partial t} = 0 \text{ and } q = f(x, \rho, t) \tag{2.54}$$

where q, ρ, x, t denote flow rate, density, distance and time, respectively. The first equation states the flow conservation on the section; the second states the relation between flow rate and density along the road section and the curve of this relation is usually called *fundamental diagram*, shown in Figure 2.1. Various assumptions and theories have been proposed for certain forms of the $q - \rho$ relation and thus the LWR model can have different solutions as well. Because of its capability to depict the formation and dissipation of shockwaves, LWR model has also been introduced into

control studies.

In (Stephanopoulos et al. 1979) the authors incorporated the linear speed-density relation (Greenshields 1934) and the resulting parabola fundamental diagram (flow-density curve) in the signal control design. A very complicated system is consequently built for isolated intersection (Michalopoulos, Stephanopoulos & Stephanopoulos 1981) but the system could not possibly be extended to network control because of its analytical nature.

Recently researchers begin to make use of a finite difference solution scheme to the LWR model, the so-called cell transmission model (CTM) (Daganzo 1994)(Daganzo 1995), to perform traffic control studies. Firstly the scheme was devised for freeway links, and then it was extended to network flow dynamics (Daganzo 1995) by decomposing complex network topologies into simple merges and diverges and specifying the flow updating rules at these merges and diverges. Lo adapted the scheme to model the flow dynamics at signalized intersections(Lo 1999). He then modified the scheme into a linear program and his modification technique is similar to a many-to-one dynamic system optimal assignment formulation (Ziliaskopoulos 2000).

Another type of linear transformation of the CTM scheme has also been used to study the global optimal ramp metering strategies(Gomez & Horowitz 2004*a*)(Gomez & Horowitz 2004*b*). In this study, influence parameters were added to the modeling of ramp merging flows instead of the original rule-based CTM scheme at merges. A linear program can then be rendered equivalent to the original non-linear program arising from the concave function (c.f. Equation 3.1). However, the underlying assumptions of no queue spillback from mainline to affect the onramp flows were too restrictive under certain circumstances, which compromises the modeling power from the network perspective.

Different from simple queuing theory of S-Q or V-Q models, LWR model

depicts the free-flow and congestion flow regimes and the transition between any traffic states in a holistic and coherent manner. Consequently, LWR model based traffic flow dynamics can more accurately predict the queue formation and its dissipation due to control actions. Therefore, this model and its solution schemes are more suitable to study the control strategies in a more general context.

2.5.2 Control Plan Computation Algorithms

Two types of optimization methods exist for computation of control plan: 1) the control problem is formulated as a mathematical program and solved via available linear and non-linear programming solvers and, 2) the control problem is described by complex system models and solved via heuristic searching algorithms. Selection of the optimization methods is highly tied to the underlying traffic flow models.

Mathematical Programming in Traffic Control Design

In (Stephanedes & Chang 1993), a forward time centered space treatment method was applied to modify the kinematic wave model and to depict the flow evolution within a corridor. Based on this treatment method, a non-linear program was built and solved by conjugate gradient algorithm to compute the dynamic ramp metering rates. The same solution method was also applied in (Kotsialos & Papageorgiou 2004), where the traffic flow on the freeway and ramps were modeled using a second-order continuum flow model(Messmer & Papageorgiou 1990). Because of the availability of highly efficient LP solvers, store-and-forward approach has also been used to model both the freeway and the urban street traffic(Papageorgiou 1995)(Diakaki et al. 2000). The resulting linear system is characterized by its sparsity and for such linear systems special LP solution algorithms exist(Banos &

Papageorgiou 1995). In (Gomez & Horowitz 2004*a*)(Gomez & Horowitz 2004*b*), the original cell transmission model was modified and led to a linear system that can compute the optimal ramp metering globally.

In a word, mathematical programming methods usually require the traffic flow models to be simplified so that the gradient information can be computed. Such simplification often compromises the underlying traffic flow models or it only computes the control plan for ad hoc situations.

2.5.3 Heuristics Method

In contrast to mathematical programs, heuristic searching algorithms, or known as "soft computing" techniques, does not rely on computation of gradient information so that more realistic and complex traffic systems can be used. Many heuristic methods have been adapted in traffic control studies; we only briefly review a few here.

Genetic algorithm (GA) is perhaps the most widely used heuristic searching algorithm in many other fields besides traffic control. GA emulates the biological evolution process of "Survival of the best fitness"(Goldberg 1989). Firstly, the control parameters to be optimized are selected and coded into chromosomes. A chromosome is a string resembling the genes of an individual. Usually, the parameter values coded in the chromosome are represented as series of unsigned decimal or binary digits. GA process starts by randomly generating a number of individuals in the population, each bearing a string of chromosome representing a feasible solution. These individuals form the initial population set, or the first generation in the evolution. From here GA evaluates each individual of its fitness, which is the objective function defined beforehand to measure the effectiveness of corresponding control plan. Each chromosome is then decoded into the actual control parameters. The

system performance is evaluated with these values evaluate the their fitness.

To emulate the evolution process, selected chromosomes are mated and manipulated to produce offspring. Usually a ranking scheme is applied to select chromosomes for reproduction, which means those with higher fitness function values have more chances to have 'children'. In the birth process, the chromosomes of parents undergo crossover and mutation to exchange or modify their gene codes purposefully. Crossover operation exchanges the parents' gene codes at random locations of the chromosome string, while mutation operation modify the gene codes randomly, namely up or down, in a small magnitude. In addition, an elitism strategy often sees its place in the reproduction process, to keep the chromosome with the highest fitness value for comparison with those computed from the next generations. If the individuals in the next generation fail to outperform the elite, its chromosome will be reinserted into the population of the next generation. The reproduction cycle will repeat itself until the stopping criteria are met.

GA-based computation methods have mainly been used for urban signal systems. Foy et al (Foy, Benekohal & Goldberg 1992) used GA to optimize the cycle length, green split and phasing order for a four intersection network. Hadi and Wallace (Haid & Wallace 1993) applied GA in conjunction with TRANSYT-7F hill-climbing procedure to optimize all four signal timing decision variables. Both studies took either microscopic simulation or external flow dynamics model to calculate the efficiency performance measure. Lo's study (Lo, Chang & Chan 2001)applied the cell transmission model (CTM) to evaluate alternative control plans.

Artificial neural network (ANN) takes a different approach to model the non-linear relationship between control plan and the system performance under the plan. The commonly used multi-layer feed-forward ANN comprises layers of neurons interconnected by forward connections. The strength of these connections are

characterized by their real-numbered weights. A neuron receives information from the ones in the preceding layer, processes the information on a internal model, and passes the processed information to the successive layer. The processing at each neuron takes a simple form of only including a summation operator, a local bias and a internal transfer function. However, because of the elaborate layer structure and the large amount of interactions between neurons on various layers, the processing as a whole behaves in a very complex manner and can lead to an acceptable replication between the input control parameters and the output performance after the *training* process. One application of the ANN was seen in (Zhang & Richie 1997). The study also applied the LWR model to depict the traffic evolution over the network. The training of the neurons used a back propagation algorithm, where a nonlinear system was easily solved to minimize the gap between the target density and the modeled density on a road section.

Although heuristic optimization methods can generally reach a global or near global optimal solution, one unavoidable feature must be mentioned. That is, using heuristic searching algorithms such as GA and ANN generally requires a large number of system performance evaluation before the stop criteria are met. In genetic algorithm, it is the number of generations and the number of *chromosomes* or candidate control plans that determine the computational overhead. In the case of ANN, it is the training process that forces the system to perform the desirable connection between the input traffic demand the output control plan. In this sense, finding a method that benefits from the advantages of both mathematical programs and heuristic method will be desirable.

2.6 Summary

Centering the theme of the dissertation study, three aspects concerning integrated corridor control are reviewed, the general control objectives and performance measures in terms of system efficiency and user equity, prevalent control methods and programs, and the underlying traffic flow model and control plan computation algorithms. The review has revealed that no previous method for traffic control considers both the overall performance of the corridor system and the fairness among individual travelers as well. The integration of control measures has been mostly done in an ad hoc way; and the traffic evolution realism has been compromised for either mathematical tractability or computation overhead reduction. Considering these deficiencies, we recognize the importance of computing a both efficient and equitable corridor control plan based on realistic traffic dynamics modeling. The study will investigate the three aspects successively.

Chapter 3

Fundamentals: Traffic flow Dynamics

3.1 Introduction

This chapter presents the building blocks of the study. Network flow dynamics model is the fundamental component in a reliable traffic control program, since the system performance under control will be evaluated with the model output. To be specific, a good traffic flow dynamics model will accurately estimate the travel costs of the system as a whole (efficiency) and those of individual travelers (equity). Since the travel cost calculation depends on the queuing evolution over the network, accurate estimation of the vehicle queues and prediction of its propagation and dissipation under control is vital to the success of a control system.

In Section (2.5.1), we concluded that the kinematic wave model, represented by the Lighthill-Whitham-Richards (LWR) model, is able to depict the density transition between various flow regimes and thus able to capture the shockwave mechanisms. We will then apply this model in our study and use one finite difference

solution scheme to model the traffic evolution over a general corridor network. We will first illustrate how the solution scheme, dubbed cell transmission model (CTM) by its author, is carried out, and then modify the model to accommodate the control measures including signal controls and ramp meters. Different from the previous studies, the modification is for general networks instead of ad hoc layouts of intersections or junctions.

Based on the flow dynamics model at general road sections (links) and controlled junctions (nodes), we are able to model the interaction between traffic demand, road supply and control plans. With all the necessary components ready, we will develop a study tool to investigate the process of traffic dynamics within a general transportation corridor. To be consistent with other literature in the field of dynamic traffic assignment, this process is generally called dynamic network loading (DNL). That is, given a time-dependent origin-destination (O-D) trip matrix, the DNL process will map the trip demand onto the network and retrieve the trip costs for every traveler groups (time-dependent O-D trips). With this trip cost information, the system performance, measured as system efficiency and user equity, will be quantified so that the desirable control plan could be selected.

3.2 Modeling Ordinary Link Dynamics With Cell Transmission Model

In this section we continue the discussion in section (2.5.1) and briefly present the link dynamics mechanism based on the cell transmission model(CTM). The following notations and symbols are used in this section.

Table 3.1: Notations for Cell Transmission Model Based Dynamic Network Loading Model

q:	link flow
q_{max}:	maximum flow rate or saturation flow rate
k:	density
k_j:	jam density
v:	link free flow speed
u:	backward propagation speed
Ç:	set of cells, ordinary($Ç_O$), diverging ($Ç_D$), merging($Ç_M$), source($Ç_R$) and sink($Ç_S$);
$Ç_G$:	set of controlled cells;
C_s^j:	the end cell of a link l_j that leads to an urban signal
n_i^t:	number of vehicles in cell i at time interval t
N_i^t:	maximum number of vehicles that can cell i can hold at time interval t
y_{ij}^t:	number of vehicles moving from cell i to cell j at time interval t
Ę:	set of cell connectors: ordinary($Ę_O$), diverging ($Ę_D$), merging($Ę_M$), source($Ę_R$) and sink($Ę_S$)
Q_i^t:	maximum number of vehicles that can flow into or out of cell i during time interval t
$\Gamma(i)$:	set of successor cells to i
$\Gamma^{-1}(i)$:	set of predecessor cells to i
δ_i^t:	ratio u/v for each cell i and time interval t
T:	loading horizon
t:	loading interval index
ϕ_l:	loading interval length

The CTM scheme discretizes the entire time horizon T into adjustable small *loading intervals* ϕ_l. Based on the loading interval, the model divides every link of the network into homogeneous segments called cells, in a way that the cell length is set equal to the distance traversed by one typical vehicle at free flow speed in one loading interval. If the relation of the flow q and the density k is in the form:

$$q = \min\{vk, q_{max}, u(k_j - k)\}, \qquad \forall 0 \leq k \leq k_j \tag{3.1}$$

the LWR model of highway link flows can be approximated by a set of difference

equations with the current system state being updated every loading interval. In doing so, the fundamental diagram as in (2.1) is simplified into a trapezoidal one (Figure 3.1), and the topology of the network is divided into cells according to the rules specified in below.

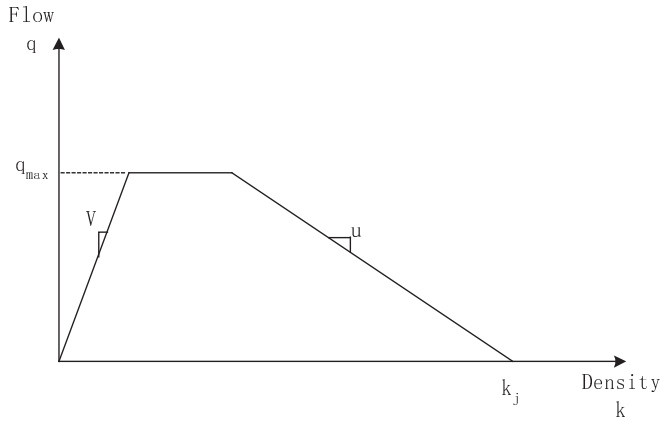


Figure 3.1: Trapezoidal Flow-Density Relationship used for Cell Transmission Model

The set of cells is divided in five sub-sets, ordinary($Ç_O$), diverging ($Ç_D$), merging($Ç_M$), source($Ç_R$) and sink($Ç_S$). They are illustrated in Figure (3.2). Ordinary cells are the ones that have only one predecessor cell and one successor cell ($|\Gamma| = 1, |\Gamma^{-1}| = 1$); diverging cells are the ones with one predecessor but two or more successor cells($|\Gamma| > 1, |\Gamma^{-1}| = 1$); merging cells are the ones with one successor cell but two more more predecessor cells($|\Gamma| = 1, |\Gamma^{-1}| > 1$); source cells are the ones with only one successor cell but no predecessor cell($|\Gamma| = 11, |\Gamma^{-1}| = 0$) and sink cells are those with only one predecessor but no successor cell($|\Gamma| = 0, |\Gamma^{-1}| = 1$).

Similar to this classification, the cell connectors, which only denote the traf-

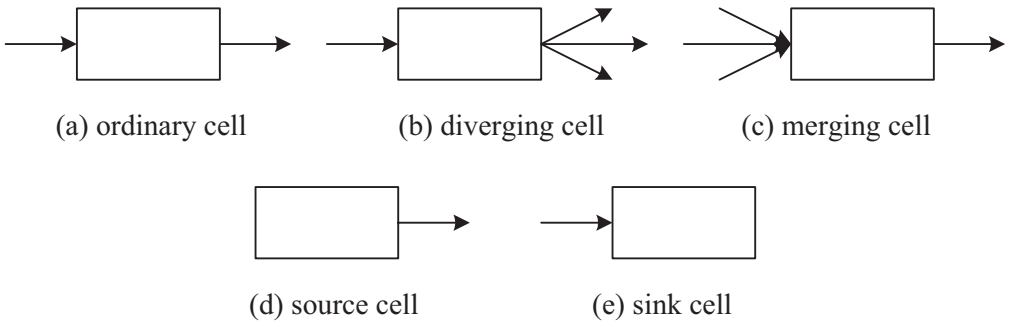


Figure 3.2: Categorization of Cells

fic direction and carry no flow, and thus differ from real network links, are also categorized into corresponding five sub-sets, ordinary($\underset{\sim}{\mathrm{E}}_O$), diverging ($\underset{\sim}{\mathrm{E}}_D$), merging($\underset{\sim}{\mathrm{E}}_M$), source($\underset{\sim}{\mathrm{E}}_R$) and sink($\underset{\sim}{\mathrm{E}}_S$). Also as in Figure (3.2), ordinary connectors are those with the beginning and ending cells are both ordinary; diverging are those with diverging cells as the beginning; merging are those with merging cell as the ending; source are those with the source cells as the beginning and sink are those with sink cells as the ending. one important rule of dividing cells is that merging cells cannot be connected to diverging cells directly; complex topological layout must break into simple merges and diverges.

By representing the network with the above segmentation and division, the traffic flow dynamics can be well represented (Daganzo 1995), and the flow updating rules are as follows.

Ordinary cells and cell connectors

The flow updating rule of ordinary cells and cell connectors (Figure 3.2-a) is represented by the following difference equations:

$$n_i^t = n_i^{t-1} + y_{ki}^{t-1} - y_{ij}^{t-1} \qquad k \in \Gamma_{-1}(i), j \in \Gamma(i), \forall i \in \underset{\sim}{\mathrm{C}}_O, \forall t \in T \tag{3.2}$$

and the flux

$$y_{ki}^t = \min\{n_k^t, \min\{Q_i^t, Q_k^t\}, \delta_i^t(N_i^t - n_i^t)\}, \quad \forall (k,i) \in \underset{\sim}{\mathrm{E}}_O, \forall t \in T \tag{3.3}$$

The first equation states the cell flow conservation, and the second equation states that the flow between two successive ordinary cells is bounded by three quantities, i.e., the number of vehicles in the current cell, the remaining holding capacity at the ending cell, and the minimum of the maximum flow that can get out of the beginning cell and the maximum flow that can get into the ending cell.

Diverging cells and cell connectors

The flow updating rule for diverging cells and cell connectors (Figure 3.2-b) reads:

$$n_i^t = n_i^{t-1} + y_{ki}^{t-1} - \sum_{j \in \Gamma(i)} y_{ij}^{t-1} \quad k \in \Gamma_{-1}(i), j \in \Gamma(i) \tag{3.4}$$

The inflow y_{ki}^t is given by the ordinary cell constraint of (3.3). It is assumed that the splitting proportion for each downstream cell is exogenously determined via the assignment model. As CTM can update the diverging flows with any allowable splitting proportions constrained by (3.6-3.7), it is interesting to note the following linear program that aims to maximize the overall outflux at diverging cells through manipulating the outfluxes y_{ij}^t

$$\max \sum_{\forall j \in \Gamma(i)} y_{ij}^t \tag{3.5}$$

s.t.

$$y_{ij}^t \le Q_j^t, y_{ij}^t \le \delta_j^t(N_j^t - n_j^t), \qquad \forall j \in \Gamma(i) \tag{3.6}$$

$$\sum_{\forall j \in \Gamma(j)} y_{ij}^t \le n_i^t, \sum_{\forall j \in \Gamma(j)} y_{ij}^t \le Q_i^t, \qquad \forall t \in T \tag{3.7}$$

As showed in (Ziliaskopoulos 2000), the constraints (3.6) and (3.7) are consistent with Daganzo's original version of diverging model (Daganzo 1995). The essential

meaning of the above constraint is that the diverging flow is also bounded by the following quantities: the overall sending flow demand of the preceding cell, the receiving holding capacity of each ending cell, and the flow capacity of the diverging cell connectors. In his work, Ziliaskopoulos used the above LP as a subroutine to solve the dynamic system optimal (DSO) assignment problem; here we list the LP formulation for the purpose of generalization, and our splitting proportion of $y_{ij}(t)$ is determined by the assignment methods to be introduced in the next chapters.

Merging cells and cell connectors

The merging flows (Figure 3.2-c) will be specified by the following:

$$n_i^t = n_i^{t-1} + \sum_{k \in \Gamma^{-1}(i)} y_{ki}^{t-1} - y_{ij}^{t-1} \qquad k \in \Gamma_{-1}(i), j \in \Gamma(i), \forall i \in \text{Ç}_M, \forall t \in T \tag{3.8}$$

The outflow y_{ij}^t is given by the ordinary cell constraint of (3.3). Similar to the treatment of diverging flows, one linear program can also be formulated to solve y_{ki}^t for maximization of throughput through the merge:

$$\max \sum_{\forall k \in \Gamma^{-1}(i)} y_{ki}^t \tag{3.9}$$

s.t.

$$y_{ki}^t \le n_k^t, y_{ki}^t \le Q_k^t, \qquad \forall k \in \Gamma(i)^{-1} \tag{3.10}$$

$$\sum_{\forall k \in \Gamma^{-1}(i)} y_{ki}^t \le Q_i^t, \sum_{\forall k \in \Gamma^{-1}(i)} y_{ki}^t \le \delta_i^t (N_i^t - n_i^t), \qquad \forall i \in \text{Ç}_M, \forall t \in T \tag{3.11}$$

The three constraints specify the maximum receiving holding capacity of the ending cell, the overall sending flows and the flow capacities of the merging connectors.

Source and sink cells and cell connectors

CTM boundary conditions compromise the states of source and sink cells and their initial values of all cells. The sink cells are designed to have infinite holding

capacities ($N_i^t = \infty, \forall i \in \text{Ç}_S, \forall t \in T$) and allow infinite influx ($Q_i^t = \infty, \forall i \in \text{Ç}_S, \forall t \in T$). Thus the input flow to a sink cell is actually determined by its preceding cell.

Source cells (Figure 3.2-d) are set to have infinite holding capacity ($N_i^t = \infty, \forall i \in \text{Ç}_R, \forall t \in T$) but finite outflow. The relationship is determined by the following equation:

$$n_i^t = n_i^{t-1} + d_i^{t-1} - y_{ij}^{t-1}, j \in \Gamma(i), \forall i \in \text{Ç}_R, \forall t \in T \tag{3.12}$$

where d_i^{t-1} is the demand at source cell i in time interval t. And the initial values n_i^0 for all cells can be set to the traffic conditions of the network at the beginning of the loading period.

We must note, however, the flow updating rules of merging and diverging cells and associated cell connectors are only for the road segments without controls. Unless at some specific locations where the road segments on either side become heterogeneous, e.g., lane drops or expansions, traffic controls are inevitable, no matter it is a signal or priority rule control, e.g., STOP signs or yield signs. As reviewed in the previous chapter, the flow dynamics at these junctions are vital to accurately evaluate the system performances. As a central theme in this dissertation study, the flow updating rules at controlled junctions will be studied in detail in the next section.

3.3 Flow Updating Rules at Controlled Junctions

The flow updating at various types of junctions are different, and they are classified into three categories in this section, urban signal intersections, ramp meters and priority rule controlled junctions.

3.3.1 Flow Updating at Signalized Urban Intersections

In (Lo 1999), Lo shows that CTM can be deployed to model the flow updates at urban intersections with a few modifications, and the major one is to make the maximum flow in equation (3.1) time-dependent in accordance with the signal timing. If the flow capacity q_{max} in equation (3.1) is replaced by a one that depends on the signal timing variable,

$$q_{max}^t = \begin{cases} q_{max} & t \in \text{ green} \\ 0 & \text{otherwise} \end{cases} \tag{3.13}$$

where it switches between q_{max} (green phase) and zero, the end cell of an intersection approach will serve as a functioning signal, and the flow dynamics still approximates the kinematic wave model. At a typical intersection, the traffic is grouped into movements that go through the conflicting area alternatively during their green time. A generalized four-leg intersection with all vehicular movements can be illustrated in Figure 3.3.

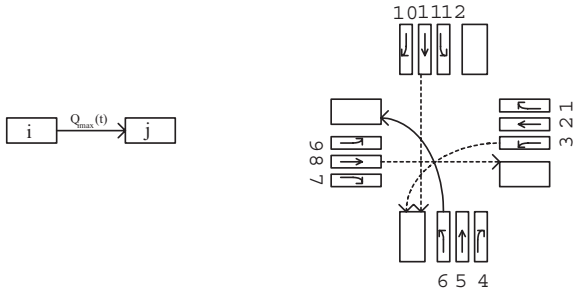


Figure 3.3: A General Representation of Cell-based Intersection Movements

Signalized Diverges

The flow diverges at an intersection occur where the traffic stream on a single link splits into left turn, through or right turn movements. It is a common practice to enlarge the intersection behind the stop line to store the turning vehicles temporarily for waiting to be serviced by certain phases, therefore, this feature is accommodated in the model as follows. Denote the end cell C_s^j of a link l_j approaching a signalized intersection, and the flow conservation equation then reads:

$$n_s(t+1) = \sum_{m=L,R,T} n_{s-1}^m(t) + y_{s-1,s}(t) - \sum_{m=L,R,T} y_s^m(t) \tag{3.14}$$

The superscripts of L, R, T denote the left turn, right turn and through movement, respectively. The cell C_{s-1}^j is the preceding cell of C_s^j . The flux into and out of cell C_s^j are stated as:

$$y_{s-1,s}(t+1) = \min\{n_{s-1}(t), Q_{s,max}, \delta_s(N_s - n_s)\} \tag{3.15}$$

$$y_{s,s+1}^m(t+1) = \min\{n_s^m(t), q_{s,max}^t, \delta_{s+1}(N_{s+1}^m(t) - n_{s+1}^m(t))\}, m = L, R, T \tag{3.16}$$

Note that $N_{s+1}^m(t), m = L, R, T$ in equation (3.15), the storage capacities for various movements, ensures that different sizes of turning bays can be modeled accurately. It also implies extra feeding data input to the model.

Signalized Merges

In this study, the right turns are explicitly considered in the signal timing optimization. This is due to two practical considerations. First, this study is aimed at large scale applications and thus the detailed priority information might not be viable at the modeling stage. Second, in many places the merging flows are also treated as conflicts. In Figure 3.3, for example, movement 4 may not be in the same

phase with movement 8, because there will be conflicts between these two streams. In this way, the flow updating at intersections is simplified to be the same as a set of coupled consecutive links. The flow updating rules will then be:

$$n_{s+1}(t+1) = n_{s+1}(t) + y_s(t) - y_{s+1}(t) \tag{3.17}$$

where $(s+1)$ is the start cell index for the downstream link, i.e., the first cell of the downstream link that receives the stream with cell index of s serviced by the signal. The incoming flux $y_s(t)$ is then determined by the signal timing plan:

$$y_s(t) = \min\{n_s(t), q^t_{s,max}, \delta_s(N_s - n_s)\} \tag{3.18}$$

and $y_{s+1}(t)$ is determined by y^t_{ks} in (3.3). One notes that this simplified method has also been used throughout Lo's study (Lo 1999)(Lo 2001).

3.3.2 Metered Freeway Onramp

Modeling ramp meters has only one control variable to deal with, the metering rate R^t_j at on-ramp j at time t. For notion simplification, the ramp subscript j is omitted in the following development. One generic updating rule is applied to represent the flow dynamics at a freeway merge section(Jin & Zhang 2003):

$$D^t_R = \min(D^t_R, R^t, q_{max}) \tag{3.19}$$

$$D^t = D^t_M + D^t_R \tag{3.20}$$

$$S^t = \min(S^t_M, D^t) \tag{3.21}$$

$$f^t_M = \frac{D^t_M}{D^t} S^t \tag{3.22}$$

$$f^t_R = \frac{D^t_R}{D^t} S^t \tag{3.23}$$

where the ramp metering R^t is embedded, and the rest of the notations apply:

D_R^t: ramp demand at time t;
D^t: demand upon the beginning cell of the link downstream of the ramp;
D_M^t: demand on mainline competing with the ramp demand;
S_M^t: supply of the beginning cell of the downstream link;
S^t: total service flow rate ;
f_R^t: outflow from ramp ;
f_M^t: outflow from upstream mainline;

The modification mainly lies in two aspects: (i) the ramp demand to the merge point is bounded not only by actual demand and the flow capacity, but also by the metering rate executed at that time step (Equation 3.19); (ii) in the overflow or congestion situation, the freeway mainline and ramp flows will be distributed proportionally to their relative demand (Equations 3.21-3.23) . The ramp metering takes effect in the form of R^t.

3.3.3 Priority Controls

Stop Sign

Stop sign control operates on a "first-come-first-serve" basis; the essence is the ordered flow discharge. An All-way STOP sign intersection node has one set of incoming link $i \in \mathcal{L}_{in}, i = 1, ..., L$. We maintain a numerical array of right-of-way, $ROW_i(t) = i, i = 1, ..., L$, corresponding to the priority order of each incoming link i at loading interval t. At each loading interval, the approach with the lowest $ROW_i(t)$ will have the right-of-way to release vehicles. Flow updating at All-way STOP sign is to determine the approach I with the lowest $ROW_i(t)$ at the current interval and then to update the flows following the ordinary cell transmission rules (3.2-3.3).

Denote d_{ij} as the demand from incoming link i to outgoing link j; s_j as the supply of the link j. At each loading interval, the critical step is to determine the next approach I to release flow. The algorithm reads as follows:

Flow updating at All-way Stop Sign Control

Initialize $ROW_i(0) = \infty, i \in \mathcal{L}_{in}$.

All-way stop: if $\sum_i \sum_j d_{ij}(t-1) = 0, \forall t \in T$, return; else, go to next step.

Determine approach I.

Iterate through $i \in \mathcal{L}_{in}$:

$$\begin{aligned} &ROW_i(t) = L && \text{if } ROW_i(t-1) = 0, \\ &ROW_i(t) = L && \text{if } \textstyle\sum_j d_{ij} = 0, \\ &ROW_i(t) = ROW_i(t) - 1 && \text{if } \textstyle\sum_j d_{ij} > 0 \end{aligned}$$

Determine the minimum index of ROW:

$$I(t) = \min_i \{ROW_i(t)\}$$

Let

$$ROW_I(t) = 0.$$

Flow discharge follows (3.2) and (3.3) for link $I(t)$.

Two issues need further clarification here. It is easy to verify the validity of the step *Determine approach* I in the case of immediate flow discharge when only one approach has demand and all the others have none. When the node has only two-way stop signs, the flow updating will use a similar treatment to yield sign, which will be described in the next section.

Yield Sign

The updating rule for competing flows at yield sign control is essentially merges under priority rules. Under yield sign, the yielding flow will only be able to take the remainder of the available space at each loading interval. At any loading

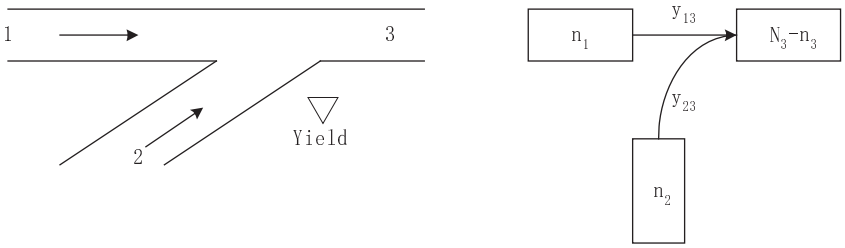


Figure 3.4: Flow Updating by Priority Control: Yield Sign

interval t, the flow updates at a yield sign will be specified by:

$$y_{13}^t = \min\{n_1^t, N_3 - n_3^t, q_{1,max}\} \tag{3.24}$$

$$y_{23}^t = \min\{n_2^t, \max\{N_3 - n_3^t - y_{13}^t, 0\}, q_{2,max}\} \tag{3.25}$$

where the flow on approach 1 has the priority and the flow on 2 has to yield to flow 1.

3.3.4 Traffic Demand Input

In the model, the traffic demand is given externally at any source link j:

$$Q_{j1}(t) = \sum_r \sum_s D^{r,s}(t), \forall (r,s) \in \{(r,s) | r \in \mathcal{R}, s \in \mathcal{S}\} \tag{3.26}$$

where $Q_{j1}(t)$ is the sum of the time-dependent demands entering the source node j. Then $Q_{j1}(t)$ will be transferred to d_i^t in the equation (3.12) to start their journey.

3.4 Adaptation of Basic Control Methods

Controls at individual intersections or ramp meters are the building block for a network control. It will be the focus of this section to adapt two basic signal

controllers, pre-timed controller and vehicle actuated controller, and their typical control methods into the network dynamics model established in the above sections.

The first issue is the specification of control variables after discretization and division of the network from CTM scheme as above. The control parameters in any control plan will have to be adapted first. One simple rounding scheme is applied in this study to approximate the control parameters to their nearest integer times of the loading interval. For example, if the split of one phase is 44 seconds and the loading interval is 3 seconds, then the phase split will be rounded to 45 seconds. Regarding the yellow time and loss time, the simplified of using only effective green and effective red is applied, so as to fit in Equation (3.13).

3.4.1 Development of Signal Controllers within DNL

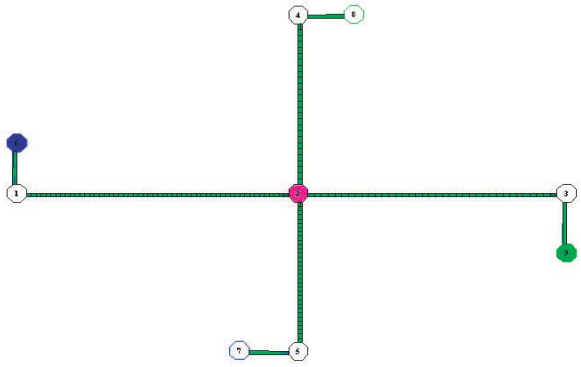
Pre-timed Controllers

Pre-timed signal controller and its control plan are characterized by cyclic repeat of a fixed sequence of phases, and the duration of the phases are also constant values. This is the simplest controller and the control variables can be easily embedded into the framework specified in (3.13-3.18).

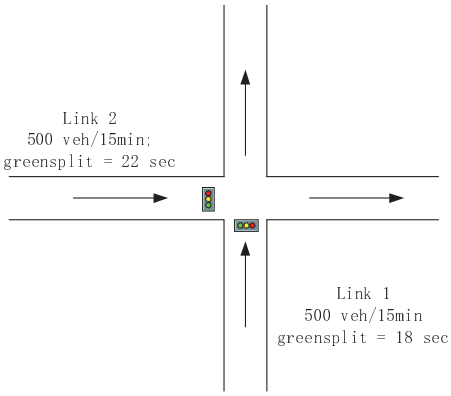
The following simple two-approach intersection (Figure 3.5) can best illustrate the effectiveness of using the above DNL platform to model the signal control. The intersection has two conflicting flows. The traffic demand and signal control plan is also illustrated in the figure. Equal number of demands are released uniformly onto the tail end of both links. Also for illustration purpose, the green split is designed to favor the horizontal link (Link 2). The loading results are shown in Figure 3.6 and Figure 3.7.

The link dynamics figures (Figure 3.6 and Figure 3.7) clearly indicate how the traffic density changes are induced from the intermittent service of the signals.

For Link 2, the signal cannot serve the traffic in a timely manner in that the same level traffic demand was less favored by the split (18 seconds vs. 22 seconds). Therefore, the gap between the arrival curve and the departure curve at this approach gets wider until all the traffic has been loaded (15 minutes). This is reflected in the traffic states behind the stop line. The link density is in a constant transition and the queues build up as well, illustrated by the density changes in Figure(3.7). While all dynamic traffic models can get to Figure(3.6), it is only advantageous to apply LWR and consequently the CTM scheme to model the density transitions as illustrated here.



(a) Cell Division at the Two-approach Intersection



(b) Demand and Control Settings for the Pre-timed Controller

Figure 3.5: Pre-timed Controller Experimentation on a Simple Two-approach Intersection

Vehicle-actuated Signal Controllers

Vehicle actuated (VA) traffic signal specifications are introduced in most traffic engineering text. A typical VA controller specification can be tailored as follows in our framework. An intersection has I predefined phases, $P_i \in |P|_I, i = 1, 2, ..., I$, where the ranking order of i denotes the phasing sequence. For every phase P_i, the minimum and maximum greensplit is determined as $G_{i,min}$ and $G_{i,max}$, respectively. The *vehicle extension* (c.f. Section 2.2.1) is also set to be the multiple(s)

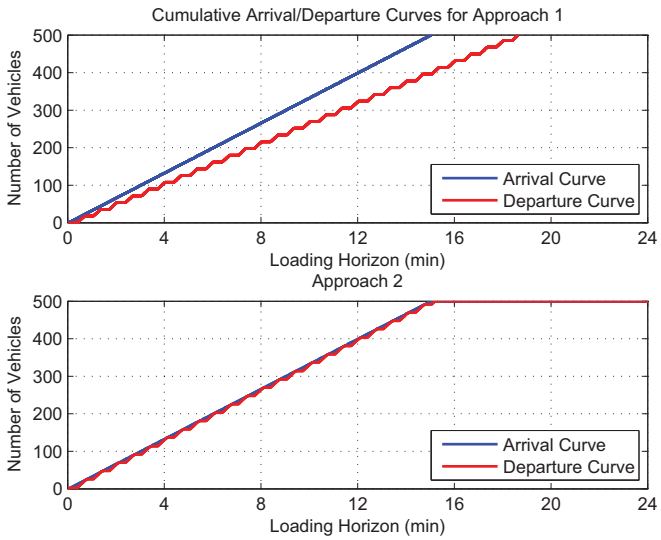


Figure 3.6: Cumulative Arrival/Departure Curves at the Two-approach Intersection

of the loading interval. One VA algorithm can be illustrated as follows.

At loading interval t, when a phase P_i is under green and has already gone through a green duration of G_i^t since the green starts, then the signal states will be determined by the following "if-then" conditions:

Algorithm: Vehicle-Actuated Signal Control Process ("Free-running")

IF $G_i^t < G_{i,min}$

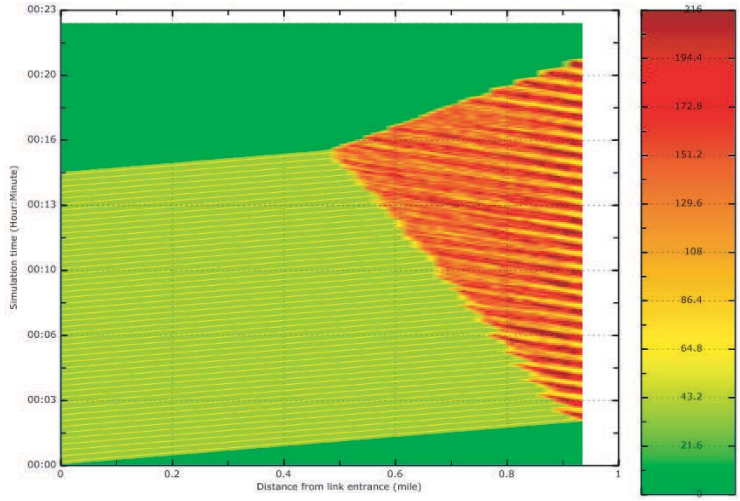
$$G_i^{t+1} = G_i^t + \phi_l,\ t = t + 1$$

Break and GoTo next interval;

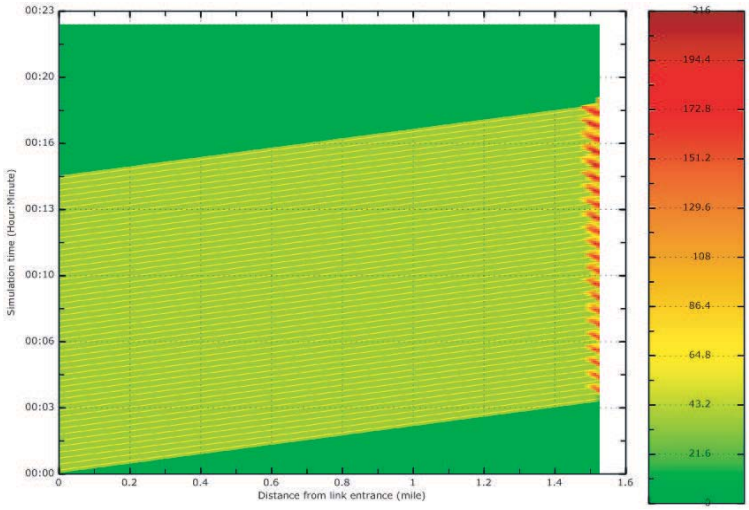
ELSE IF $G_i^n < G_{i,max}$,

IF the Demand n_i^t *upon the phase* k $n_i^t > 0$

$G_i^{t+1} = G_i^t + \phi_l$, and $t = t + 1$,



(a) Density Transitions (Link 2)



(b) Density Transitions (Link 1)

Figure 3.7: Traffic States (Density) Transitions Behind the Stop Line

ELSE

$$G_i^t = 0, P_i = P_{i+1}, G_i^{t+1} = G_i^t + \phi_l, \text{ and } t = t + 1,$$

ELSE

$$G_i^t = 0, P_i = P_{i+1}, G_i^{t+1} = G_i^t + \phi_l, \text{ and } t = t + 1,$$

The above VA process is generally called "free-running", as all phases possess *presence detectors* that can provide the demand n_i^t as in the above algorithm. No phase is superior to the others; that is, no major traffic flow directions are specified. More elaborate settings are adaptable with slight changes from above.

Signal Coordination

The signal coordination is characterized by the offset Δ specified at the signal groups of each intersection. That is, the start of the green time of the first phase of the signal groups. Similar to other control parameters, the offset is also normalized to be the multiples of the loading interval.

The on-ramp meters are nothing but a specific case of the above signal controllers, since there is only one signal is present at each of the ramp metering junction, following the flow updating rules in Equation(3.19-3.23). Different ramp metering algorithms are adapted as well, with more details specified in the next chapter when we present the local synchronization control schemes.

3.5 Computation of Travel Cost

3.5.1 Calculation of Link Travel Cost

The fundamental diagram indicates two regions that traffic flow status can fall into, the free flow region and the forced flow region. Once the flow falls in the forced flow region, the vehicles will not operate at the free flow speed any more, and delays are incurred to the vehicles. Within the DNL framework built on cell

transmission model, at a given time t, a vehicle, *or* flow quantum, can either move to the next cell along the trip, or it can stay in the cell. Then at the cell level, the delay is calculated as the following:

$$d_i(t) = d(n_i(t) - y_i(t)) = (n_i(t) - y_i(t)) \times \phi_l \tag{3.27}$$

where $d_i(t)$ is the delay occurring at cell i at loading interval t, and n_i, y_i has the same meaning as before. At the link level, the travel time is the free flow journey time, expressed as the integer multiples of loading interval, plus the delay incurred on the link. Given link l and its cells that are ordered from link entry to exit as $i = 1, ...K$, the link traversal time of a certain vehicle entering the link at time t will then be calculated from the following function:

$$\tau_l^1(t) = d_1(t) + \phi_l \tag{3.28}$$

$$\tau_l^2(t) = d_2(\tau_l^1(t)) + \phi_l, ... \tag{3.29}$$

$$\tau_l(t) = d_K(\tau_l^{K-1}(t)) + \phi_l \tag{3.30}$$

In a simple form, it would be written as:

$$\tau_l(t) = \sum_{i=2}^{K} d_i(c_l^{i-1}(t)) + K\phi_l \tag{3.31}$$

where $\tau_l^i(t)$ denotes the time the vehicles spend within the cell i on link l when entering the cell at time t.

3.5.2 Calculation of Path Travel Cost

Similar to the calculation of link traversal time, the path travel time for a certain travel groups, the travelers with the same origin-destination in this study, can also be calculated recursively from the dynamic network loading results. Consider a

path p consisting of sorted nodes $N^(r,s)_p = (r, 1, ..., s-1, s)$, from r to s. When a user departs from origin r at time t,

$$c^{r,1}(t) = \tau_{r,1}(t) \tag{3.32}$$

$$c^{r,2}(t) = c^{r,1}(t) + \tau_{1,2}(t + c^{r,1}(t)), ... \tag{3.33}$$

$$c^{r,i}(t) = c^{r,i-1}(t) + \tau_{i-1,i}(t + c_p^{r,i-1}(t)), ... \tag{3.34}$$

$$c_p^{r,s}(t) = c_p^{r,s-1}(t) + \tau_{s-1,s}(t + c^{r,s-1}(t)) \tag{3.35}$$

where $\tau_{i,j}$ is the actual link travel time on link (i, j) calculated from Equation (3.31).

For the entire system, the total travel time is the summation over all O-D pairs through the entire analysis horizon:

$$TTT = \sum_{(r,s)} \sum_{p} \sum_{t} c_p^{r,s}(t) \times f_p^{r,s}(t) \tag{3.36}$$

TTT from (3.36) will be our system efficiency measure.

For later analysis, we also give the following definition of relative path travel cost.

Definition 3.1 *Relative path travel cost is the ratio of the travel cost of certain path with regard to its nominal travel cost. The nominal travel cost can be that under equilibrium flow pattern or free flow conditions.*

Relative path travel cost is a better measure to compare the path travel costs between different O-D trips. For a path p with sorted nodes $N_p^{r,s} = (r, 1, ..., s-1, s)$, the nominal path travel cost can be defined as the sum of the free flow travel time traversing all associated links:

$$\tau_{p,0}^{r,s} = \sum_{i=r}^{s-1} \tau_0^{i,i+1} \tag{3.37}$$

The relative path travel cost is then computed as:

$$d_p^{r,s}(t) = \frac{c_p^{r,s}(t)}{\tau_{p,0}^{r,s}} \tag{3.38}$$

In the later development, a variation of (3.38) will also be used:

$$D_p^{r,s}(t) = \frac{c_p^{r,s}(t) - \tau_{p,0}^{r,s}}{\tau_{p,0}^{r,s}} \tag{3.39}$$

where $D_p^{r,s}(t)$ is the relative path travel delay. To note that minimization of (3.39) is equivalent to that of (3.38) when the path is fixed.

3.6 Measuring User Equity

As reviewed in Section 2.1, user equity is defined in two broad category, horizontal equity and vertical equity. Translated in the corridor control context, improving the horizontal equity will imply that the traveler groups experience equal delays without regard to their trip lengths. On the other hand, improving vertical equity implies that the travel costs of different O-D pairs will be proportional to their nominal travel cost, e.g., travel cost under free-flow conditions. Continuing Section 3.5.2, we can immediately conclude that path travel cost and relative path travel cost (RPTC) correspond to the horizontal equity and vertical equity, respectively.

As implied in the review, the variables of measuring the user equity of a control system can be classified into two categories, the aggregate ones and the disaggregate ones. Aggregate equity characterizes the overall benefits distribution with respect to all user groups. Representative aggregate equity measures include Gini Coefficient (Gini 1936) and the associated Lorenz Curve(Lorenz 1905).

In this study, traveler groups are classified according to their characteristic of origin-destination . For all traveler groups according to their departure time and

origin-destination characteristics $\forall(r,s) \in (\mathcal{R},\mathcal{S})$, the path travel costs (3.37) or RPTC (3.38) can be reordered in their ascending order:

$$c_{(1)} < c_{(2)} < ... < c_{(\mathcal{W})}, \tag{3.40}$$

and

$$d_{(1)} < d_{(2)} < ... < d_{(\mathcal{W})}, \tag{3.41}$$

where $\mathcal{W}$ denotes the total number of (time-dependent) O-D pairs in $(\mathcal{R},\mathcal{S})$. Their corresponding number of trips are denoted as $f_{(w)}, i = 1, 2...\mathcal{W}$ successively.

The aggregate and disaggregate user equity measures are defined in the following section.

3.6.1 Aggregate Equity Measures

Three aggregate measures can be defined here:

- Gini Coefficient
- Relative Mean Difference (RMD)
- Mean Difference (MD)

Gini Coefficient and Lorenz Curve

Referring to Figure 3.8, the Lorenz Curve concerning the path travel cost is drawn from reorganizing the path travel cost. Then the Lorenz Curve is a cumulative distribution of a population, drawn to show how much percentage $y\%$ of the total travel delay is experienced by bottom $x\%$ percentage of the driving population. After reordering the path travel cost according to (3.40) or (3.41), the Lorenz Curve for the transportation corridor system will be easily drawn as in Figure 3.8. Once the

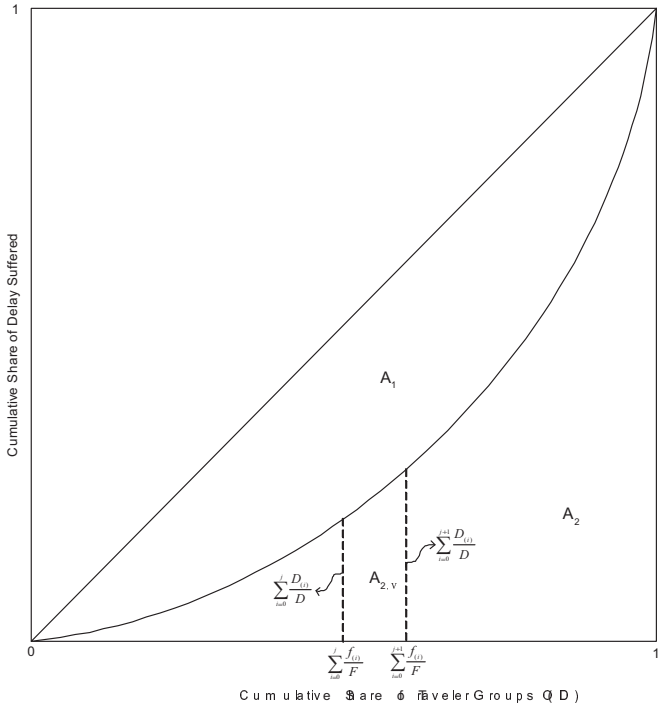


Figure 3.8: Lorenz Curve and Gini Coefficient

Lorenz Curve is drawn, the Gini Coefficient will then be calculated as:

$$Gini = \frac{A_1}{A_1 + A_2} = 1 - 2A_2 \tag{3.42}$$

Differentiating the components, we have

$$A_{2,v} = \left[\sum_{w=0}^{v} \frac{D_{(w)}}{D} \frac{f_{(w)}}{F} + \sum_{w=0}^{v+1} \frac{D_{(w)}}{D} \frac{f_{(w)}}{F} \right] \bullet \frac{f_{(v+1)}/F}{2} \tag{3.43}$$

where

$$D = \sum_{i=1}^{W} D_{(w)} \qquad F = \sum_{w=1}^{\mathcal{W}} f_{(w)} \tag{3.44}$$

$$\begin{aligned} A_2 &= \sum_{v=0}^{W-1} A_{2,v} \\ &= \sum_{v=0}^{W-1} \left[\left(\sum_{w=0}^{v} \frac{D_{(w)} f_{(w)}}{DF} + \sum_{w=0}^{v+1} \frac{D_{(w)} f_{(w)}}{DF} \right) \frac{f_{(v+1)}/F}{2} \right] \\ &= \frac{1}{DF^2} \left[\sum_{v=0}^{W-1} \sum_{w=0}^{v} D_{(w)} f_{(w)} f_{(v+1)} + \sum_{v=0}^{W-1} D_{(v+1)} \frac{f_{(v+1)}^2}{2} \right] \end{aligned}$$

Resulting in:

$$Gini = 1 - 2\frac{1}{DF^2} \left[\sum_{v=0}^{W-1} \sum_{w=0}^{v} D_{(w)} f_{(w)} f_{(v+1)} + \sum_{v=0}^{W-1} D_{(v+1)} \frac{f_{(v+1)}^2}{2} \right] \tag{3.45}$$

Mean Difference and Relative Mean Difference

Relative mean difference (RMD) is considered an estimate of the Gini Coefficient, and statistics text shows that it is approximately twice as large as Gini Coefficient. Its calculation is as follows:

$$RMD(D) = \frac{\sum_{v=1}^{\mathcal{W}} \sum_{w=1}^{\mathcal{W}} |D_w - D_v|}{(n-1) \sum_{w=1}^{W} D_w} \tag{3.46}$$

Similar to Gini Coefficient, it is a dimensionless measure.

For the sake of comparison, the absolute mean difference is introduced as well:

$$MD = \sum_{w=1}^{\mathcal{W}} \sum_{w'=1}^{\mathcal{W}} |D_w - D_{w'}| \tag{3.47}$$

Apparently Mean Difference (MD) has the same unit as the path travel costs. In this sense, it may become a convenient measurement to be linearized with TTT so as to balance the efficiency and equity measures in the later chapter.

3.6.2 Disaggregate Equity Measures

Different from aggregate equity measures, disaggregate ones focuses on individual traveler groups. Mostly it is concerned with the most disadvantaged traveler groups as has been used in a few studies as the only equity measure(Meng & Yang 2002)(Chen & Yang 2004). It is called the *critical trip cost ratio* and simply the (relative) path travel cost of the last traveler group after the costs are reordered (3.40). That is,

$$CR = c_{(\mathcal{W})}, \textit{or } CR = d_{(\mathcal{W})} \tag{3.48}$$

depending on whether horizontal equity or vertical equity is concerned.

Another disaggregate measure quantifies the *range* of the distribution of the travel costs:

$$RG = c_{(\mathcal{W})} - c_{(1)}, \textit{or } RG = d_{(\mathcal{W})} - d_{(1)} \tag{3.49}$$

depending on which type of equities is concerned. The range RG is sometimes a weak complement to CR, since the best travel cost that can be achieved is the free-flow travel time. When vertical equity is concerned and the free-flow travel cost as nominal cost, the range is equivalent to the critical trip cost ratio.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

Experimental Investigation of Corridor Control Strategies

4.1 Introduction

In this chapter, we investigate the efficiency and equity performances of the integrated control strategies that take the traffic information from only the local environment. Usually this type of control strategies is not rigorously formulated as either mathematical programs, instead they are generally rule-based control logic that are composed of "if-then" conditions. These types of control strategies generated large interests from practitioners as well as researchers, since these strategies are less data-intensive, less sensitive to detection failures, and thus more robust.

The primary goal of these local strategies is to prevent the queues from spreading beyond the local scope, e.g., an interchange area. Besides the goal of congestion alleviation, maximum waiting time and maximum queue length constraints are generally imposed to improve user equity (Zhang & Levinson 2003). The past research generally focused on one specific control rules (Tian et al. 2002), but no

generalized forms were provided to guide the design of these control strategies. In this chapter we build a more general form of coordination control schemes to locally integrate the signal control and ramp metering.

The congestion initiated at closely spaced highway junctions and intersections, e.g., freeway interchange areas, may spread and severely degrade the operational efficiency of the whole network if not handled timely and properly. From this observation, we propose this local synchronization control (LSC) scheme. The scheme manages queues at those critical locations through coordinating neighboring intersection traffic signals and freeway on-ramp meters wherever available. By reducing the amount of traffic feeding into and increasing the amount of traffic discharging from heavily queued sections, the scheme can prevent a queue from evolving into the seed of a gridlock and thus improve the overall system performance.

To examine its effectiveness, other control strategies including isolated control and global optimal control are also developed and compared. As the central theme, their relative performances in both efficiency and equity are also compared using a contrived corridor network.

4.2 Local Synchronization Control Schemes

Illustrated in the flowchart of Figure 4.1, the proposed LSC is centered in monitoring the traffic operations and particularly the vehicle queuing on critical links. Usually such queues that are not dissipated timely will evolve into local or even network wide gridlock or crawling conditions. When such a potential spillback is detected, normal traffic operations will be superseded by local synchronization control (LSC) schemes, where the control actions are synchronized to discharge the queuing traffic and reduce the feeding traffic simultaneously. The normal operations will resume after the critical queue is cleared.

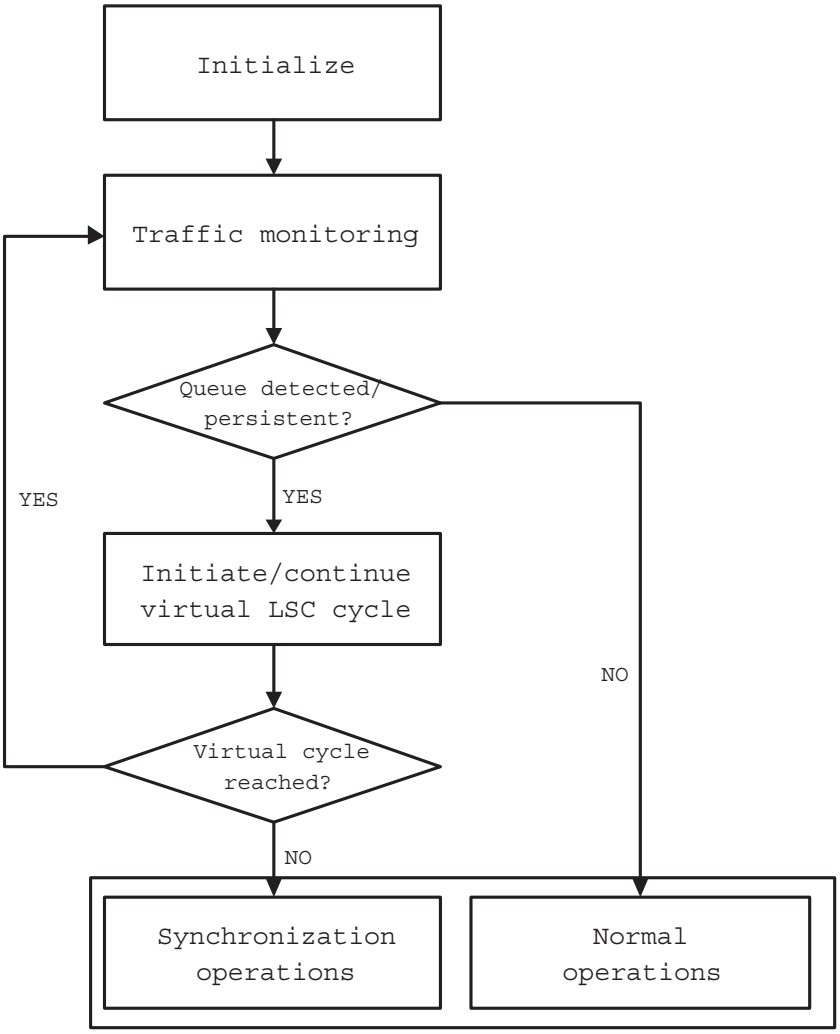
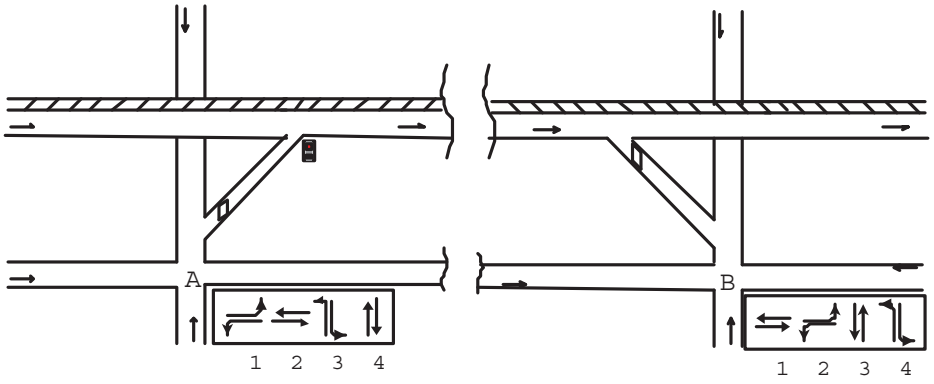


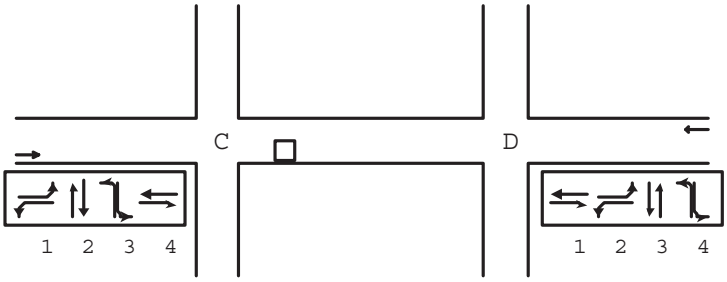
Figure 4.1: Flowchart of Local Synchronization Control Schemes

Figure 4.2: Possible Local Synchronization Control Units



(a) On-ramp priority

(b) Off-ramp priority



(c) Intersection internal metering

Typical road sections that require synchronization treatment are those short ones controlled on either one end or both ends and carry large interfacing flows. Sample sections are on-ramps with either on-ramp meters, or signals that controls the feeding traffic, or both; off-ramps leading to a signalized intersection; and short road sections connecting two tandem signals. These sections are commonly seen in any freeway interchange area. Correspondingly, three synchronization operations are developed in this study (Figure 4.2) as follows.

On-ramp priority (4.2a). The queue detector is positioned at the upstream end of the metered on ramp. When the spill back is detected, the synchronization operation will be triggered: 1) setting the current meter off, i.e., ramp traffic can compete freely with the freeway traffic for the right of way; 2) reducing the maximum green time (or duration) of the phases that feed to the ramp traffic. For example, in Figure 4.2a, it means that the maximum green time phase 1 and 4 at intersection A will be reduced by an Adjustment factor to be explained next.

Off-ramp priority (Figure 4.2b). The queue detector is positioned at the upstream end of the off ramp. Once triggered, the maximum green time (or duration) that discharge the off ramp flow (phase 3 and 4 of intersection B in Figure 4.2b) will be increased by the Adjustment factor.

Street internal metering ("gating") (Figure 4.2c). This operation considers the queues building up on a critical street section. Typical sections are the cross street links within interchange areas that have signals controlling the interfacing traffic between the freeway and the street. The queue detector is positioned on the upstream end of the critical section; once triggered, the feeding phases (phase C3&C4) will become shorter by the Adjustment factor whereas the

discharging phases (phase D1& D2) will become longer.

It is shown that the synchronization schemes apply "if-then" rules to control the traffic flow at these critical sites. Rule-based control methods are generally parametric in that they often require one or a set of adjustment factors specified to fine-tune the control rules. Herein the LSC scheme is characterized by three important factors as well:

- Queue detector position. To set up any LSC scheme, the primary step is to determine the critical queuing location, beyond which the spillback will start to block other traffic and should be avoided. This task needs the observation or knowledge of how local congestion evolves during the study period.

  In the DNL tool established in the previous chapter, the queue detection is modeled as tracking the occupancy changes at the detection locations. One can notice that if the traffic flow dynamics is modeled using cell transmission model, the occupancy will be naturally emulated as the ratio of the number of vehicles to the holding capacity at the location of interest. That is,

  $$O_i^t = \frac{n_i(t)}{N_i} \tag{4.1}$$

  Where $n_i(t)$ and N_i are the same notation as in (3.2-3.12), and O_i^t is the occupancy of bottleneck location designated by the cell i, corresponding to the occupancy in Equation (2.34).

- Virtual cycle of synchronization operation (c.f. Figure 4.1). Virtual cycle specifies the duration of synchronization operations. Once virtual cycle is triggered, the synchronization operations will continue until it reaches the virtual cycle. Another virtual cycle will be initiated afterwards if the queue persists or otherwise normal operations will resume. This virtual cycle will then be determined

in conjunction with the queue detection: it equals the time period required to discharge the queue under the synchronization scheme. In this CTM based evaluation platform, this parameter is set equal to the number of intervals for the queued traffic to traverse the entire queued sections.

- Adjustment factor of synchronization intensity. The adjustment factor determines how much the LSC unit will meter the affected phases, i.e., how long the discharging phase will be increased and how long the feeding phases decreased. It is set to avoid the side effect from an aggressive LSC strategy, e.g., reducing the feeding phases to its minimum green time all the time (Tian et al. 2002). If the volume of traffic feeding to the area is high, the LSC unit cannot be too restrictive and thus the adjustment factor cannot be high. Experiments showed that the adjustment factor of one third of the maximum duration is appropriate to produce satisfactory synchronization effect.

4.3 Genetic Algorithm Based Global Optimal Control: System Efficiency Only

The local synchronization control schemes take only the local traffic information as input and act locally as well. It is not known how well the rule-based control strategies can perform for the entire system. We thus develop a global optimal control strategy to compare with the rule-based strategies. In this section, a genetic algorithm based integrated corridor control program is built, partly following the modeling structure in (Lo et al. 2001) as reviewed in Section(2.5.3).

A typical procedure to perform GA computation is illustrated in Figure 4.3.

The manipulations in genetic algorithm, including gene-coding, replace-

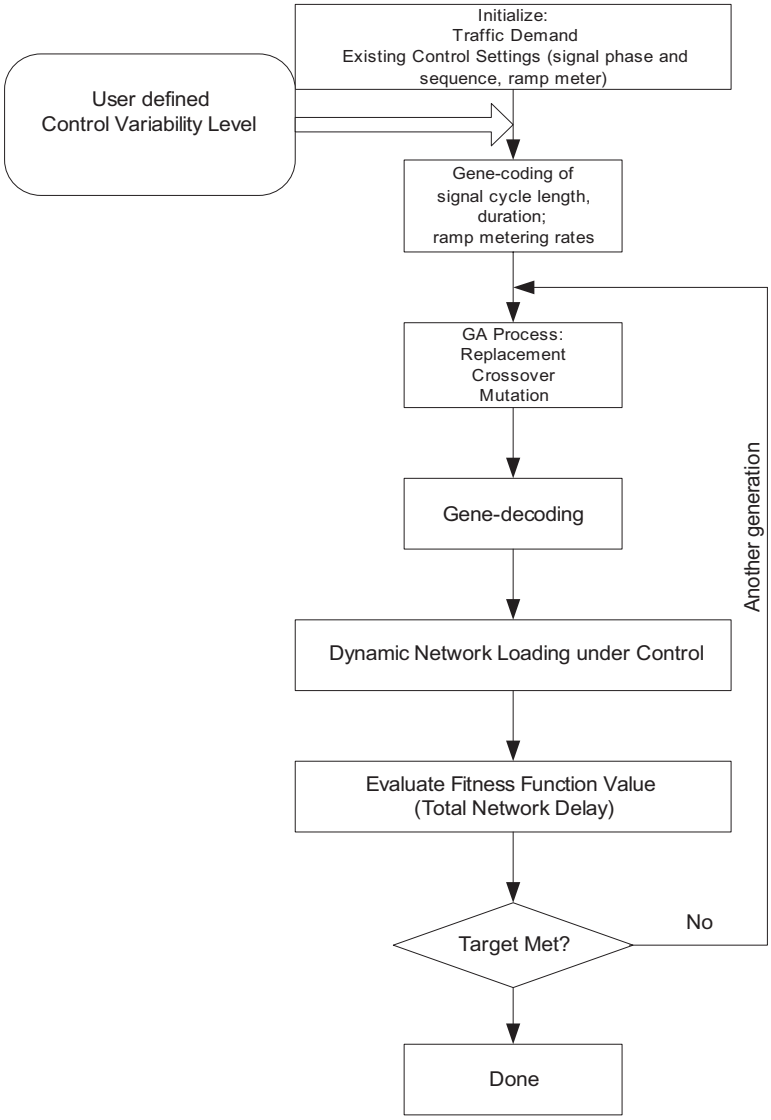


Figure 4.3: Using Genetic Algorithm to Optimize Integrated Corridor Control Plan (System Efficiency Only)

ment, crossover and mutation, have been documented in detail in text (Goldberg 1989) and is not necessary to repeat here. Nevertheless, we applied a few noteworthy features in the implementation for solving the integrated corridor control problem.

Firstly, control plan is expected to be adaptive to network flow conditions. The GA control module applies a flexible structure to allow variable control update periods. For example, a 15-minute updating period is expected to perform better than one hour, since the demand pattern may well differ from one 15-minute interval to another within the same hour.

Secondly, various control variability levels are allowed, that is, the set of control variables to be optimized is not fixed. For example, three levels of variability have been tested: 1) optimizing only the signal greensplit and ramp metering rates within the corridor; 2) optimizing the greensplit, cycle length and phase sequencing and ramp metering rates; 3) optimizing all control variables, namely greensplit, cycle and phasing order, offset and ramp metering rates. The tests confirmed that higher control flexibility can improve the system efficiency.

Lastly, real-valued gene-coding is applied to improve the performance of genetic algorithm itself. Most genetic algorithm based applications apply binary coding method: the values of genes in the chromosome are coded in a binary string for genetic operations. However, as argued in (Wright 1991), real-valued coding is more promising in terms of computational efficiency, because 1) no conversion between decimal and binary is necessary, 2) greater freedom is gained to use different genetic operators and, 3) higher mutation rates can be utilized so as to better explore the search space. In the developed GA based corridor control module, therefore, real-valued gene-coding is applied instead of binary one.

The stop criterion or target is usually set to be the maximum number

of generations reached, or the sliding average fitness value of $\mathfrak{N}$ generations not exceeding a threshold ϵ.

4.4 Numerical Experimentation with LSC and Global Optimal Control

To understand the congestion evolution within a corridor network, numerous experiments have been carried out on both sample networks and real networks of various sizes. Herein one case study is selected to investigate the effectiveness of the proposed LSC schemes. The network is a contrived one to compare LSC with other control strategies so as to better reveal their improvements.

4.4.1 Simple freeway-frontage road corridor: LSC vs. global optimal control

The contrived network keeps all major features of a freeway corridor. For instance, only one direction of the freeway and the arterial on the same direction are retained, and the major connections (link numbered 15, 18 in Figure 4.4) are both ways. A bottleneck link is at downstream of the freeway sections, namely link 6, where the number of lanes drops from three to two.

To ease the investigation, the network segments are classified based on both facility types and traveler groups. They are shown in Table 4.1 as follows.

4.4.2 Demand Scenarios

Efficiency and equity effectiveness of various control strategies can only be evaluated under certain traffic demand loads. We design a series of traffic scenarios to test all three control strategies in concern.

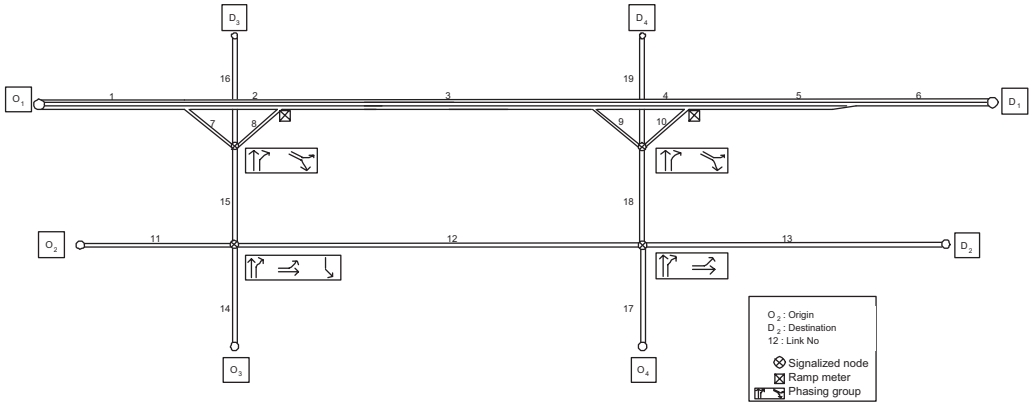


Figure 4.4: LSC Test Network Layout

Table 4.1: Network Segmentation and Division

Facility Classifications	**Links**
Freeway segments	1,2,3,4,5,6
Arterial segments	11,12, 13, 14, 15, 16, 17, 18, 19
Freeway-arterial interface segments (ramps)	7,8,9,10
Traveler groups (traffic demand)	**O-D pairs**
Freeway demand	$O_1 - D_1$
Freeway-arterial demand	$O_1 - D_2, O_1 - D_3, O_1 - D_4, O_2 - D_1, O_3 - D_1, O_4 - D_1$
Arterial demand	$O_2 - D_2, O_2 - D_3, O_2 - D_4, O_3 - D_2 O_3 - D_3 O_3 - D_4 O_4 - D_2 O_4 - D_3 O_4 - D_4$

Table 4.2: Basic Traffic Demand Structure

	Destination			
Origin	**D1 (13)**	**D2 (8)**	**D3 (19)**	**D4 (21)**
O1 (4*)	2000	200	200	300
O2 (20)	400	200	100	400
O3 (24)	300	200	150	350
O4 (25)	750	0	200	150

- Off-peak situation. All of the facilities in the network are loaded under capacity.
- Peak time situation. Some of the facilities, especially the bottleneck section, will be overloaded.
- Incident situation. An incident, which occurs after the network loading starts one hour and lasts half an hour, is created on freeway link 3 between two interchanges. The link capacity is decreased by 60 percent during the half hour to simulate the incident effect.

The peak demand scenario does not change the shape of the demand structure of off-peak scenario. Instead, it increased all the values in Table 4.2 by a "growth factor" of 1.1. The value is selected in that the bottleneck section will be overloaded under traffic peaking period. The incident demand takes the same structure and values as in the off-peak scenario.

4.4.3 Demand Release Patterns

All testing scenarios span four hours so that any phenomena of congestion formation and dissipation can be fully captured. To investigate the impact of traffic fluctuation, three demand release patterns are tested: uniform, triangle and reversed triangle (Figure 6). All OD pairs will take such demand release patterns to test

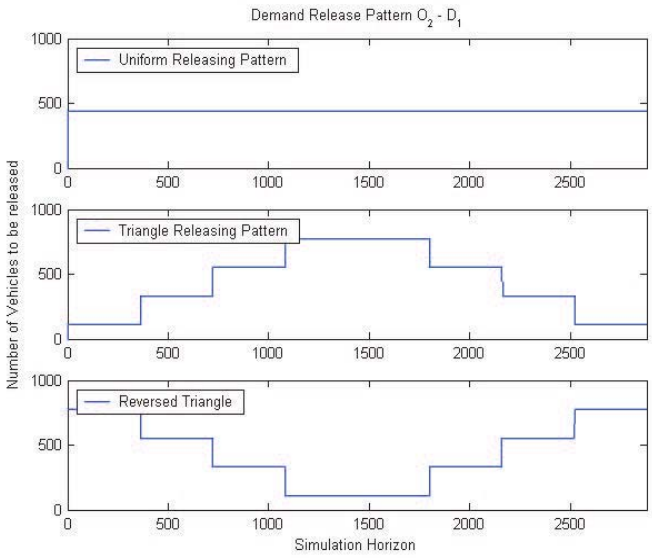


Figure 4.5: Traffic Demand Release Patterns: Uniform, Triangle and Reversed-triangle (O2-D1)

the control strategies. They reflect various traffic load patterns and intensities in the corridor network. For example, a four-hour triangle one may well represent the traffic tiding and ebbing during evening commute peak period. All O-D pairs take the same release pattern as in O2-D1.

4.4.4 Control Strategies

Three control strategies are tested: (1) the isolated adaptive (IA) control, (2) local synchronization control (LSC) and (3) GA-based global optimal control.

Basic settings. The basic control settings include the two signals at the frontage arterial, the signals of the two interchange areas and the two ramp meters. The frontage arterial signals are pre-timed:

- Intersection to the left: greensplits (30, 20, 10), offset 0;

- Intersection to the right: greensplits (28, 16, 16), offset 12 seconds.

The above splits are calculated through Webster's (HCM) method (2.12, 2.15-2.18), and a common cycle 60 is used.

Isolated control settings. The two intersections in the interchange areas are controlled by vehicle actuated controllers, the algorithm is listed in Section 3.4.1, and the minimum and maximum green time for the phases of the two controllers are set to be 4 seconds and 20 seconds.

Both ramp meters take ALINEA (2.34) and both settings are assumed to be the same in Table 4.3. Both ramp meters will then be used to manage the same downstream bottleneck.

Table 4.3: ALINEA parameters for the sample network

Bottleneck Link:	Link #5
Update cycle:	60 seconds
Parameter K_r :	70 veh/hr
Detector Location:	0.15 miles upstream of the bottleneck
Initial Rate:	720 veh/hr

LSC settings. Four LSC units are defined in the network: both the on-ramp (Figure 4.2-a) and off-ramp (Figure 4.2-b) are defined for each of the interchange areas. The following settings apply to both types of LSC units. queue detectors for on-ramp units are positioned 0.18 miles upstream the merge point. Correspondingly, the virtual cycle would be 14.4 seconds. Since a loading interval ϕ_l is universally used, the virtual cycle would be approximated by 7 loading intervals. The intensity is set to be 2, implying that each time the maximum green time of the refrained phases will be decreased by two loading intervals, namely 4 seconds. The two off-ramp units share similar queuing detector and

intensity settings; the maximum green of favored phases will then be increased by 4 seconds when LSC is initiated.

Genetic algorithm settings. As the focus is to examine the effectiveness of different control strategies in alleviating congestion, only system efficiency measures has been considered. In particular, total network travel time (3.36) has been selected as the only fitness function for GA-based optimal control plan computation. The green splits for all phases at all signalized intersections and the ramp metering rates are selected and coded for GA operations (Figure 4.3). The following GA parameters are taken in the gene manipulation process in the algorithm.

Table 4.4: Genetic algorithm parameters for the sample network

chromosome size:	decimal string of 28 digits
Population size:	40
Mutation rate:	0.02
Max generation # :	80
# of sliding average $\mathfrak{N}$:	20
ϵ value:	0.001

4.4.5 Driver's route choice and vehicle routing

Under both off-peak and peak scenarios, all travelers are assumed to take shortest paths that are obvious choices in the network, since travel time of freeway sections is shorter than parallel frontage roads. For instance, instead of

$$\text{Link list: } 1 \rightarrow 7 \rightarrow 15 \rightarrow 12 \rightarrow 13$$

O1-D2 will take the route

$$\text{Link list: } 1 \rightarrow 2 \rightarrow 3 \rightarrow 9 \rightarrow 18 \rightarrow 13$$

In the incident scenario, however, a pre-determined portion of traffic diversion is assumed, because the incident would affect three O-D pairs, O1-D1, O2-D4, and O3-D4. Ten percent of the demand from these three O-D pairs is diverted off their original routes of taking freeway to take the parallel arterial link 12. For instance, the remaining majority of the freeway trips (O1-D1) (90%) will continue to take the freeway route

$$\text{Link list: } 1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6$$

and the rest ten percent will take the alternative detour route

$$\text{Link list: } 1 \rightarrow 15 \rightarrow 12 \rightarrow 18 \rightarrow 10 \rightarrow 5 \rightarrow 6$$

4.4.6 Numerical results

The system efficiency and user equity measures of the 27 scenarios are summarized in Table 4.5 and Table 4.6. Since most existing corridor controls are isolated ones that do not integrate signal controllers and ramp meters, the nominal trip cost is defined as the cost under isolated control. This treatment simplifies the comparison among these strategies.

Figure 4.6 showed the convergence of fitness value (FV) function for a few scenarios. All processes converge to the stable (near-global) solution within $1,600$ DNL evaluations. Each scenario took about 40 minutes to converge on a Pentium 3.2G CPU with 1G RAM.

Comparing corresponding cells in Table 4.5 and Table 4.6, we can find that there is a clear trend of total travel time decrease from IA, LS to GO; however, there is no equally clear trend reversed for Gini Coefficient. Under various controls and different demand scenarios, the aggregate equity vary among IA, LS and GO. Expectedly, GA-based global optimal has the best efficiency measures, but it does

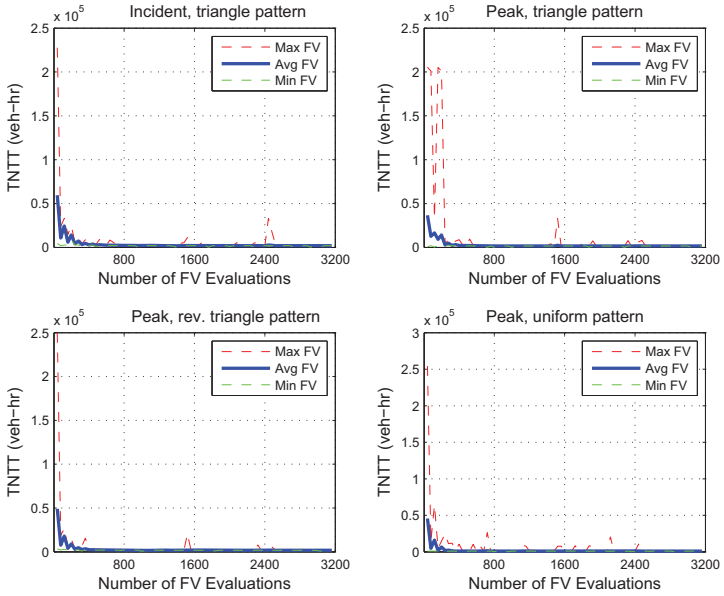


Figure 4.6: GA Based Corridor Control Plan Computation Convergence: Sample Network under Various Conditions

not always performs worst in terms of aggregate equity. In fact its aggregate equity performance is better than IA in most testing cases. On the other hand, LS performs either better than or at least similar to IA and GO with only few exceptions(peak traffic, uniform loading). The equity implication of the control rule of queue manipulation is translated into the Gini Index measure.

The Lorenz Curves of travelers' delay under different control strategies are drawn for the scenario of peak traffic of uniform load pattern (Figure 4.7).

In the disaggregate equity analysis, the most disadvantaged travelers' groups are the surface streets traffic. In Table 4.6, critical trip cost ratios show that the major direction of the corridor network traffic, i.e., the freeway mainline and the parallel arterial, mostly better off. However, the cross street traffic is the most dis-

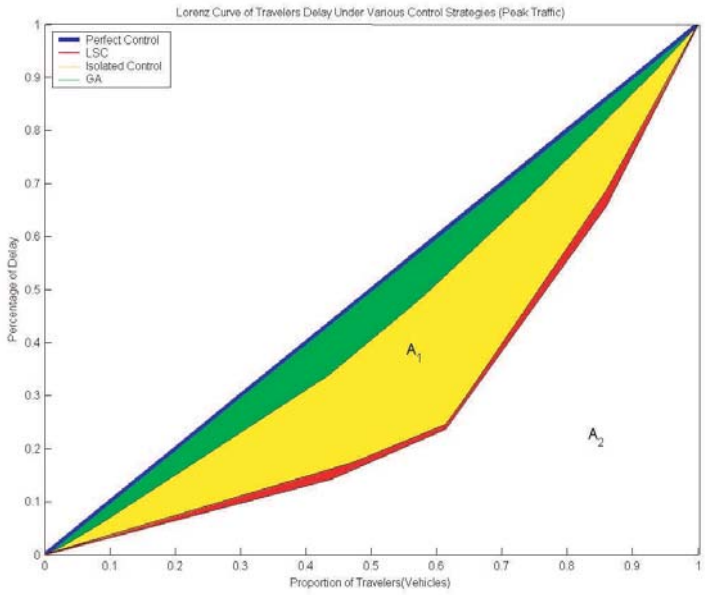


Figure 4.7: Lorenz Curve of Various Control Strategies

Table 4.5: System Efficiency (TTT) Comparison Under Various Demand Scenarios

	Control Strategies		
Demand scenarios	IA	LS	GO
Off-peak Traffic	893.9 (U) 1876.3 (T) 1188.4 (R)	892.4 (U) 1674.5 (T) 1174.0 (R)	874.3 (U) 1613.6 (T) 1108.9 (R)
Peak Traffic	985.9 (U) 4192.3 (T) 2135.5 (R)	985.4 (U) 3609.4 (T) 1869.0 (R)	966.0 (U) 3579.2 (T) 1809.9 (R)
Incident	946.9 (U) 3845.4 (T) 1535.7 (R)	946.0 (U) 1932.3 (T) 1208.0 (R)	912.3 (U) 1860.9 (T) 1181.5 (R)

advantaged from LS and GO under various testing scenarios. Under LS and GO, over half of the scenarios (10 out of 18) have the most disadvantaged traveler group of the cross street traffic O4-D4.

Generally, global optimal control has the worst disaggregate equity measure. Under global optimal control, the trip travel time of some traveler groups are even quadrupled than in isolated control. This implies that global optimal control reaches the optimal in the price that some travelers are severely penalized. It also indicates the importance of incorporating disaggregate equity measures in corridor control system design.

It is also noted that various loading patterns can lead to drastic change of network efficiency measures even if the total demand remains the same. In the case of off-peak scenarios, For example, the total travel time under a triangle loading is twice as much as that of uniform loading. Adaptive controls like the proposed LSC schemes are much needed in such situations.

Table 4.6: Control Equity Comparison Under Various Demand Scenarios Control Strategies

	Control Strategies		
	IA	LS	GO
Gini Coefficient			
Off-peak Traffic	0.24 (U)	0.24 (U)	0.17 (U)
	0.28 (T)	0.21 (T)	0.27 (T)
	0.24 (R)	0.21 (R)	0.23 (R)
Peak Traffic	0.23(U)	0.26 (U)	0.17 (U)
	0.34 (T)	0.29(T)	0.37(T)
	0.30 (R)	0.22 (R)	0.28 (R)
Incident (Reroute)	0.39 (U)	0.18 (U)	0.24 (U)
	0.58 (T)	0.26 (T)	0.22 (T)
	0.39 (R)	0.19 (R)	0.21 (R)
Critical Trip Cost ratio			
Off-peak Traffic	1	1.01 [O3-D1]	1.05 [O3-D4]
	1	1.81 [O4-D4]	2.22 [O4-D4]
	1	1.25 [O4-D4]	2.22 [O4-D4]
Peak Traffic	1	1.01 [O1-D4]	1.05 [O4-D4]
	1	3.3 [O4-D2]	3.7 [O4-D2]
	1	2.1 [O4-D4]	2.3 [O4-D4]
Incident (Reroute)	1	1.02 [O1-D4]	0.97 [O4-D4]
	1	3.5 [O4-D4]	3.7 [O4-D4]
	1	2.0 [O4-D4]	1.8 [O4-D4]

4.4.7 Concluding Remarks

Both integrated control strategies of rule-based local synchronization control (LSC) schemes and GA-based global optimal control perform better than most prevalent isolated adaptive control strategies. It clearly shows that integration of control measures can improve the corridor control operations.

The proposed LSC schemes perform satisfactorily in most scenarios. Under incident scenarios where the congested traffic is diverted to alternative routes, the LSC schemes can prevent the congestion from spreading over the network and thus

improves the system efficiency considerably. Surprisingly LSC is not inferior to the GA-based global optimal control plans in terms of both system efficiency and user equity in many cases.

The experimentations also indicate that at least system efficiency and user equity measures are not all competing dimensions: better system efficiency does not imply worse the equity measures. This urges us to formulate the integrated corridor control problem in a holistic manner, where balancing both the system efficiency and user equity measures must be considered at the same time.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 5

Efficiency-Equity Solution Framework to Integrated Corridor Control Problem

This chapter develops a solution framework to solve the integrated corridor control(ICC) problem so as to balance both the system efficiency and the user equity of the corridor network.

5.1 An Integrated Corridor Control Design Framework

For either off-line planning or online operating purposes, the dynamic traffic management systems (DTMS) need both components of traffic control and traffic assignment. One simplified version of DTMS can be illustrated in the flowchart of Figure 5.1. DTMS will compute and publicize the control plans based on surveillance and prediction of the network users' behavior. The traffic management center collects real-time traffic information and/or synthesize the historical information for

both control design and traffic prediction. On the one hand, the traffic control loop computes the control plan and the control plan is transmitted to the control devices for implementation. On the other hand, the network users adjust their travel decisions based on the observed control settings.

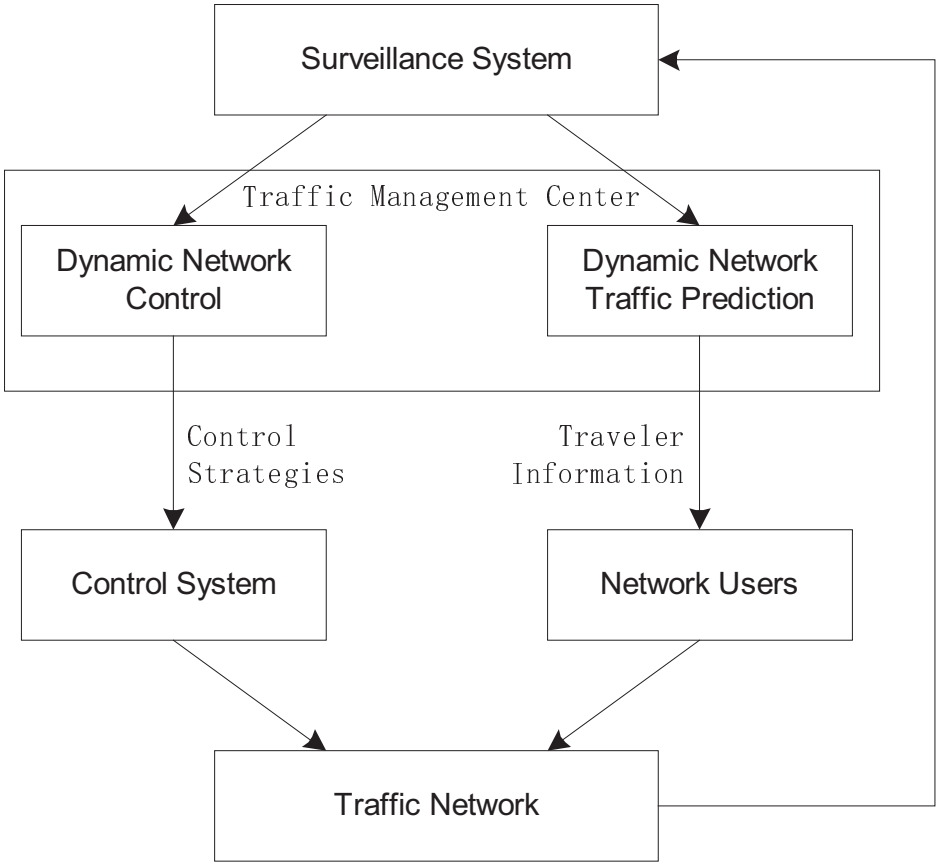


Figure 5.1: An Integrated Dynamic Traffic Management System (DTMS) (*Adapted from Chen 1998*)

The abstract framework in Figure 5.1 can be adapted for both off-line planning and online operation applications. For example, an off-line application can

translate the framework into a modulized version in Figure 5.2.

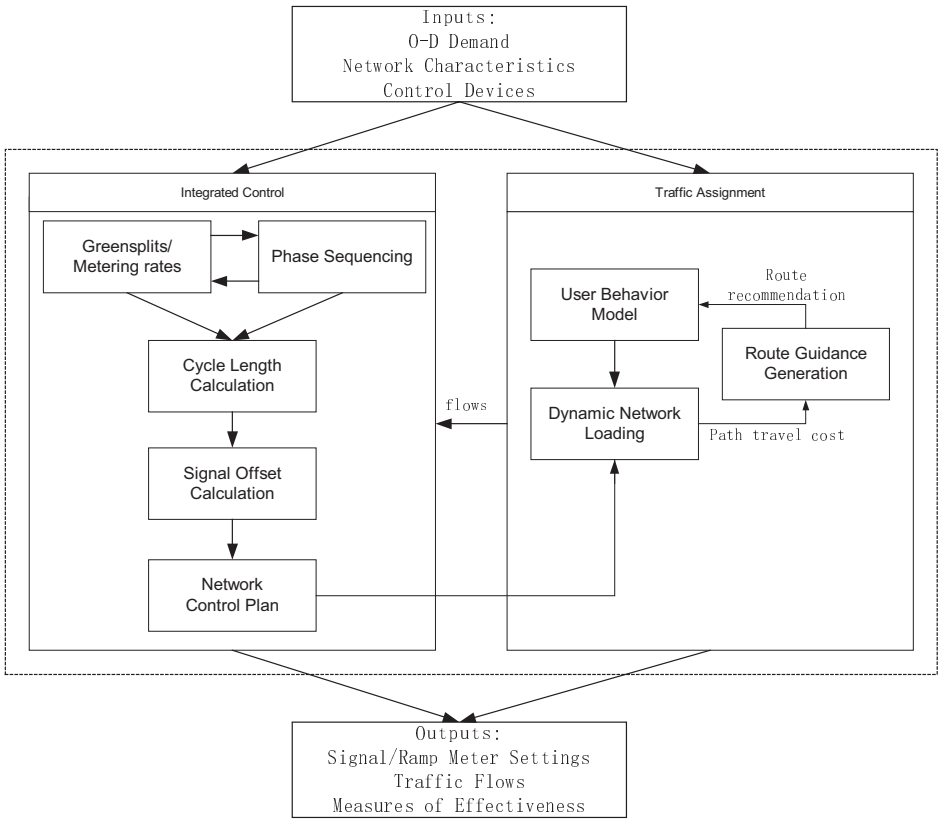


Figure 5.2: An Off-line Corridor Traffic Management System

This system in Figure 5.2 is designed for two purposes. The first one is to directly compute the time-of-day control plan for any off-line control programs as in the literature review; and the other one is to dynamically allocate the capacity for conflicting flows for planning purposes. As an off-line system, the traffic demand is assumed available for the entire analysis horizon. On the contrary, the traffic demand input for an online application will have to be estimated in a real-time manner, and together with the control plan, must be cast into the *rolling horizon* scheme for

periodic update. In this study, we focus on the off-line application systems. The development of the proposed efficiency-equity solution to the ICC problem will be based on the off-line solution framework as in Figure 5.2.

Regarding the efficiency-equity control in this study, the program requires better differentiation of the network users, and it has to consider their relative benefits and losses from certain control strategies. We start building the framework from development of the bi-criterion control objective functions in the next section.

5.2 Balance Efficiency and Equity in Control Objectives

Development of balanced bi-criterion control objectives closely follow the findings in the previous chapters. In Chapter 3, individual efficiency and equity measures have been presented based the DNL model (Section (3.6). The comparison study in Chapter 4 showed that neither efficiency measure nor equity measure only can fully capture the effectiveness of any control system. The results also indicated that aggregate and disaggregate equity measures are not supplementary equity measures; rather they must be present simultaneously so that the fairness among network travelers is not compromised or distorted.

In Chapter 1, our version of efficient and equitable system has been defined. According to this definition, mean difference (MD) and the critical trip cost ratio (CR) are selected to represent aggregate and disaggregate equity respectively. Since Gini Coefficient has been widely used as an aggregate equity measure with clear meaning, the measure will also be computed to supplement the analysis. We note here that based on the needs specific to every situation, the definitions of both efficiency and equity measures, equity ones in particular, can vary. Consequently, various efficiency and equity measures can be be selected.

From the definition in Chapter 1, the system with the minimal TTT will be

considered efficient whereas the system with both minimal MD and minimal CR will be considered equitable. Therefore, with the same set of constraints, optimization of any of the above measurements will lead to one corridor system that addresses one aspect of its performance. For example, if TTT is minimized, the system will be an efficient one, whereas minimization of MD leads to an equitable system at aggregate level. A balanced efficient-equitable system would then require that TTT, MD and CR be combined into a single objective. Thanks to the fact that MD also has the same units as TTT, the following objective may be defined:

$$TTT + W_M MD \tag{5.1}$$

where W_M is the weight for MD.

However, consider that efficiency measure TTT and equity measure MD cannot be known a priori when choosing the weight W_M, the above linearized combination of efficiency and equity is not always appropriate. Regarding this deficiency, a metric distance measure is proposed to balance all three measures equally:

$$\alpha = \left(\left(\frac{TTT - TTT_{min}}{TTT_{max} - TTT_{min}} \right)^{\beta} + \left(\frac{CR - CR_{min}}{CR_{max} - CR_{min}} \right)^{\beta} + \left(\frac{MD - MD_{min}}{MD_{max} - MD_{min}} \right)^{\beta} \right)^{\frac{1}{\beta}} \tag{5.2}$$

where β is the power of the metric distance α that measures the effectiveness of the control system in balancing the three measures. Minimizing the metric distance α is to balance both the efficiency measure (TTT) and the equity measure (MD and CR). We apply $\beta = 2$ in this study. A series of the metric measures α will form the Pareto efficient frontier as discussed in Chapter 1.

5.3 Traffic Assignment and Nominal Travel Cost

The problem of mapping the time-dependent demand onto the network to obtain the network flow pattern falls in the category of dynamic traffic assignment (DTA). As decomposed in Figure 5.2, the structure of the DTA problem contains the following components:

- User Behavior Model
- Route Guidance Generation
- Dynamic Network Loading

The user behavior model takes the time-dependent O-D and route guidance information as input and assigns the trip rates to a set of paths available to each O-D trip. The set of time-dependent paths (3.32) will feed into the dynamic network loading model so as to obtain the time-dependent link flows and link/path travel costs as calculated in (3.28-3.32).

Herein one time-dependent shortest path calculation method (Chabini 1998) is used to update the shortest paths by relying on an space-time expanded network (STEN)(Nie, Nie & Zhang 2004). The STEN is the expansion of the static node-link network by the time dimension where the time-dependent link travel costs obtained from dynamic network loading is added. Chabini's algorithm scanns all nodes in STEN in a reversed chronological order, so it is often called the decreasing order of time algorithm. Because of its reduced computation complexity, this algorithm is chosen to compute the time-dependent shortest path for loading the path flows.

The well accepted user behavioral models are the user equilibrium assumptions, where all travelers will choose their paths based on full knowledge of the network traffic evolution(Wardrop 1952)(Friesz, D.Bernstein, Smith & Wie 1993)(Peeta

& Jayakrishnan 2001). the resulting flow pattern is one that no traveler can benefit from unilaterally changing the route. Mathematically solid and physically close to reality as these models are, the assumptions are usually considered strong in the real world systems. Therefore, more realistic stochastic dynamic route choice behavior models have been developed(Cascetta et al. 1996) and more complex multi-user class dynamic traffic assignment models are used(Chen 1998). Along this line, the integrated control-assignment problem will naturally take the bi-level structure and the solution will be the *mutual consistent* points between dynamic traffic control (DTC) and dynamic traffic assignment, as the past research in the review(Chen 1998).

As the focus of this study is to investigate the system efficiency and user equity of the control strategies, an efficiency-equity balanced control strategy will exist with any network flow pattern from a reasonable user behavior model. In this sense, we simplify the bi-level structure to a single level one where only one iteration of the assignment process will be used, namely the all-or-nothing assignment based on the updated time-dependent shortest path.

Correspondingly, the free flow travel cost will be taken as the nominal path travel cost in calculation of the relative path travel cost (Equation 3.37 in Section 3.5.1).

5.4 Computation of Dynamic Traffic Control Plan

An off-line control design system maintains the cycle length and phase structure as most systems do (Robertson 1969*b*)(Hunt, Robertson, Bretherton & Winton 1981). The dynamic control plan computation follows the two level structure (Figure 5.2). At the first level, the green splits and metering rates will be computed based on the information obtained from dynamic network loading. The information includes the time-dependent link flows as well as the traffic queues. Since the green

splits at urban signals are tied with corresponding phases and phase sequencing, the phase selection and phase sequencing will be computed iteratively with green splits and metering rates.

After the first level of optimization is finished, the cycle length will be adjusted for each intersection so that the network or the predefined sub-regions will have the same cycle length. Afterwards the offset for each signal controller at the intersection will be computed to provide the best progression through the network. In the next sections, the conceptual design is provided and the concrete formulation and solution algorithms will be developed in the next chapter.

5.4.1 Green splits and Metering Rates

Green splits of the signal phases and metering rates for ramp meters are the most important control variables within the control system, since the green splits and metering rates allocate the right-of-way to the conflicting traffic flows and thus determine the link and movement capacities. Computation of the green splits and metering rates is formulated as a mathematical program with the following structure:

Minimize *TTT, MD, CR* or α

s. t.

Traffic demand input (3.26);

Dynamic link flow pattern from assignment and DNL (3.1-3.25);

Phase and phase sequencing constraint (6.15);

Cycle length (6.12);

Min and max green/metering rate constraint (6.13-6.14).

Similar programs are also built for the subsequent phase selection and offset optimization processes.

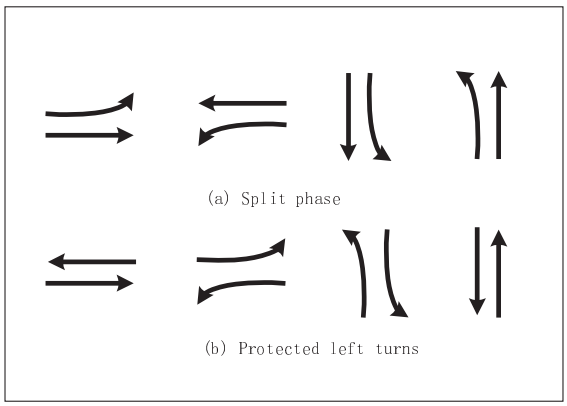


Figure 5.3: Phase Diagram at An Intersection

5.4.2 Phases and Phase Sequencing

A phase is the group of compatible traffic movements at an intersection controlled by the signal groups that have the same green-amber-red sequences and durations. Phase sequences affect the waiting time of various movements, left turning and the opposing through traffic in particular, phase sequencing is also one set of variables to be determined. Note that *permissive* or *protected* left turns have different flow updating rules, and the permissive left turns would follow the priority control rules specified in Section 3.3.3. For it is not controllable through signals, we herein only consider protected left turns in the program.

Previous studies have developed phase grouping rules for complicated intersections based on enumeration techniques(Stoffers 2003). Since only vehicular traffic is treated in this study, the phase library is composed as Figure 5.3 based on the NEMA phasing specifications. The sequencing order will then become a combinatorial optimization problem and this also follows the program structure specified for optimizing green splits and metering rates. Solving for the best phasing diagram and phase sequencing is also performed based on programs as conceptualized previously.

5.4.3 Signal Coordination and Cycle Length

From developing the dynamic network loading model, we know that offset is an important control variable that provides the coordination between signal controllers when properly set. Signal coordination can be provided within either pre-defined subregions or the entire corridor network itself. The coordinated signal controllers usually require the same common cycle length or half (double) of the common cycle length. After the first level of phasing and greensplits/metering rates optimization is done, the best set of green splits and metering rates will be kept and the common cycle lengths for sub-regions or for the network will be scaled and adjusted in proportional to its original splits, based on the longest cycle length required by individual intersections. Afterwards the offset will become the decision variable in a similar program as above until the best system performance is reached.

Chapter 6

Heuristic Solution Algorithm: Simultaneous Perturbation Stochastic Approximation

6.1 Theoretical Basis and Critique of Heuristic Searching Algorithms

Solving complex optimization problems that have no available gradient information generally resorts to heuristic searching methods. Commonly used methods include genetic algorithm, simulated annealing, artificial neural network, finite approximation among other variations of the above methods. Behind these methods are the theories of stochastic searching. The most fundamental one comes from a simple rule that states the following:

Theorem 6.1 *Suppose that θ^* is the unique minimizer of L on the domain Θ in the*

sense that $L(\theta^*) = inf_{\theta \in \times} L(\theta), L(\theta^*) > -\infty$, *and*

$$\inf_{\theta \in \Theta, \|\theta - \theta^*\| \geq \eta} L(\theta) > L(\theta^*) \tag{6.1}$$

for all $\eta > 0$, *Suppose further that for any* $\eta > 0$ *and for all* k, *there exists a function* $\delta(\eta) > 0$ *such that*

$$P[\theta_{new}(k) : L(\theta_new(k)) < L(\theta^*) + \eta] \geq \delta(\eta). \tag{6.2}$$

Then, for algorithm A with noise-free loss measurements, $\hat{\theta}_k \to \theta^*$ *a.s. as* $k \to \infty$.

A brief proof can be found in (Spall 2003). In this sense, all methods that can probabilistically span the entire search space will be able to find the optimal solution, and therefore more algorithms can be designed to treat special types of problems including genetic algorithm, which has been used to compute the globally optimal efficient corridor control plan in Chapter 4.

As recalled from the literature review, mathematical programming methods that solve the control problem usually require the traffic flow models to be simplified so that the gradient information can be computed. Such simplification often compromises the underlying traffic flow models. On the other hand, heuristic optimization methods such as the genetic algorithm can search for a near-global optimal control plan, allowing more realistic representation of traffic flow. However, heuristic methods usually need a large number of evaluations of system performance and usually lead to high computational costs. Therefore, it is desirable for a method to preserve the modeling power of more advanced underlying traffic flow models while being able to compute the control plan in a computationally efficient manner. We explore a stochastic approximation technique that can be viewed as a compromise of the above two types of approaches. The method is called simultaneous perturbation stochastic approximation (SPSA), which has been used in other fields (Spall 1998) and shown

satisfying performances. In the next section, we will introduce the general application procedure of the method and later develop one heuristic optimization algorithm to solve the integrated corridor control problem.

6.2 Simultaneous Perturbation Stochastic Approximation

For a general SPSA procedure, the general objective function $L(\theta)$ as specified by (3.36, 3.45, or 3.46) in Chapter 3 is a scalar-valued performance measure of the system, and the θ is a continuous-valued $p-$dimensional vector of parameters, i.e., the vector of control variable θ (vector 1.1) in the corridor control context. It could happen that noises ϵ occur[1] when measuring the system performance measure $z(\theta)$:

$$z(\theta) = L(\theta) + \epsilon \tag{6.3}$$

The SPSA method starts from an initial guess of θ (one feasible solution) and by a sequential "simultaneous perturbation" over the solution trajectory, the approximation of the gradient $\varphi(\theta) \equiv \frac{\partial L(\theta)}{\partial \theta}$ will converge to zero, under several regularity conditions (section IV) over the sequence.

Assuming that $L(\theta)$ is differentiable over θ and the minimum is obtained at a zero point of the gradient, i.e.,

$$\varphi(\theta) = \left| \frac{\partial L(\theta)}{\partial \theta} \right|_{\theta=\theta^*} = 0 \tag{6.4}$$

The recursive updating of θ takes the standard form:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{\varphi}(\hat{\theta}_k) \tag{6.5}$$

[1] As a matter of fact, The SPSA method performs superior when system performance evaluation is contaminated by observation or measurement noises.

where the gain sequence $\{a_k\}$ must satisfy certain regularity conditions.

The perturbation is performed upon evaluating $\varphi(\hat{\theta}_k)$. First define a $p-$dimensional mutually independent mean-zero random variable vector $\Delta_k \in R^p$ $\{\Delta_{k1}, \cdots, \Delta_{kp}\}$ satisfying certain conditions the most important of which is that $E(|\Delta_{ki}^{-1}|)$ is above bounded by some constant α_1, $E(|\Delta_{ki}^{-1}|) \leq \alpha_1$. An optimal distribution of Δ_k is symmetric Bernouli (Sadegh & Spall 1998), i.e., $P(\Delta_{ki} = \pm 1) = \frac{1}{2}$. Furthermore, $\{\Delta_k\}$ is a mutually independent sequence independent of $\hat{\theta}_0, \hat{\theta}_1, \cdots, \hat{\theta}_k$. Let

$$z_k^{(+)}(\theta_k) = L(\hat{\theta}_k + c_k\Delta_k) + \epsilon_k^{(+)} \tag{6.6}$$

$$z_k^{(-)}(\theta_k) = L(\hat{\theta}_k - c_k\Delta_k) + \epsilon_k^{(-)} \tag{6.7}$$

where c_k is a positive scalar satisfying the regularity conditions, and $z_k^{(+)}(\theta_k), z_k^{(-)}(\theta_k)$ are the measurements of the system under the perturbation $\hat{\theta_k} + c_k\Delta_k$ and $\hat{\theta_k} - c_k\Delta_k$, respectively.

And the approximation of the gradient will be:

$$\hat{\varphi}_k(\hat{\theta}_k) = \frac{z^{(+)_k} - z_k^{(-)}}{2c_k} \begin{bmatrix} \Delta_{ki}^{-1} \\ \vdots \\ \Delta_{kp}^{-1} \end{bmatrix} \tag{6.8}$$

Spall (Spall 1992) shows that by recursively updating θ_k, the gradient will converge to a zero point, which implies that $L(\theta)$ can reach at least a local minimum. He also argues that it is unlikely that the approximation will settle down at a maximum or a saddle point because of the stochastic nature of the algorithm. The regularity conditions are introduced in the following section.

6.3 Regularity Conditions assuring Convergence

Five assumptions are made upon the gain sequence a_k when forcing θ_k to converge almost surely to θ^*:

A1: $a_k, c_k > 0 \forall k; a_k \to 0, c_k \to 0$ as $k \to \infty$;

$\sum_{k=0}^{\infty} a_k = \infty; \sum_{k=0}^{\infty} (\frac{a_k}{c_k})^2 = 0$

A2: For some $\alpha_0, \alpha_1, \alpha_2 > 0$ and $\forall k, E\epsilon^{(\pm)^2} \leq \alpha_0, EL(\hat{\theta} \pm \bar{\Delta_k})^2 \leq \alpha_1$, and $E\Delta_{kl}^{-2} \leq \alpha_2, l = 1, \cdots, p$.

A3: $||\hat{\theta_k}|| < \infty, \forall k$;

A4: θ^* is an asymptotically stable solution of the differential equation $dx(t)/dt = -g(t)$.

A5: Let $D(\theta^*) = \{x_0 : \lim_{t\to\infty} x(t|t_0) = \theta^*\}$ where $x(t|x_0)$ denotes the solution to the differential equation of A4 based on initial conditions x_0, there exists a compact set $S \subseteq D(\theta^*)$ such that $\hat{(}\theta_k) \in S$ infinitely often for almost all sample points.

The gain sequences of a_k and c_k generally take the power function:

$$a_k = \frac{a}{(1 + A + k)^{\alpha}}, \qquad c_k = \frac{1}{(1 + k)^{\gamma}} \tag{6.9}$$

It is argued (Chin 1997) that the asymptotically optimal values of α, γ are 1 and $\frac{1}{6}$, respectively. But Spall pointed out that $\alpha < 1.0$, usually produces better finite-sample performance(Spall 1998). Hence another set of values of 0.602 and 0.101 that are the lowest allowable to satisfy the regularity conditions above (A1-A5) are recommended.

It is observed that for most engineering problems these conditions are almost automatically satisfied with only **A3** being hard to verify for a general case (Spall 1992). In the corridor control problem, it physically implies that the transportation system leads to gridlock completely. As this could be avoided by placing the practical constraints over the control(6.12-6.14), it does not impose any difficulty in the solution as indicated in the numerical example.

6.3.1 Constrained Optimization via Stochastic Approximation

The SPSA procedure introduced above is suitable for unconstrained optimization. While most engineering problems are constrained by physical feasibility set, the optimal corridor control problem is no exception. Sadegh proposed a projection method to restrict $\theta_k \in \mathcal{R}^p$ at each iteration k to fall in the feasibility range of the control variables by simply replacing any violating $\hat{\theta_k}$ with the nearest point $\theta_k \in G(\theta)$ where $G(\theta)$ is the feasibility set of the control vector:

$$\hat{\theta_{k+1}} = P(\theta_k - a_k \hat{g}_k(\hat{\theta}_k)) \tag{6.10}$$

And the perturbed vectors $\hat{\theta_k} + c_k\Delta_k$ and $\hat{\theta_k} + c_k\Delta_k$ in evaluating of the loss function (6.6 and 6.7) will also be projected such that these two perturbed vectors must lie in the feasibility range. By forcing another restriction (6.2) over the constraints, SPSA can still converge to a Karash-Kuhn-Tucker point almost surely (Proposition 1 in (Sadegh 1997)).

Assumption 6.2 *The set* $G = \{\theta : q_i(\theta) \leq 0, i = 1, ..., s\}$ *is non-empty and bounded, and the functions* $q_i(\theta), i = 1, ..., s$, *are continuously differentiable. At each* $\theta \in col(G)$ *where col denotes the boundary, the gradients of the active constraints are linearly independent. Furthermore, there exists an* $\epsilon < 0$ *such that the set* $G^- = \{\theta : q_i(\theta) \leq r, i = 1, ..., s\}$ *is non-empty for* $\epsilon \leq r < 0$ *(i.e. the set* G *has a non-empty interior.)*

Spall (1998) observes that the gain factors a_k and c_k are key to the decaying process of the stochastic gradient approximation and the selection of the two scalars is critical to the success of searching the optima. However, because the parameter vector θ may have various numerical magnitudes, for example, the ramp metering rate $\mathbf{R}$ of hundreds of vehicles per hour and the green duration g of seconds, they have to be synchronized during the decaying process. A normalization process is thus

necessary for the decaying process. The following normalization process is applied in this study.

$$g_i^n = \frac{g_i - g_{i,min}}{g_{i,max} - g_{i,min}} \tag{6.11}$$

where g_i^n can be any control variable with the physical constraints of (6.12-6.14). Then One needs to examine whether the normalization process would affect the convergence performance of SPSA process. The following lemma checks the regularity conditions and shows that the network control problem formulated as in the previous section can be solved via the SPSA procedure.

Proposition 6.3 *A normalized version of the projection method in constrained SPSA can assure a convergence to at least a local optimum a.s.*

Proof It is trivial to verify the non-emptiness and of the control feasibility set $G(\theta)$ since any points that fall in the box constraints (6.12-6.14) will fulfill the conditions. Since all constraints including the box constraints and sum constraint (6.15) are all linear, the following equation holds:

$$\partial \frac{q_i(\theta)}{\partial \theta_j} = 1 \text{ or } -1, i = 1, ..., s, j = 1, ..., q$$

As $g_{i,max}, g_{i,min}$ are constants, the linear transformation (6.11) does not change the above argument; then Assumption 1 for the control feasibility set after the linear transformation still holds.

With the assumptions A1-A5 and the above verification of Assumption 1, we conclude that after the linear transformation (6.11) as $k \to \infty$ with $\hat{g}_k(\hat{\theta_k}) = \hat{g}_k^{SP}(P_k(\hat{\theta_k}))$,

$$\hat{\theta}_k \to KT \quad a.s.$$

End of proof.

6.3.2 The Constraints on Control Variables in Flow Models

In practice, traffic controls usually enforce some physical constraints including the maximum and minimum duration of the cycle length and green duration for any phase, and the max/min metering rates:

$$C_{i,min} \leq C_i \leq C_{i,max} \tag{6.12}$$

$$g_{i,min} \leq g_i \leq g_{i,max} \tag{6.13}$$

$$R_{i,min} \leq R_i \leq R_{i,max} \tag{6.14}$$

In this study, time-of-day traffic control plan is studied and the cycle length as well as a fixed phasing order is thus maintained for signalized intersections. Also effective green times are considered in this study, so the cycle length constraint for any intersection will be:

$$\sum_{p=1}^{P} g_p^j = C^j - PL \tag{6.15}$$

It tells that the sum of the effective green duration of the phases $p = 1...P$ at intersection j has to be equal to the available green time $C^j - PL$, i.e., the cycle length deducted by the loss time of all phases.

6.4 Solution Algorithm

The iterative SPSA solution algorithm to solve the offline network control optimization problem is as follows.

SPSA Algorithm for Integrated Network Traffic Control

Step 1: *Initialization and Coefficient Selection.*

1.0 Set iterator $k = 0$.

1.1 Generate the control vector and normalize it via (6.11) as θ^N.

1.2 Pick an initial feasible solution of θ_0^N,

1.3 Select nonnegative coefficients $a, c, A, \alpha,$ *and* γ.

Step 2: *Simultaneous Perturbation.*

Generate a $p-$*dimensional random perturbation vector* Δ_k, *where each component is mutually independent Bernoulli* ± 1 *distributed with probability of* $\frac{1}{2}$ *for each* ± 1 *outcome.*

Step 3: *Loss Function Evaluation by Dynamic Network Loading.*

3.1 Perturb the normalized control vector with $\hat{\theta}_k^N \pm c_k \Delta_k$;

3.2 Project the perturbed control vectors onto $G^N(\theta_k)$ *from (6.10);*

3.3 Transform the projected control vector back to the real valued control variables;

3.4 Evaluate the system performances by loading the demand onto the network under both set of control variables and obtain (6.6- 6.7);

Step 4: *Gradient Approximation.*

Calculate the approximated gradient from (6.8).

Step 5: *Control Update.*

Update $\hat{\theta}_k$ *with (6.5).*

Step 6: *convergence criteria check.*

Check if the maximum iteration number has achieved or is met. If yes, stop. If not, set $k = k + 1$ *and go to step 2.*

Some notes on the algorithm: In step 1.2, the initial solution can be obtained from either setting $\theta_0^N = 0(1)$, which corresponding to minimum (maximum) controls or using other calculation methods, e.g., HCM or Webster's. As about selecting the coefficients, according to Spall (Spall 1988), both sets are fine: either a practically effective and theoretically valid set of $\alpha = 0.602, \gamma = 0.101$ or a asymptotically optimal values of $\alpha = 1, \gamma = \frac{1}{6}$ can be used.

6.5 Numerical Experimentations with SPSA for Efficient System Control

6.5.1 A Simple Network to Investigate the Convergence Performance

When we further analyze the above SPSA method using numerical examples, two issues are of particular interest, its ability to reach a optimal solution, and the efficiency of obtaining the optimal solution compared to other methods. Two network examples are provided to examine both issues separately.

To examine the algorithm's ability to converge to a stable optimal solution, a typical diamond interchange area is selected within which one ramp meter and one intersection exist to control the flow (Figure 6.1). The freeway mainline is from 1 to 2, and the crossing surface street is both ways. The phasing diagram is also shown in the figure; the right turning flow at the intersection, i.e., the off-ramp flow onto the surface street (in the direction of 4 to 3), has the right-of-way all the time since it is compatible with any of the movements in the two phases.

The demand is given in the following table. For simplification, only fifteen minutes of demand are made; to simulate the time variations, the traffic demand is loaded in a isosceles triangular form in three-minute intervals. During this period,

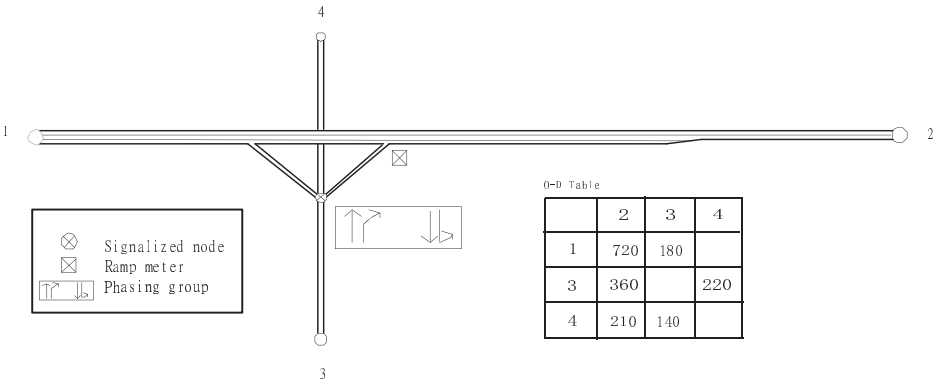


Figure 6.1: SPSA Test Network Layout

Table 6.1: Trip Rates Table for Sample Network I

	2	3	4
1	720	180	
3	360		220
4	210	140	

the metering rate at the on-ramp is simply set to be fixed. One may notice that only two independent control variables are present in the sample network, corresponding to the normalization procedure in (6.11), namely the green ratio g_1 for phase 1 (the green ratio for phase 2 will be the $1 - g_1$ if we omit the loss time for the time being) and the metering rate R. The maximum/minimum green time is set to be 50 and 10 seconds respectively, and the range for the metering rates is set to be 300 - 1,500 vph. To locate the possible global optimal control plan, an exhaustive search through the feasibility solution space is performed. In this example, the network loading interval t is one second; thus the increment of the phase duration is set equal to t, while an increment of 20 vph is selected to scan the range of metering rates. Therefore, the exhaustive search goes through a total number of 2,400 (40×60) $g_i - R$

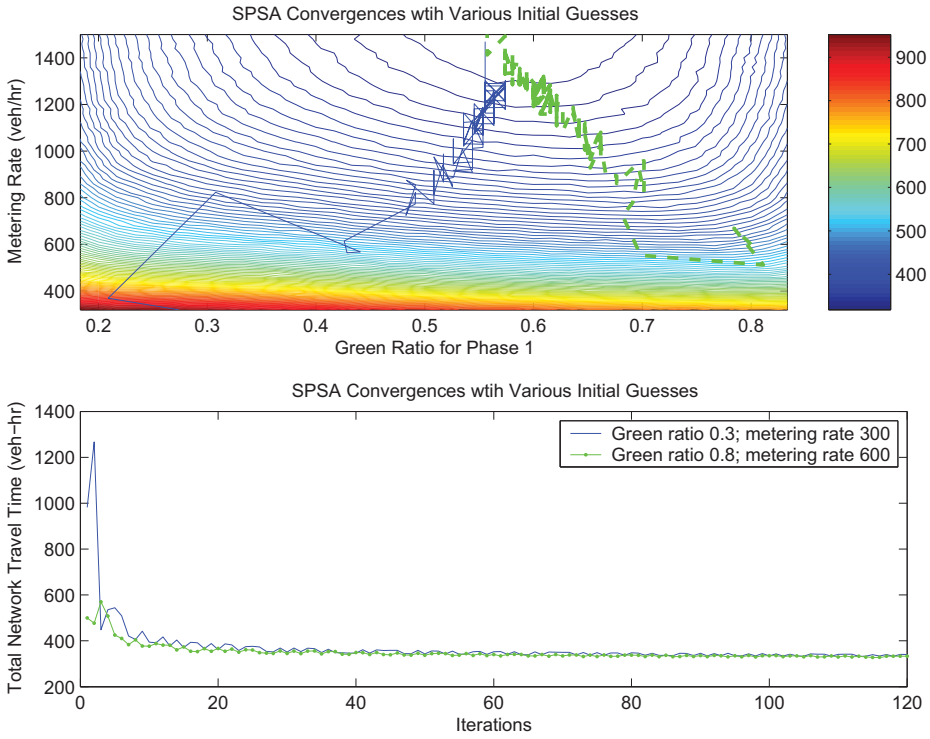


Figure 6.2: SPSA Convergence under Various Initial Feasible Solutions

combinations, and the contour of the objective function (total travel time) under various combinations is plotted in Figure 6.2a. The contour implies only one global optimal solution exist in the search space; and it is reached at $(g_1, R) = (0.61, 1500)$, with a total travel time of 312 veh-hr. In the example, the optimal metering rate is the upper bound of the feasible range, that is, allowing as many flows as possible into the freeway mainline during this 15-minute period. Two SPSA processes with different initial feasible "guesses" (θ_0) are experimented and plotted in Figure 6.2. The first starts with $(g_1, R) = (0.30, 300)$ and stabilizes itself at $(g_1, R) = (0.58, 1473)$; the second starts with $(g_1, R) = (0.80, 600)$ and stabilizes itself at$(g_1, R) = (0.57, 1498)$.

Generally, both processes can reach the near-global optimal solution along a different path. Figure 6.2-b shows that the convergence process of SPSA has a feature of quick drop at the first few iterations; after less than forty iterations in the example, both processes reach the region that contains global optimal solution.

6.5.2 A Real Network to Investigate the Effectiveness of SPSA Algorithm

Network Background and Preliminary Work

The second test network is SR-81 corridor in Dallas Fort Worth. A DynaSmart-P network has been developed elsewhere, and it is converted into the above CTM-based network representation in this study.

Due to the differences in the network representation (e.g., the travel demand release mechanisms are different) and lack of further data support, the network was slightly modified. In the converted Fort Worth network, 46 intersections are signal controlled and two on ramps are metered. The signals are most vehicle-actuated controllers; since we herein consider time-of-day control, all controllers are assumed pre-timed but with the same phasing sequence and phase diagrams as specified in the original network, and the green splits (metering rates) as well as offsets between adjacent intersections are subject to optimization.

The preliminary experimentations indicate that the network flow pattern is sensitive to the control settings. In this study controllers are different from their on site counterparts, and because the intersections are closely located and a peak demand is adopted, local or network wide gridlock scenarios very often occur when the green splits are arbitrarily configured. An arbitrary control setting cannot act as θ_0 if it leads to a gridlock because the performance index or objective function value cannot be evaluated under gridlock scenario. A “good” control setting that can at

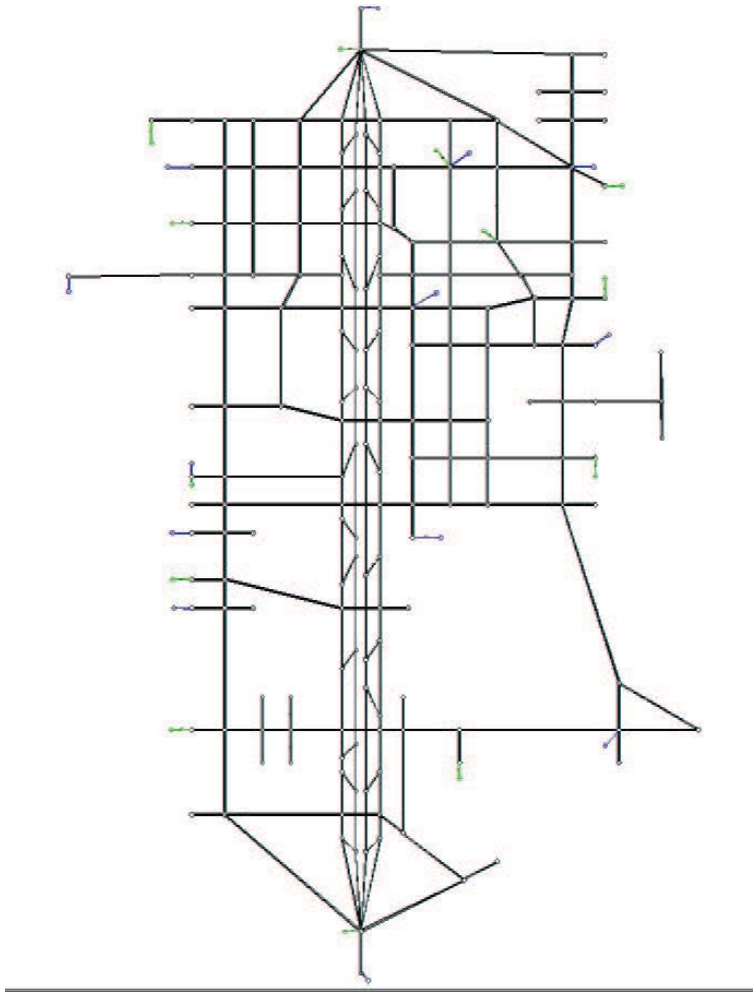


Figure 6.3: Dallas Fort Worth Network Layout

least allow the traffic flows smoothly through the network must be found before the SPSA optimization process could start.

The signal timing design procedure in HCM is followed to come up with such an initial feasible solution. First the demand is loaded onto the network with no controls taking action and the network flow pattern is obtained, i.e., the time-dependent volumes for every movement at each signalized intersection are calculated. Then the cycle length and duration for each phase is computed under the "equi-saturation" logic(Webster 1958). The resulting timing plans and an initial offset of zero for each intersection and a no-meter ($R = q_{max}$ in equation 3.19) solution is found to lead to a smooth flow through the network. Then this solution is taken as θ_0 for the successive SPSA optimization process.

SPSA Application in a Real Network

The SPSA optimization process takes a two-level architecture: first the green splits and metering rates for every signal group at each intersection are computed first; then the offsets are optimized based on the computed green splits. At the first level of optimization a total of 255 control variables are coded as the vector θ. The total network travel time is computed as the objective function. The objective function reduction process is shown in Figure (6.5.2).

Figure (6.5.2) indicates that a total reduction of 5% is gained from optimizing the green splits and metering rates using SPSA method. This also implies that the conventional design method of processing the intersections without considering the flows over the network as in HCM cannot really reach an optimal solution for the network.

The second level of offsets optimization is also conducted using the SPSA algorithm. All offsets are in reference to the start of the overall analysis period; and

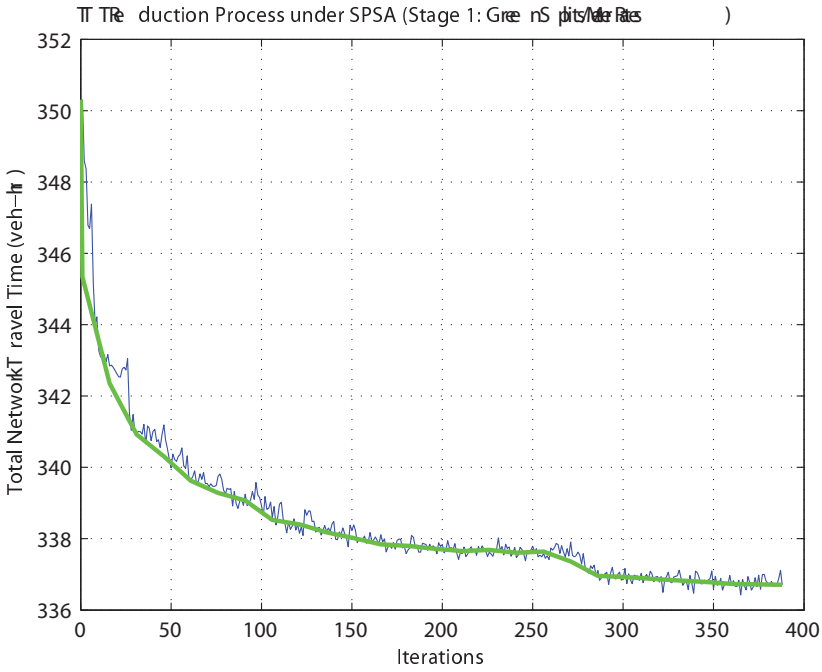


Figure 6.4: SPSA based Optimization of Green Splits/ Metering Rates for Fort Worth Network

the obtained green splits are scaled to take the same cycle length (determined by the critical intersection with the longest cycle length from level 1 optimization). The offsets for all 46 intersections are coded as θ_0.

To study the efficiency in obtaining the optimal solution, the classic "Hill-climbing" algorithm applied in TRANSYT(Robertson 1969a) is also implemented to compute the offset for each intersection and compared with SPSA optimization process. The heuristic method of Hill-climbing proceeds as a sequential adjustments of offset at each intersection. First a step size is selected, the adjustments are performed by a line search to find an improved global objective function that is also computed from network loading. The adjustments are incremental by the selected step size as long as the search improves the objective function. If the search degrades the objective function, the adjustments reverse direction and continues in the other direction at the same step size. In this way, a relatively optimum offset is achieved for the intersection. Then the search proceeds to the remaining intersections. An optimization decision is made for each of several step sizes.

Table 6.2: Computation Performances of Various Algorithms for Dallas Fortworth Network: GA, SPSA and Hill-Climbing(HC)

	Green splits and metering rates			Offsets		
Method	# of evaluations	$z(\theta_0)$	$z(\theta^*)$	# of evaluations	$z(\theta_0)$	$z(\theta^*)$
SPSA	388	364.4	323.1	172	317.2	308.2
GA	3,200	377.9	317.2	/	/	/
HC	/	/	/	470	317.2	311.1

The results are shown in Figure(6.5.2) and Table (6.2). Firstly, the level II optimization of offsets from both SPSA and Hill-climbing methods harvests another 3% of network travel time savings, and the solution from SPSA is slightly better than that from Hill-climbing. Secondly, while Hill-climbing method takes 470 network loading runs, SPSA stabilizes at the same level of objective function with only 180

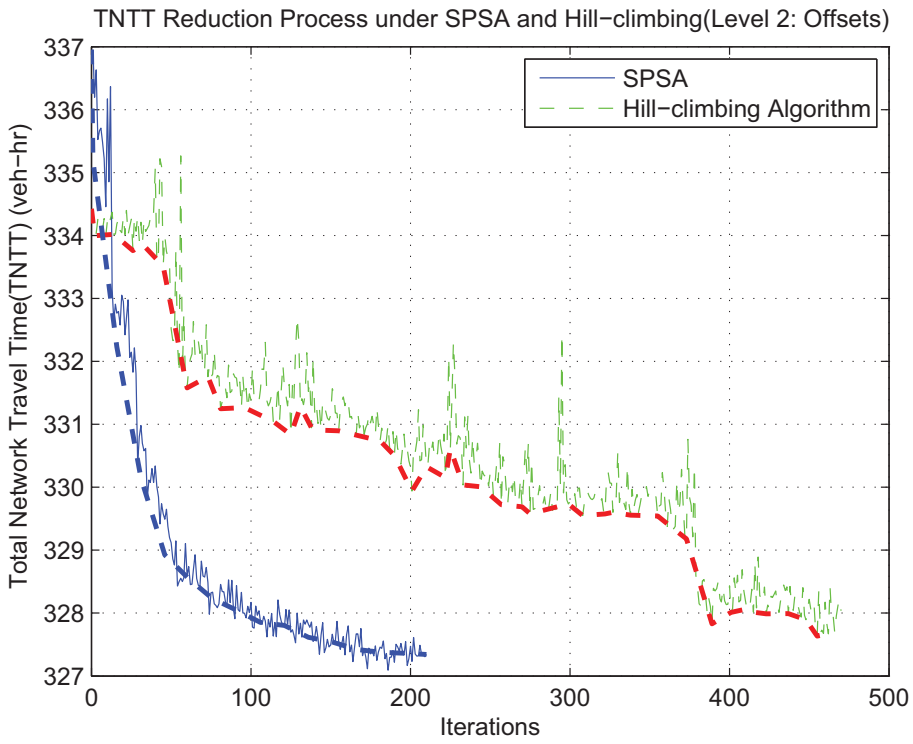


Figure 6.5: SPSA Based Optimization of Offsets for Fort Worth Network

iterations. To note that each SPSA iteration updates the previous solution with two perturbations and each perturbation needs evaluation, SPSA method saves one-third of the computational resources in this 46-variable problem.

System Efficiency Under Various Demand Levels

To test the improvement of the global optimal control from heuristic optimization method, the traffic loads have been uniformly changed and the same test procedures were done as shown in the following table (6.3).

Table 6.3: System Efficiency Improvements Under Different Traffic Loads: Fortworth Network

	Green splits and metering rates			Offsets			Overall improvement
Demand	$z(\theta_0)$	$z(\theta^*)$	Δ	$z(\theta_0)$	$z(\theta^*)$	Δ	
100%	364.4	323.1	11%	323.1	308.2	4%	15%
80%	313.2	256.8	18%	256.8	250.6	2%	20%

It is easily seen that the total network travel time has seen more improvement under light traffic than near-saturation traffic.

6.5.3 Practical Guidelines for Applying SPSA in Optimal Integrated Corridor Control

Choosing appropriate parameters for the gain sequence a_k and c_k is crucial to the performance of SPSA process. In (Spall 1998), Spall provided guidelines for the choice of the related parameters, i.e., α, γ, a, A and c.

With the Bernoulli ± 1 distribution for Δ_k, c can be set at a level approximately equal to the standard deviation of the measurement noise in $z(\theta)$ so that the magnitude of the approximated gradient $\hat{g}_k(\hat{\theta}_k)$ does not go excessively large. In our study, the system performance evaluation is deterministic under CTM based network loading model; in these cases of perfect measurement of $L(\theta)$, c can be some small positive number. In our experimentations with various networks in the normalization scheme, it is found that 0.05 provides acceptable results, namely the change in each element of θ in the initial iterations is in the magnitude of around five percent.

It is also suggested that a "stability constant" A should be used for the sequence of a_k and A and a should be chosen together to ensure the practical performance. A useful guideline for choosing A is found to be 10% (or less) of the maximum

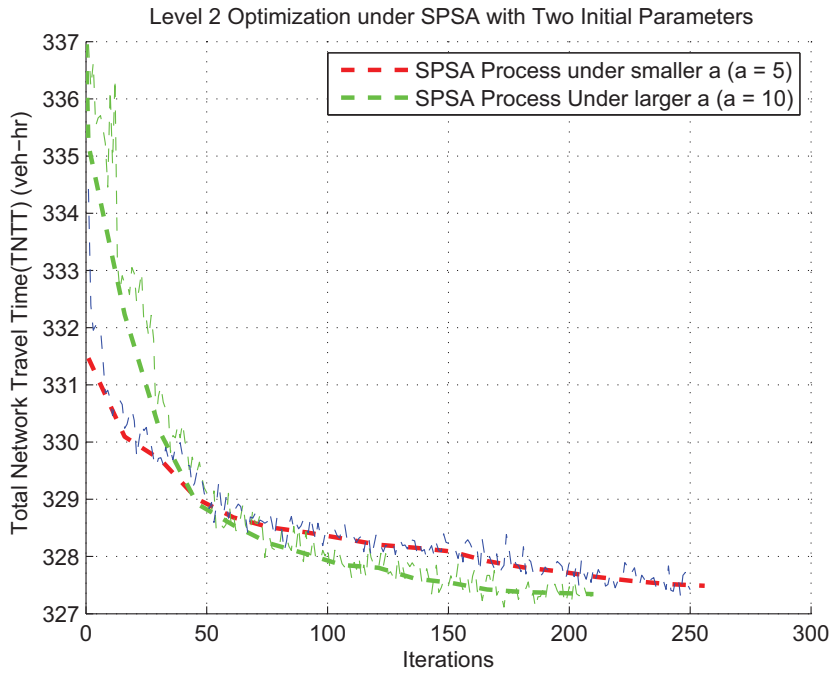


Figure 6.6: Total Network Travel Time Under Hill-climbing and SPSA Algorithms

number of expected or allowed iterations. Meanwhile, Spall (Spall 1998) suggests to run a few preliminary replications of $\hat{g}_0(\hat{\theta}_0)$, and choosing a such that $\frac{a}{(A+1)^\alpha}$ times the magnitudes of $\hat{g}_0(\hat{\theta}_0)$ should be approximately equal to the smallest change in θ. It is found that a larger a could lead to faster convergence to the optimal solution, but it may also run into the risk of too often reaching infeasible solutions (gridlock in our corridor control context) and it has to be projected into the feasible solution region again. Figure 6.6 shows two SPSA processes with various a in the Fort Worth network example. A "greedy" a could reach the solution quicker, but during the process more gridlock conditions have been encountered and those iterations have to be discarded.

Following the above guidelines, we have found that the initial changes in phase duration or offset values can generally provide smoother SPSA process when they are in the magnitude of 3-4 seconds.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 7

Numerical Experimentation with Efficiency-Equity Solution to Integrated Corridor Control

In this chapter we present two examples to examine the effectiveness of the proposed efficiency-equity solution to the integrated corridor control problem using the algorithms developed from previous chapters. The first example is a simple many-to-one freeway corridor with four ramp meters controlling the flow into the freeway. This example is contrived to examine the changes of the efficiency and equity measures and the changes of the control plans under different control objectives. Based on the formulated bi-criterion objective as specified in the previous chapters, various weights are given to the efficiency and equity objectives functions to form different programs. Their performances of system efficiency and user equity are then compared and analyzed. Similar programs are also constructed and solved for the real network, and more detailed analysis is performed, aiming to gain better

understanding of the interaction between the system efficiency and user equity within a general corridor network.

7.1 A Simple Linear City Case Study

A highly simplified freeway corridor is contrived as shown in Figure 7.1. This network has five origins and only one destination and is a many-to-one network. Hence only one route is available for any O-D pairs.

The following geometric features are also specified:

- Node 2, 3, 4 and 5 are all metered node
- Freeway links are uniformly one mile long, and have free flow speed (FFS) of 60 mph; ramp links are all 0.4 mile long and has a FFS of 40 mph;

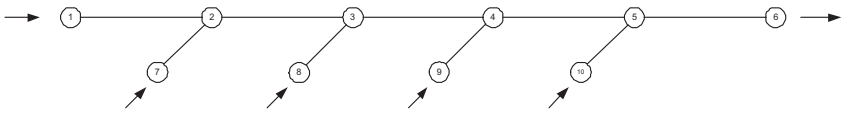


Figure 7.1: The Linear City Case Network Layout

The demand is set as in Table (7.1) with a peaking release pattern as in Figure (7.2).

Table 7.1: Trip Rate Table for the Linear City Network

		Destination 6
Origins	1	1620
	7	450
	8	450
	9	450
	10	450

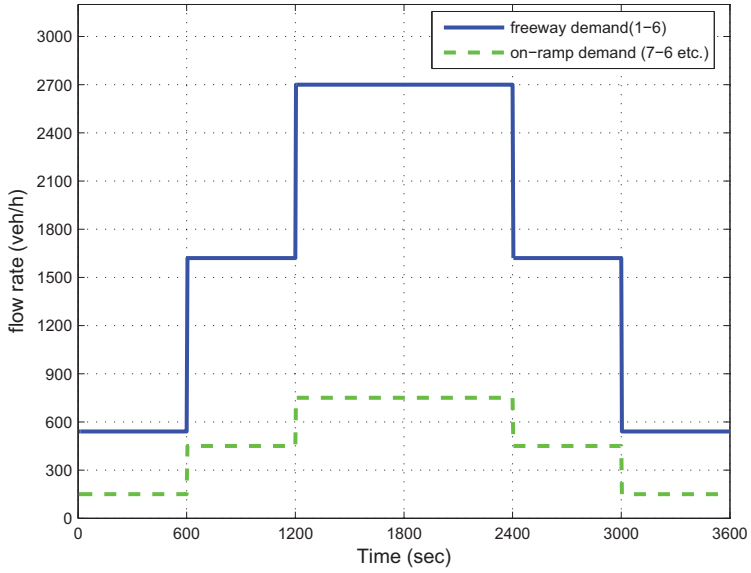


Figure 7.2: The Peaking Demand Pattern for the Linear City Network

7.1.1 Various Control Objective Specifications

Corresponding to the specifications in Chapter (5), the control programs with various control objectives have been set up on the linear city network:

- Criterion I: total travel time (TTT) only. In this network minimization of TTT it is equivalent to minimization of total delay.
- Criterion II: Mean Difference only. As calculated in (3.47).
- Criterion III: Critical Cost Ratio (CR) only. As calculated in (3.48).
- Criterion IV: Balanced combined measure α as calculated in (5.2).

For the sake of simplicity, only one ramp metering rate for each meter is defined for the entire loading horizon. The heuristic optimization processes using

Table 7.2: Efficiency-Equity Measures Under Various Objectives in the Linear City Network

Scenarios	TTT (veh-hr)	MD (veh-hr)	Gini	CR	Most Disadvantaged O-D
No Control	609.1	112.8	0.54	3.52	1-6
Criterion I	441.4	45.6	0.50	2.71	7-6
Criterion II	445.4	40.1	0.51	2.48	1-6
Criterion III	443.4	43.1	0.49	2.29	1-6
Criterion IV	444.2	40.5	0.51	2.41	1-6

SPSA method are initiated to optimize the control objectives listed above. Table (7.2) listed the efficiency and equity measures as well as the most disadvantaged traveler groups corresponding to the critical cost ratio.

We have the following observations from Table (7.2).

- The most significant finding is that the total travel time shows little variation under different control plans after optimization. It is true that with the objective of minimizing the total travel time, the total system efficiency is the best; but when optimizing other objective functions, including aggregate and disaggregate equity measures (MD or CR) and the balanced measure α, the increase of the total travel time is trivial. The total travel time convergence processes under various control objectives can be found in Figure 7.3. Compared to the uncontrolled scenario, all four control scenarios have reached improved system efficiency; and surprisingly they all at similar numerical levels.

- With MD or CR as the sole objective, the optimization processes can reach the best aggregate and disaggregate equity performances, respectively. As noted

above, the changes of the total travel time is trivial while significant differences of the equity measures are observed, considering the magnitude and layout of the network.

- When the simple corridor is uncontrolled, the ramp traffic from the farthest ramp is the most disadvantaged as they have to experience all delays caused by the downstream junction congestion during the peaking. However, when under control with equity introduced, the most disadvantage traveler group becomes again the freeway traffic, i.e., the travelers with longest travel distance (travel time). Compared to the uncontrolled case, the minimization of the total travel time renders the ramp traffic from node 7 to be the most disadvantaged. This finding confirms the common critique to the prevalent ramp metering strategies as the experiences of the Twin Cities travelers.

- Gini Coefficient. The commonly used Gini Coefficient is found to be insensitive indicator in this probelm. Nevertheless, it provides the measure that is easy to understand and a common ground for comparison in the context of social welfare.

Equity Elasticity

One interesting question arises when looking at the relative gains and losses on either system efficiency and user equity of various control programs: can the losses of efficiency losses be justified by the gains in equity? In economics this question is answered generally by the concept of *elasticity*, i.e., the ratio the proportional change of one variable with respect to that of another variable, e.g., the price elasticity of demand. In this section, we develop the following *efficiency elasticity* of equity (EEE)

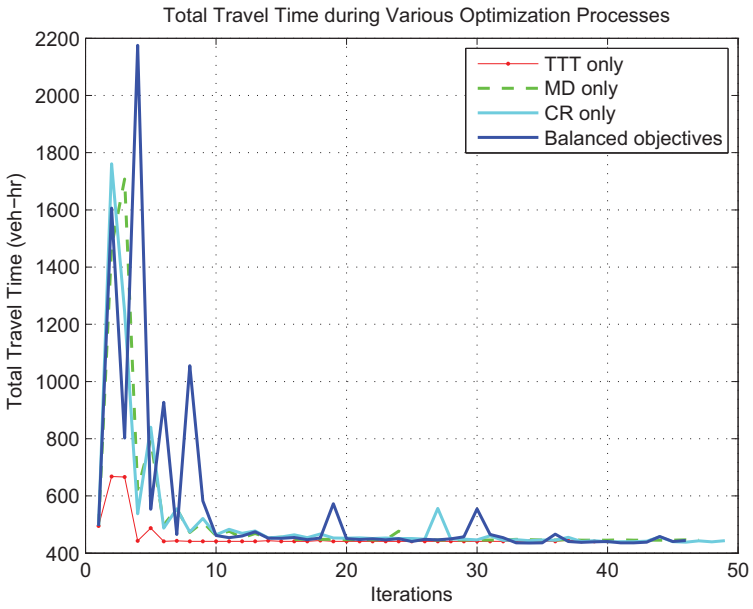


Figure 7.3: The Changes of TTT under Various Optimization Processes

to answer the question. EEE is defined as follows:

$$E_{TTT,\Psi} = \left| \frac{\partial ln\Psi}{\partial lnTTT} \right| = \left| \frac{\partial \Psi}{\partial TTT} \cdot \frac{TTT}{\Psi} \right|$$

where Ψ refers to any of the equity measures including *MD, RMD, CR* or *Gini Coefficient*. When no formulae are available to characterize both variables, $E_{TTT,\Psi}$ can also be defined in percentage changes as:

$$E_{TTT,\Psi} \simeq \frac{\%\Delta\Psi}{\%\Delta TTT} = \frac{(\Psi - \Psi_0)/\Psi_0}{(TTT - TTT_0)/TTT_0} \tag{7.1}$$

Following the arguments in elasticity, here are the interpretations of the equity elasticity:

- $E_{TTT,\Psi} = 0$: perfectly inelastic;

- $0 < E_{TTT,\Psi} < 1$: relatively inelastic;
- $E_{TTT,\Psi} = 1$: unitary elastic;
- $< 1 E_{TTT,\Psi} < \infty$: relatively elastic;
- $E_{TTT,\Psi} = \infty$: perfectly elastic.

When the equity is inelastic w.r.t. efficiency, the losses in efficiency after introducing equity cannot be compensated for by the gains in equity and it would be arguable to do so. But when the equity is elastic, the gains in equity will be justified. The larger the equity elasticity, the better the balance will be.

For this analysis, the best possible efficiency performance, TTT in Criterion I (C-I), will be selected as TTT_0 and the elasticities of various equity measures under different programs are computed in the following table.

Table 7.3: MD, CR & Gini Coefficient Equity Elasticity: Linear City

	MD	CR	Gini Coefficient
C-II	13.3	2.2	9.4
C-III	12.1	4.4	34.2
C-IV	17.6	3.2	17.5

Table 7.3 clearly shows that in the linear city case, introduction of the equity into control is well justified, since all EEE are relatively elastic.

A Simple Sensitivity Analysis: Downstream Bottleneck

Changes of the model parameters and resulting efficiency and equity performances can better indicate the interaction between two measurement dimensions. A typical but simple bottleneck scenario is generated in this section. With all other settings remaining the same, the only network change is made at the link between

node 5 and 6, where its capacity is decreased from 2000 vph to 1700 vph. From the simple demand-capacity analysis it is expected that the bottleneck is overloaded starting from the second assignment period (Figure 7.2).

The same work procedures followed and the relative efficiency and equity changes are summarized in Table 7.1.1.

Table 7.4: Efficiency-Equity Measures In the Linear City with Bottleneck

Scenarios	TTT	MD	CR	Gini Coefficient	Most Disadv. O-D
C-I	621.4	187.5	7.2	0.5	10-6
C-II	645.4	157.6	4.1	0.47	1-6
C-III	645.6	178.3	3.8	0.43	1-6
C-IV	645.5	158.5	3.9	0.46	1-6

The same set of conclusions can still be drawn as have been done from Table 7.2. In addition, the most noticeable variation of the corresponding cells from Table 7.2 to Table are the most disadvantaged traveler groups under different optimization goals. In Table 7.2 the most disadvantaged group is 7-6 when optimizing TTT only, while the group becomes 10-6 when the bottleneck is present. It implies that the metering is more strict with the upstream ramp when only peaking is present as indicated in Figure 7.2. In comparison, the metering will be more strict with the downstream ramp that closer to the bottleneck in the presence of the bottleneck. But in all other optimization after introducing equity measures, the freeway or long distance traveler group unanimously becomes the most disadvantaged one.

In lieu with Table 7.3, the same set of efficiency elasticities of equity can be computed as in the following table.

Table 7.5 indicates that all EEE are still elastic in the presence of bottleneck; but it is also noticed that the elasticities are in general smaller in Table 7.3. It is then anticipated that the equity gains are smaller when the network congestion spans

Table 7.5: MD, CR & Gini Coefficient Equity Elasticity: Linear City with the Bottleneck

Scenario	MD	CR	Gini Coefficient
C-II	4.1	11.1	1.6
C-III	1.3	12.1	3.6
C-IV	4.0	11.8	2.1

longer.

7.1.2 A Short Discussion

First, the simple example confirms our belief that under similar level of system efficiency, various equity performances can be obtained when the new dimension of control objective is introduced. This is an encouraging finding, since it reveals the potential of introducing the equity measures to re-distribute the delays among traveler groups, as seen from the above analysis. After the introduction of equity measures in the control objective, the most disadvantaged traveler group becomes the one with the longest trip length. Within the framework of various equity measures discussed in the review section (2.1), introducing equity measures into control design improves the vertical equity of the system. The sensitivity analysis with the bottleneck confirms the findings as well. The losses of system efficiency can be justified by the gains in user equity from the elasticity computation. Next we proceed to test the mechanisms in a more realistic network.

7.2 Efficiency-Equity Control Experiments on a Real Corridor Network

In this section, we continue to use the Dallas Fort Worth network to investigate the balanced efficiency and equity corridor control design program. Only system efficiency measures were selected when developing the heuristic optimization algorithm in the previous chapter.

Similar to the first many-to-one freeway example, various experiments are set up to examine the performances under different control programs. As compared in the previous chapter, genetic algorithm is more reliable to reach a better solution and thus selected to compute the corresponding control plans.

Three of the convergence processes were shown in Figure 7.4-7.5. From the figures it can be seen that the changes of total travel time (TTT) is independent of those of the equity measures, disaggregate measure of critical trip cost ratio in particular. While during optimization of TTT only as in Figure 7.4, the changes of critical trip cost ratio is in random oscillation in general, and vice vera (Figure 7.5).

Similar findings can be drawn from Table 7.6 as in Table 7.2. With various control programs, the total travel time (TTT) vary within the range of less than 7%. However, the variation of the relative mean difference and critical trip cost ratios change significantly. It clearly indicates that introduction of the equity measures improves the network performance. This is particularly true when applying the balanced control objective of α, where the TTT, RMD and CR are balanced compared to other scenarios.

Figure 7.8-7.10 indicate the most disadvantaged O-D pairs under various control optimization programs. Under all four control programs, the most disadvantaged traveler groups are the ones that are generally perpendicular to the major

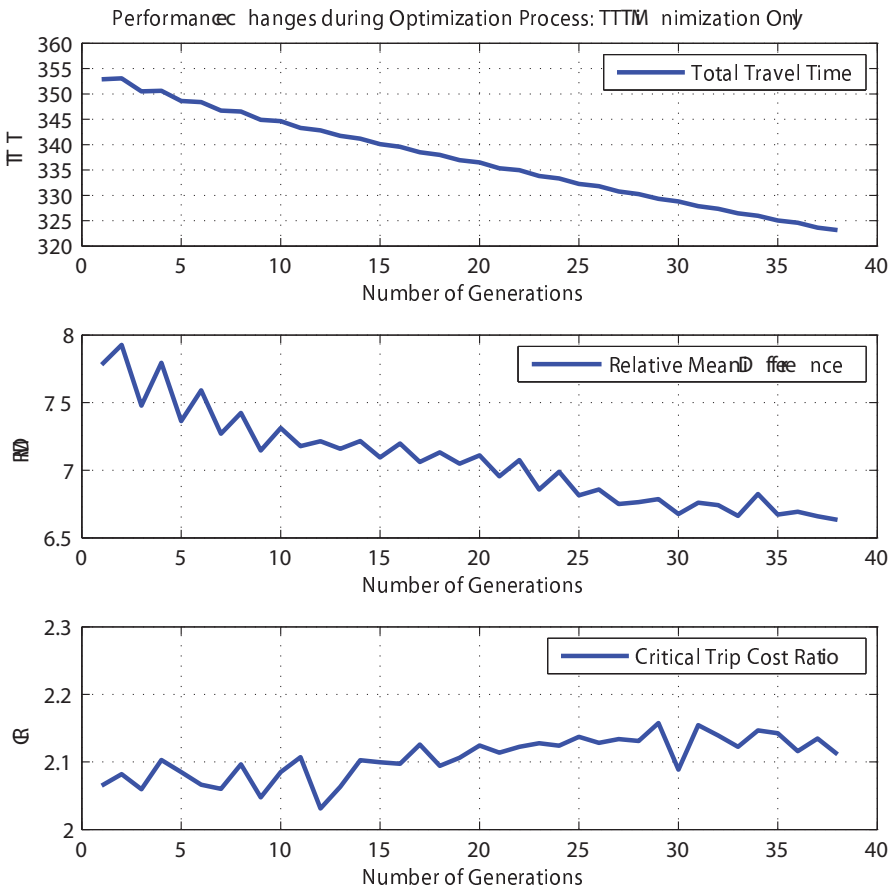


Figure 7.4: Convergence Processes of Efficiency and Equity Measures: Under Efficiency Optimization Only

travel direction; recalling the illustrative example in Chapter 4, it implies that the traffic control will favor the major corridor direction, even though efforts were made to improve the system efficiency and user equity.

Table 7.7 and Figure 7.11 summarized the dispersion of the relative path travel cost along major routes between all O-D pairs in the network. The most

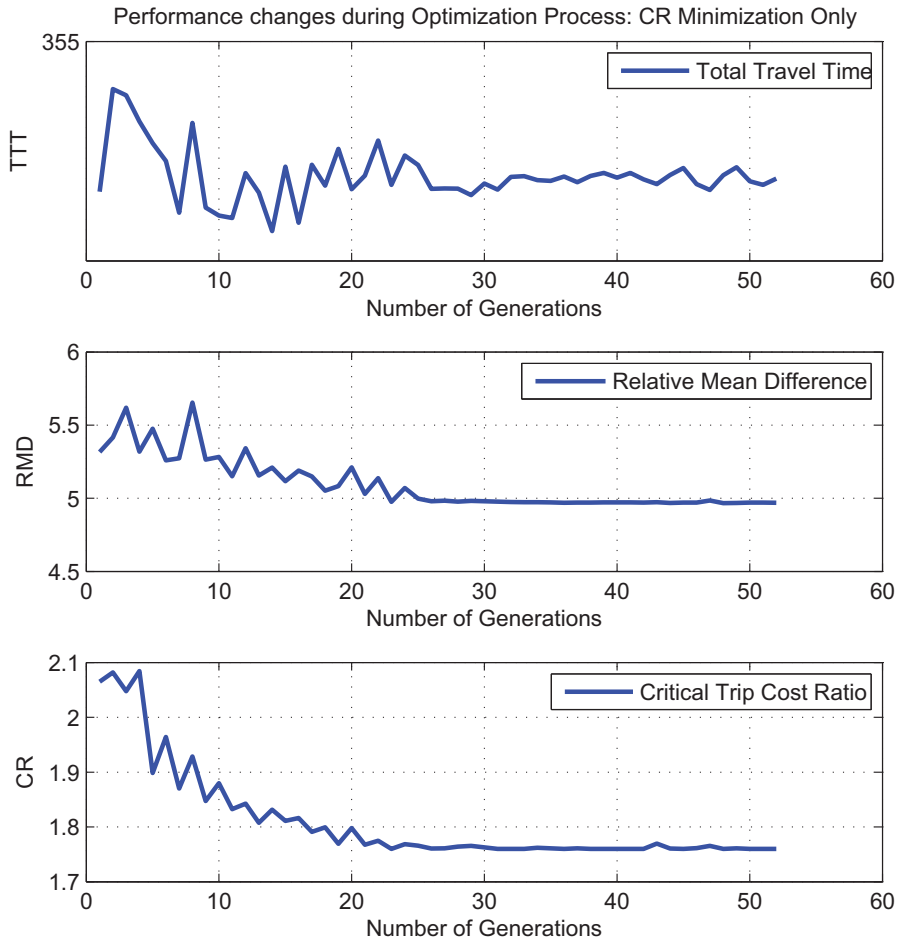


Figure 7.5: Convergence Processes of Efficiency and Equity Measures: Under Disaggregate Efficiency Optimization Only

noticeable comparison from Table 7.7 is that optimization of critical trip cost ratio (CR) do not reduce the variation of the relative path travel cost; in fact the average relative path travel cost when optimizing CR only is even higher than that of optimizing TTT only. This finding is rather counterintuitive and rebuts some past

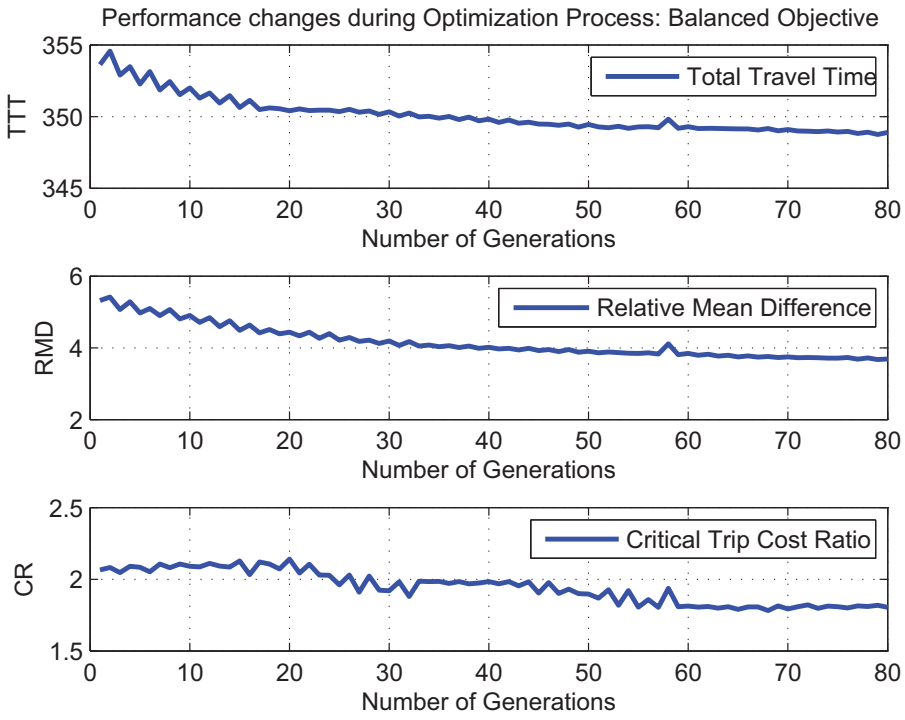


Figure 7.6: Convergence Processes of Efficiency and Equity Measures: Under Balanced Objective α

Table 7.6: Efficiency-Equity Measures Under Various Objectives: Fortworth Network

Scenarios	TTT (veh-hr)	Mean Difference (MD) of RPTC	Gini Coefficient	Critical Trip Cost Ratio (CR)	Most Disadvantaged O-D
TD only (I)	323.1	6.63	0.85	2.1	204-203
RMD only (II)	343.4	4.75	0.65	1.85	204-203
CR only (III)	346.5	6.27	0.73	1.75	187-196
α (IV)	342.8	4.83	0.71	1.76	189-194

studies such as (Chen & Yang 2004)(Meng & Yang 2002). In their studies, the "equity" measure was solely relying on the disaggregate measures of the critical trip cost

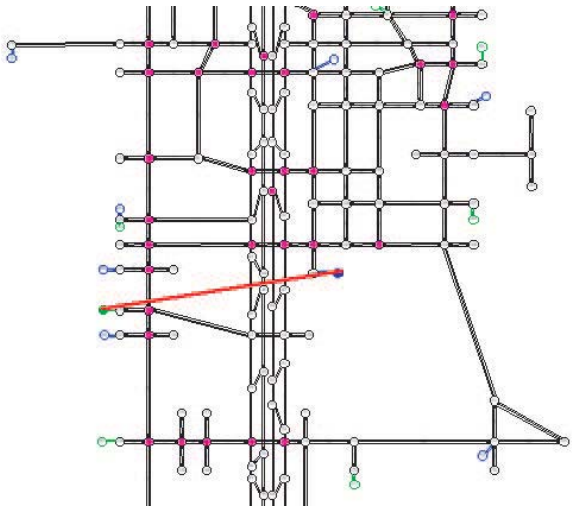


Figure 7.7: Most Disadvantaged O-D Pair(204-203): Under Efficiency Optimization Only

Figure 7.8: Most Disadvantaged O-D Pair(204-203): Under Aggregate Equity Optimization Only

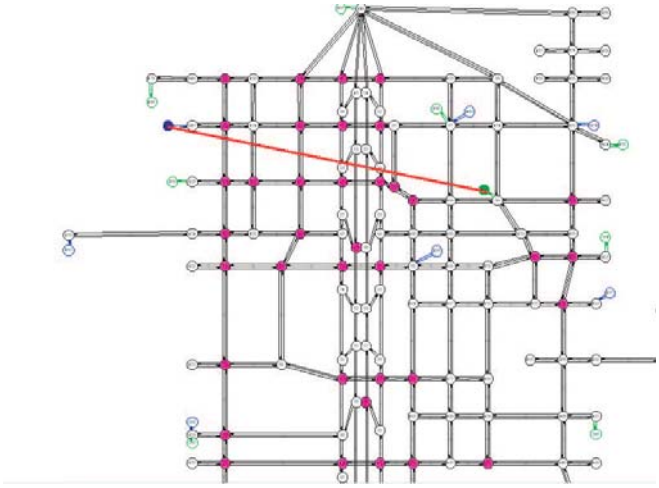


Figure 7.9: Most Disadvantaged O-D Pair (187-196): Under Disaggregate Equity Optimization Only

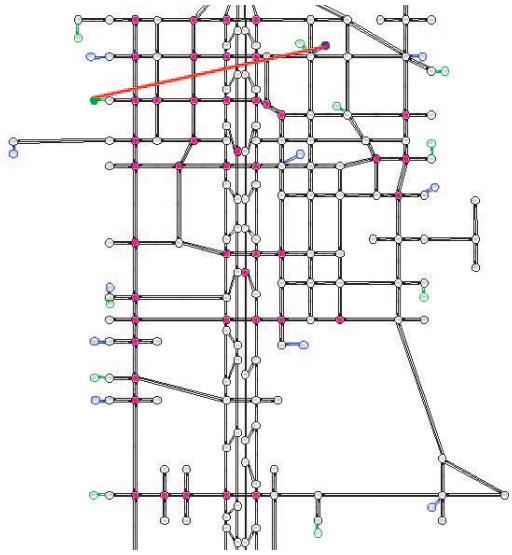


Figure 7.10: Most Disadvantaged O-D Pair (189-194): Under Balanced Efficiency-Equity Optimization

Table 7.7: Dispersion Statistics of the Relative Path Travel Cost

Scenarios	TTT Only (I)	RMD Only(II)	CR Only (III)	α (IV)
Average RMD	1.27	1.24	1.28	1.26
Std. Dev.	0.18	0.12	0.15	0.13

ratio. From Figure 7.11 it can also be seen that even the maximum trip cost ratio is confined, the dispersion of the RPTC was not improved when optimizing CR only. In contrast, both balanced α and minimization of RMD only could lead to a more even distribution of the RPTC.

Similar elasticity analysis is also performed with the efficiency and equity measures in the network. The EEEs are tabulated as follows. As these elasticities

Table 7.8: MD, CR & Gini Coefficient Equity Elasticity: Dalls Network

Scenario	MD	CR	Gini Coefficient
C-II	4.5	3.7	1.9
C-III	0.7	1.9	2.3
C-IV	4.5	2.7	2.7

are obtained with this real network, they are more demonstrative than the results in Table 7.3 and Table 7.5, which are more illustrative qualitatively. In Table 7.8, except for the row of C-III, all elasticities are greater than one, implying that optimization of MD and the balanced objective α can be justified. The C-III program is to minimize CR; it again implies that this disaggregate measure may not be able to fully characterize the control equity aspect in this network.

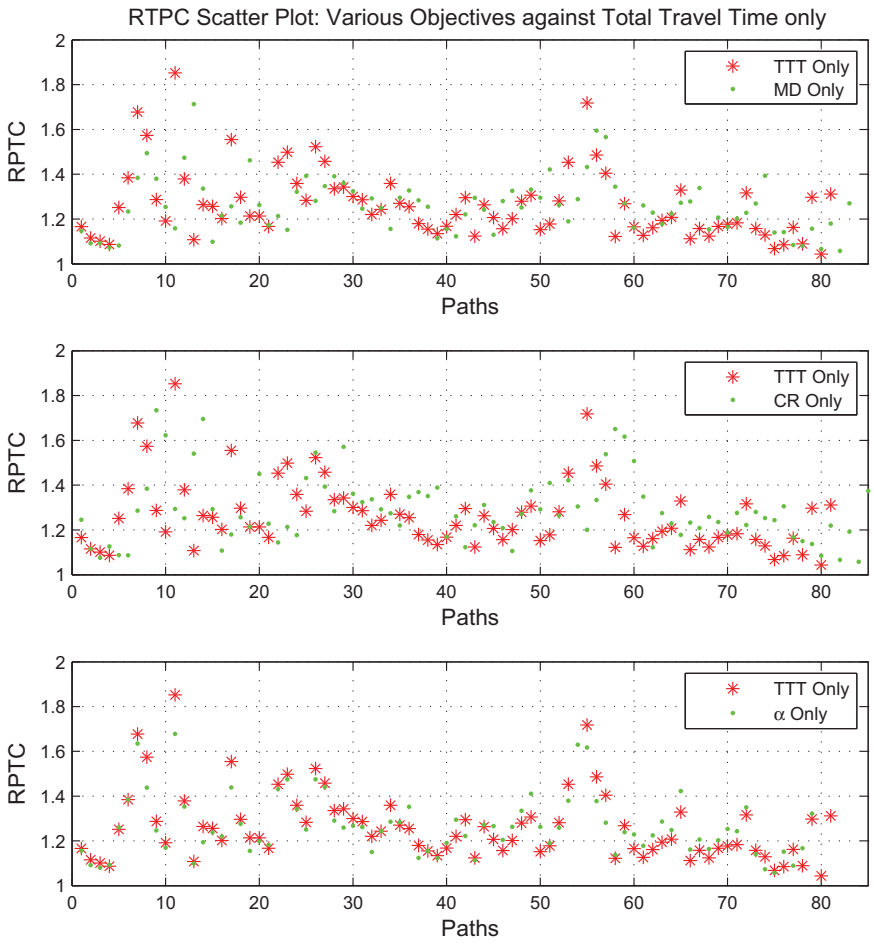


Figure 7.11: Relative Path Travel Cost (RPTC) Scatter Plots under Various Control Objectives After Optimization

7.3 Summary

Two examples are used to illustrate the effectiveness of the proposed bi-criterion integrated corridor control design programs. In the experiments, system

efficiency is represented by the total travel time, while the user equity is analyzed at both aggregate level (mean difference and Gini Coefficient) and disaggregate level. When optimizing only one single objective, TTT, MD or CR, the other measures cannot be guaranteed to converge to an acceptable level. On the contrary, balanced objective that incorporates both efficiency and equity measures sees more satisfactory results. The results indicate that some common practice such as measuring equity solely by disaggregate measures is misleading. Designing corridor control programs, therefore, will have to consider equity measures at both aggregate and disaggregate levels. Compared to efficiency measures optimization only, introducing user equity into optimal corridor control may slightly degrade the system efficiency, but the solution will harvest significant equity improvements. The efficiency elasticity of equity analysis justifies the introduction of equity into corridor control.

Chapter 8

Conclusions

As professionals manage ever busier transportation corridors, interest is being generated among the traveling public. The distributive effect of traffic management systems among users will become more prominent. Incorporation of user equity measures into designing new or updating existing control systems will be increasingly required. The proposed dissertation research is perhaps the first systematic study to develop network traffic control with balanced efficiency and equity through integration of urban signals and ramp metering in a coherent manner. This chapter summarizes the major research work in this dissertation and defines the future research potential.

8.1 Summary of Major Research Work

Two components are fundamental to an integrated corridor control program: 1) the underlying dynamic traffic flow model and, 2) the control plan computation methods. Through an extensive literature review, we find that most past studies relied on the two queuing models or their variations to depict the dynamic

traffic evolution over the network, namely point-queue or vertical queue model and spatial queue or horizontal queue model. While these two types of models are easy to implement and convenient for implementing gradient-based optimization methods, they cannot accurately estimate the transient queueing dynamics from the cyclic control actions. Thanks to a recent finite difference solution scheme to the well-accepted kinematic wave model, we develop the cell transmission model (CTM) based dynamic network loading (DNL) tool in Chapter 3. Meanwhile, we focus on adapting the CTM rules to model the flow updates at general urban signalized intersections and ramp meters, and priority rule controlled junctions such as STOP signs and yield signs. With this tool, we can model the general corridor network flow dynamics under control.

In Chapter 3 we also identified the system efficiency and user equity measures. The review of the equity issues in both planning and operations indicated that user equity or user fairness has been much neglected in traffic control studies, and the research has drawn conflicting and ambiguous conclusions. Based on the built dynamic traffic flow model, we formulated the user equity measures within the corridor control context and classified the horizontal equity and vertical equity measures at both aggregate and disaggregate level. Together with the typical efficiency measures, this completes the set of measurements of effectiveness (MOE) for corridor control systems.

Regarding other fundamental component of control program, we identified two classes: 1) rule-based feedback control methods and, 2) mathematical program based feed-forward control. In Chapter 4, we proposed a generalized rule-based local synchronization control scheme (LSC) to coordinate the actions of adjacent controllers that are tightly located, e.g., within a freeway interchange area. The comparison of this scheme with other strategies indicate the rule-based control methods

are surprisingly not inferior to even global optimal control methods in both efficiency and equity measures.

As reviewed in Chapter 2, when the control problem is formulated as a mathematical program, two types of solution algorithms exist to compute the control plan: 1) heuristic searching algorithms and, 2) gradient information based linear and non-linear programming methods. Heuristic searching algorithms are generally used to compute the control plan with more realistic and complex traffic flow models, while linear and non-linear programming algorithms are based on simplified flow models. Heuristic searching algorithms generally require a large number of system performance evaluations because of its stochastic searching nature. On the contrary, linear and non-linear programs can take advantage of the gradient information and reach the optimal solution much faster. In lieu of the pros and cons of both types of algorithms, we adapted the simultaneous perturbation stochastic approximation (SPSA) method to make use of the advantages of both types of algorithms. Compared to the well-accepted genetic algorithm and hill-climbing method, the developed algorithm could reach the near-global optimal solution in fewer iterations.

The efficiency-equity solution framework has been established in Chapter 6. Three control objectives, total travel time, mean difference and critical trip-cost ratio have been selected for corridor control performance measures. Meanwhile, one metric measure that combines three objectives has also been proposed as the objective function to balance the efficiency and equity simultaneously. Assuming the corridor operation can be improved for any given traffic flow pattern, one simple user behavior model of route selection has been used. The control plans were computed using the developed SPSA algorithm as well as GA algorithm. Through extensive numerical tests in Chapter 4, 5, and 7, the major findings concerning the corridor system efficiency and user equity are as follows:

- Optimization of system efficiency measures only can generally lead to undesirable user equity performances. It implies that some traveler groups will sacrifice to compensate for other traveler groups when the most efficient control plan is implemented. In particular, the most disadvantaged traveler groups will be penalized. As confirmed in both illustrative and real networks, the most disadvantaged traveler groups are those that do not travel along the major traffic direction (the direction parallel to the freeway and the frontage roads).

- Rule-based local synchronization control schemes are not inferior in terms of both efficiency and equity measures. As the major goal is to prevent the queue spillback from spreading locally or even globally, the schemes could improve the efficiency compared to prevalent isolated control strategies. The LSC schemes can also improve the equity aspects compared to the global optimal control that addresses system efficiency only. More rule-based control methods can be derived from the schemes to cope with different congestion patterns.

- Integrated and network-wide coordination control perform better when the network experiences light or medium traffic load. When the network reaches near-saturation state, the benefits from integrated and coordination control diminish. Real network tests indicate generally 15% system efficiency improvement in light-medium traffic, but the improvement will drop to less than 10% when the network becomes congested. In this case, more refined rule-based coordination schemes should be developed to deal with the special situations rather than relying on general control programs.

- The system efficiency and user equity objectives are generally independent from each other. When optimizing one single control objective, the others could display random oscillation. This phenomenon confirms that an integrated

control system is possible to balance the efficiency and equity. Particularly, minimization of total travel time is irrelevant to the critical trip cost ratios; this also implies that an efficient-equitable control program will have to combine both efficiency measures and equity measures.

- The usual treatment of having only disaggregate equity measure is incomplete to model the user equity in general. With minimized critical trip cost ratio, the dispersion of the relative path travel cost can still be large, implying that only restricting the most disadvantaged group from being sacrificed too much is not enough for other traveler groups. Elasticity analysis in the numerical experiments indicated that the gains in equity may not even be justified when optimizing the disaggregate equity measure only. Therefore, equity measures must include both aggregate and disaggregate ones.

- The proposed bi-criterion control program that combines both efficiency and equity measures in its objective can generate a balanced system. While the system efficiency may be slightly degraded if there is any, the equity among travelers can be significantly improved.

8.2 Future Research Directions

This study is only an initial solution towards efficient AND equitable traffic management policies, and many more questions remain to be answered. Here are a few interesting research directions.

1. More realistic user behaviors and the resulting dynamic network flow patterns would be an immediate step to examine the effectiveness of the proposed bi-criterion integrated corridor control method. Dynamic user equilibrium (DUE) traffic assignment has been an active research area. Proactive and Reactive

DUE traffic assignment exist in the field of dynamic traffic assignment (DTA). Once extended to modeling DUE traffic flow patterns, a bi-level programming structure will take shape and the efficient-equitable control and the DTA will be solved iteratively to reach the *mutually consistent* point.

2. Development of online application using simplified traffic flow models. In this research only the off-line control programs have been developed and analyzed, partly due to the non-linearity nature of the CTM solution scheme. Nevertheless, since the entire solution framework can be adapted easily to accommodate online applications, using simplified traffic flow models and consequently adapting to online program will be a direct extension of the research. As some promising research indicates(McLean, Brader, Diakaki & Papageorgiou 1998)(Papageorgiou 1995)(Diakaki et al. 2000), the simple store-and-forward approach can be used to build computationally efficient online applications based on the proposed bi-criterion control.

3. Supplementing traveler information and various policy alternatives will be insightful to understand the deeper interaction between the generative and distributive effect of the control systems. Interesting questions arise here: how do we use traveler information to effect a more equitable and efficient transportation system; does more information bring higher efficiency and greater equity to a system overall; have the policies and measures like HOV lanes and congestion pricing achieved their goals of balancing equity and efficiency? It is expected that the insights and understanding gained from this dissertation work will help answer such questions, and that the developed framework could be adapted to accommodate more management and control measures.

4. Robust corridor control formulations and solutions will add to the vitality of con-

trol studies in general. So far most control studies assume that the collected traffic data input is accurate but this is generally not true. The more sophisticated the programs are, the more detailed detailed and accurate data they require and the more severely the traffic data input errors will affect the effectiveness of the control programs. Therefore, introducing a robust control concept into traffic control will be a very promising direction.

5. Application of the proposed research methodology and conclusions to active transportation planning processes will be good for both research and practice. The applicability of any research can only be improved in real practice; applying the efficiency-equity bi-criterion design concept into the system planning process is expected to encourage research and practice for a more desirable transportation system.

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

Al-Malik, M. (1991), An Investigation and Development of a Combined Traffic Signal Control Traffic Assignment Model, PhD thesis, Georgia Institute of Technology.

Allsop, R. E. (1971), 'Delay minimizing settings for fixed-time traffic signals at a single road junction', *Jounal Inst. Maths Applics* **8**, 164–185.

Allsop, R. E. (1974), 'Some possibilities for using traffic control to influence trip distribution and route choice', *Transportation and Traffic Theory, Proceedings of the 6th International Symposium on Transportation and Traffic Theory* pp. 345–372. University of New South Wales, Sydney, Australia.

Andrews, C., Elahi, S. & Clark, J. E. (1997), 'Evaluation of new jersey route 18 opac/mist traffic control system', *Transportation Research Record* **1603**, 150–155.

Banos, J. C. M. & Papageorgiou, M. (1995), 'A linear programming approach to large-scale linear optimal control problems', *Proceedings of the 34th Conference on Decision and Control* pp. 1115–1119.

Benesch, A. (1915), 'Regulating street traffic in Cleveland', *The American City* **XIII**, 182–184.

Bowman, M. J. (1945), 'A graphical analysis of personal income distribution in the united states', *American Economic Review* **XXXXV**(4), 607–628.

BPR (1964), Traffic assignment manual, Technical report, Bureau of Public Roads, US Department of Commerce, Urban Planning Division, Washinton D C.

Bretherton, D., Wood, K. & Raha, N. (1998), 'Traffic monitoring and congestion management in the SCOOT urban traffic control system', *Transportation Research Record* **1634**, 118–122.

Cascetta, E., Nuzzolo, A., Russo, F. & Vitetta, A. (1996), 'A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks', *Transportation and Traffic Theory, Proceedings of the 13th International Symposium on Transportation and Traffic Theory* pp. 697–711. Lyon, France.

Chabini, I. (1998), 'Discrete dynamic shortest path problems in transportation applications', *Transpn Research Record* **1645**, 170–175.

Chabini, I. & He, Y. (1998), 'A flow-based approach to the dynamic traffic assignment problem: formultaions, algorithms, and computer implementations', *Third Triennial Symposium on Transportation Analysis* .

Chang, G. L. (1993), 'A dynamic system-optimum control model for commuting traffic corridors', *Transpn. Res. -C* **1**(1), 3–22.

Chaudhary, N. A. & Messer, C. J. (1993), 'PASSER IV : A program for optimizing signal timing in grid networks', *Transportation Research Record* **1421**, 82–93.

Chen, A. & Yang, C. (2004), 'Stochastic transportation network design problem with spatial equity constraint', *Transportation Research Record* **1882**, 97–104.

Chen, O. J. (1998), Integration of Dynamic Traffic Control and Assignment, PhD Thesis, Massachusetts INstiute of Technology, Department of Civil and environmental Engineering.

Chin, D. (1997), 'Comparative study of stochastic algorithms for system optimization based on gradient approximations', *IEEE Transcations on Systems, Man, and Cybernetics* p. 244:249.

Cohen, S. (1989), *Evaluation of the Optimized Policies for Adaptive Control Strategy*, Vol. FHWA-RD-89-135, U.S. Department of Transportation Federal Highway Administration, McLean, Virginia.

Daganzo, C. F. (1994), 'The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory', *Transpn. Res. -B* **28**, 269–287.

Daganzo, C. F. (1995), 'The cell transmission model, part ii: network traffic', *Transpn. Res. -B* **29**, 79–93.

D'Ans, G. & Gazis, D. (1976), 'Optimal control of over-saturated and store-and-forward transportation networks', *Transportation Science* **10**, 1–19.

Diakaki, C. & Papageorgiou, M. (1997), 'Simulation studies of integrated corridor control in Glasgow', *Transpn Res. -C* **5**(3/4), 211–244.

Diakaki, C., Papageorgiou, M. & McLean, T. (2000), 'Integrated traffic-responsive urban corridor control strategy in Glasgow, Scotland : Application and evaluation', *Transpn Research Record* **1727**, 101–111.

Dickson, T. J. (1981), 'A note on traffic assignment and signal timings in a signal-controlled network', *Transportation Research, -B* **15**, 267–271.

Donati, F., Mauro, V., Roncolini, G. & Vallauri, M. (1984), 'A hierarchical decentralized traffic light control system', *Proceedings of 9th IFAC Triennial World Congress* .

Fang, C. & Elefteriadou, L. (1006), 'Development of an optimization methodlogy for adaptive traffic signal control at diamond interchanges', *ASCE Journal of Transportation Engineering* **132**(6), 629–637.

Feng, C.-M. & Wu, J. Y.-J. (2003), 'Highway investment planning model for equity issues', *ASCE Journal of Urban Planning & Development* **129**, 161–176.

FHWA (2005), 'Integrated corridor management system(ICMS) work plan', Website http://www.itsdocs.fhwa.dot.gov/icms/icms_workplan.htm. Accessed in July, 2006.

Fisk, C. S. (1984), 'Game theory and transportation systems modellng', *Transportation Research B* **18**(4/5), 301–313.

Foy, M., Benekohal, R. & Goldberg, D. (1992), 'Signal timing determination using genetic algorithms', *Transportation Research Record* **1365**, 108–115.

Friesz, T. (1985), 'Transportation network equilibrium, design and aggregation: key development and research opportunities', *Transpn. Res. -A* **19**, 413–427.

Friesz, T., D.Bernstein, Smith, T. & Wie, B. (1993), 'A variational inequality formulation of the dynamic network user equilibrium problem', *Operations research* **41**, 179–191.

Garrett, M. & Taylor, B. (1999), 'Reconsidering sociao equity in pulic transit', *Berkeley Planning Journal* **13**, 6–27.

Gartner, N. (1981), 'Prescription for responsive urban traffic control', *Control of Urban Traffic Systems, Proceedings of the 1981 Engineering Foundation Conference* .

Gartner, N. H. (1983), 'OPAC: A demand-responsive strategy for traffic signal control', *Transportation Research Record* **906**, 75–81.

Gartner, N. H., Little, J. D. C. & Gabby, H. (1975), 'Optimization of traffic signal settings by mixed-integer linear programming, part i: the network coordination problem', *Transpn. Science* **9**(4), 320–344.

Gartner, N. H., Stamatiadis, C. & Tarnoff, P. (1995), 'Development of advanced traffic signal control strategies for intelligent transportation systems: multilevel design', *Transportation Research Record* **1494**, 98–105.

Gartner, N. J., Little, J. D. & Gabby, H. (1976), 'Simultaneous optimization of offsets, splits, and cycle time', *Transportation Research Record* **596**, 6–15.

Gazis, D. (1964), 'Optimum control of a system of oversaturated intersections', *Operations Research* **12**, 815–831.

Gazis, D. C. (1974), *Traffic Science*, Wiley Press. Chapter 3, Traffic Control-Theory and Application.

Gini, C. (1936), 'On the measure of concentration with especial reference to income and wealth', *Cowles Commission* . Gini C. "Variabilitá e mutabilita" (1912) Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi (1955).

Goldberg, D. (1989), *Genetic algorithms in search optimization and machine learning*, Addison-Wesley Publishing Co., Reading, Mass.

Gomez, G. & Horowitz, R. (2004*a*), 'Globally optimal solutions to the onramp metering problem - Part I', *2004 IEEE Conference on Intelligent Transportation Systems* pp. 509 – 514.

Gomez, G. & Horowitz, R. (2004*b*), 'Globally optimal solutions to the onramp metering problem - Part I', *2004 IEEE Conference on Intelligent Transportation Systems* pp. 515 – 520.

Greenshields, B. (1934), 'A study of traffic capacity', *Proceedings of Highway Research Board* **14**, 448–477.

Hadj-Salem, Blosseville, H. J. & Papageorgiou, M. (1991), 'ALINEA: A local feedback control law for on-ramp metering', *Transportation Research Record* **1320**, 58–64.

Haid, M. A. & Wallace, C. (1993), 'Hybrid genetic algorithm to optimize signal phasing and timing', *Transportation Research Record* **1421**, 104–112.

Haj-Salem, H. & Papageorgiou, M. (1995), 'Ramp metering impact on urban corridor traffic: field results', *Transportation Research, -A* **29**, 303–319.

Hall, M., van Vliet, D. & Willumsen, L. (1980), 'SATURN–a simulation-assignment model for the evaluation of traffic management', *Traffic Engineering and Control* **21**(4), 168–178.

Han, B. & Reiss, R. A. (1994), 'Coordinating ramp meter operation with upstream intersection traffic signal', *Transportation Research Record* **1446**, 44–47.

Hasan, M., Jha, M. & Ben-Akiva, M. (2002), 'Evaluation of ramp control algorithms using microscopic traffic simulation', *Transpn. Res. -C* pp. 229–256.

Head, S., Mirchandani, P. & Shelby, S. (1997), 'Computational improvements for real-time implementation of the controlled optimization of phases traffic signal control strategy', *The Institute for Operations Research and the Management Sciences* . Dallas, TX.

Hegyi, A. (2004), Model Predictive Control for Integrated Traffic Control Measures, PhD Thesis, Technical University of Delft, The Netherlands.

Helbing, D. (2003), 'A section-based queuing-theoretical traffic model for congestion and travel time analysis in networks', *Journal of Physics A: Mathematical and General* **36**, L593–L598.

Henry, J. J., Farges, J. L. & Tuffal, J. (1983), 'The PRODYN real time traffic algorithm', *Proceedings of the 4th IFAC Control in Transportation Systems* .

Henry, R. (2005*a*), Signal timing on a shoestring, Research Report FHWA-HOP-07-006, Sabra, Wang & Associates, Inc. contract no: DTFH61-01-C-00183.

Henry, R. (2005*b*), Signal timing process, Research Report FHWA-HOP-07-006, Sabra, Wang & Associates, Inc. contract no: DTFH61-01-C-00183.

Hunt, P., Robertson, D., Bretherton, R. & Winton, R. (1981), SCOOT - a traffic responsive method of coordinating signals, Technical Report TRRL Laboratory Report 1014, TRRL Department of the Environment.

Jacobsen, L., Henry, K. & Mehyar, D. (1989), 'Real-time metering algorithm for centralized control', *Transpn Research Record* **1232**, 17–26.

Jahn, O., Öhring, R. H. M., Schulz, A. S. & Stier-Moses, N. E. (2005), 'System-optimal routing of traffic flows with user constraints in networks with congestion', *Operations Research* **53**(4), 600–616.

Jin, W. & Zhang, H. (2003), 'On the distribution schemes for determining flows through a merge', *Transpn. Res. -B.* **37**(6), 521–540.

Kane, J. (1964), *Famous First Facts*, New York, The H.N. Wilson Company.

Kotsialos, A. & Papageorgiou, M. (1999), 'Optimal coordinated and integrated motorway network traffic control', *Transportation and Traffic Theory, Proceedings of the 14th International Symposium on Transportation and Traffic Theory* pp. 621–644. Jerusalem, Israel.

Kotsialos, A. & Papageorgiou, M. (2004), 'Efficiency and equity properties of network-wide ramp metering with AMOC', *Transportation Research, -C* **12**, 401–420.

Kotsialos, A., Papageorgiou, M., Mangeals, M. & Haj-Salem, H. (2002), 'Coordinated and integrated control of motorway networks via nonlinear optimal control', *Transpn. Res. C* **10**, 65–84.

Kwon, E., Ambadipudi, R.-P. & Bieniek, J. (2005), 'Adaptive coordination of ramp meter and intersection signal for balanced management of a freeway corridor', *Presented at 2005 TRB Annual Meeting* .

Lakshmanan, T., Nijkamp, P., Rietveld, P. & Verhoef, E. (2001), 'Benefits and costs of transport: classification, methodologies and policies', *Papers of Regional Science* **80**, 139–164.

LeBlanc, L. & Boyce, D. (1986), 'A bi-level programming algorithm for exact solution of the network design problem with user-optimal flows', *Transpn. Res. -B* **20**, 259–265.

Levinson, D. (20003), 'Perspectives on efficiency in transportation', *International Journal of Transport Management* pp. 145–155.

Levinson, D., Zhang, L., Das, S. & Sheikh, A. (2002), 'Ramp meters on trial: evidence from the Twin Cities ramp meters shut-off', *81st TRB Annual Meeting* . Washington, D.C.

Lighthill, M. & Whitham, J. (1955), 'On kinematic waves. I. flow modeling in long rivers. II. a theory of traffic flow on long crowded roads', *Proc. Roy. Soc. A* **229**, 281–345.

Litman, T. (2007), Evaluating transportation equity: Guidance for incorporating distributional impacts in transportation planning, Research report, Victoria Transport Policy Institute, Victoria, BC, Canada.

Little, J. D. (1966), 'The synchronization of traffic signals by mixed-integer linear programming', *Operations Research* **14**, 568–594.

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**, 40–46.

Lo, H. K. (1999), 'A novel traffic signal control formulation', *Transportation Research, -A* **33**, 433–448.

Lo, H. K. (2001), 'A cell-based traffic signal formulation: strategies and benefits of dynamic timing plans', *Transportation Science* **35**(2), 148–164.

Lo, H. K., Chang, E. & Chan, Y. C. (2001), 'Dynamic network traffic control', *Transpn. Res. -B* **35**, 721–744.

Lorenz, M. (1905), 'Methods of measuring the concentration of wealth', *Publications of the American Statistical Association* **9**(70), 209–219.

Lowrie, P. (1981), 'SCATS: the sydney co-ordinated adaptive traffic system', *International Conference on Raod Traffic Signaling* .

Ma, J., Sun, L. & Sperling, D. (2005), 'Technical and social issues in implementing electronic road pricing (erp) in shanghai, china', *Proceedings of the 8th IEEE Intelligent Transportation System Conference* . Vienna, Austria.

MacGowan, J. & Fullerton, I. (1979-1980), 'Development and testing of advanced control strategies in the urban traffic control system', *Public Roads* **43**(2,3,4). Three articles.

Maher, M. J., Zhang, X. & van Vliet, D. (2001), 'A bi-level programming approach for trip matrix estimation and traffic control problems with stochastic user equilibrium link flows', *Transpn. Res. -B* **35**, 23–40.

McLean, T., Brader, C., Diakaki, C. & Papageorgiou, M. (1998), 'Urban integrated traffic control in glasgow, scotland', *Road Transport Information and Control* **21**, 243–249.

Meldrum, D. & Taylor, C. (1995), 'Freeway traffic data prediction using artificial neural networks and development of a fuzzy logic ramp metering algorithm', *Final Technical Report* . WA-RD 365-1, WSDOT, USA.

Meng, Q. & Yang, H. (2002), 'Benift distribution and equity in road network design', *Transpn. Res. -B* **36**, 19–35.

Merchant, D. & Nemhauser, G. (1978*a*), 'A model and an algorithm for the dynamic traffic assignment problems', *Transportation Science* **12**, 183–199.

Merchant, D. & Nemhauser, G. (1978*b*), 'Optimality condtions for a dynamic traffic assignment model', *Transportation Science* **12**, 200–207.

Messer, C. J., Whitson, R. H., Dudek, C. L. & Romano, E. J. (1973), 'A variable sequence multiphase progression optimization program', *Highway Research Record* **445**, 24–33.

Messmer, A. & Papageorgiou, M. (1990), 'METANET: a macroscopic simulation program for motorway networks', *Traffic Engineering and Control* **31**, 466–473.

Michalopoulos, P. G., Stephanopoulos, G. & Stephanopoulos, G. (1981), 'An application of shock wave theory to traffic signal control', *Transpn. Res. - B* **15**(1), 35–51.

Miller, A. (1963), 'Settings for fixed-cycle traffic signals', *Operations Research Quarterly* **13**, 373–386.

Mirchandani, P. & Head, L. (2001), 'A real-time traffic signal control system: architecture, algorithms and analysis', *Transportation Research, -C* **9**, 415–432.

Morgan, J. T. & Little, J. D. (1964), 'Synchronizing traffic signals for maximal bandwidth', *Operations Research* **12**, 896–912.

Nash, J. (1951), 'Noncooperative games', *Annals Math.* **54**, 286–295.

Newell, G. F. (1956), 'Statistical analysis of the flow of highway traffic through a signalized intersection', *Quarterly Applied Mathematics* **13**, 353–369.

Newell, G. F. (1998), 'The rolling horizon scheme of traffic signal control', *Transportation Research A* **32**(1), 39–44.

Nie, Y. (2006), A Variational Inequality Approach For Inferring Dynamic Origin-Destination Travel Demands, PhD Thesis, University of California at Davis, Department of Civil and environmental Engineering.

Nie, Y., Nie, X. & Zhang, H. M. (2004), 'The relative performance of time-dependent shortest path algorithms: a network expansion perspective', *Proceedings of the 8th International Conference on Applications of Advanced Technologies in Transportation* .

Olmo, P. D. & Mirchandani, P. B. (1992), 'RAELBAND: an approach for real-time coordination of traffic flows on network', *Transportation Research Record* **1494**, 106–116.

Owen, L. & Stallard, C. (1999), 'Rule-based approach to real-time distributed adaptive signal control', *Transpn Research Record* **1683**, 95–101.

Paesani, J., Perovich, P. & Khosravi, E. (1997), 'System wide adaptive ramp metering in Southern California', *Proceedings of the 7th ITS America Annual Meeting* . Washington D.C.

Papageorgiou, M. (1995), 'An integrated control approach for traffic corridors', *Transpn. Res. -C* **3**(1), 19–30.

Papageorgiou, M. (2000), 'Freeway ramp metering: an overview', *Proceedings of IEEE Intelligent Transportation Systems* .

Papageorgiou, M., Blosseville, J.-M. & Hadi-Salem, H. (1990*a*), 'Modeling and real-time control of traffic flow on the southern part of boulevard Peripherique Paris: Part I: Modeling', *Transportation Research -A* **24**, 345–359.

Papageorgiou, M., Blosseville, J.-M. & Hadi-Salem, H. (1990*b*), 'Modeling and real-time control of traffic flow on the southern part of boulevard Peripherique Paris: Part II: coordinated on-ramp metering', *Transportation Research -A* **24**, 361–370.

Peeta, S. & Jayakrishnan, R. (2001), 'Foundations of dynamic traffic assignment: the past, the present and the future', *Network and Spatial Economics* pp. 233–265.

Pooran, F. J. & Lieu, H. C. (1994), 'Evaluation of system operating strategies for ramp metering and traffic signal coordination', *Proceedings of IVHS AMERICA Annual Meeting* .

Porche, I. R. (1998), 'Dynamic traffic control: decentralized and coordinated methods', *Electrical Engineering Systems* .

Richards, P. (1956), 'Shockwaves on the highway', *Opns. Res.* **4**, 42–51.

Robertson, D. (1969*a*), TRANSYT: a traffic network study tool, Technical report LR 253, Transport Road Research Laboratory.

Robertson, D. (1969*b*), 'TRANSYT method for area traffic control', *Traffic Engineering and Control* **11**(10), 276–281.

Robertson, D. & Bretherton, R. D. (1974), 'Optimal control of an intersection for any known sequences of vehicle arrivals', *Proceedings of the 2nd IFAC Triennial World Congress* .

Sadegh, P. (1997), 'Constrained optimization via stochastic approximation with a simultaneous perturbation gradient approximation', *Automatica* **33**, 889–892.

Sadegh, P. & Spall, J. C. (1998), 'Optimal random perturbations for multivariate

stochastic approximation using a simultaneous perturbation gradient approximation', *IEEE Transcations on Automatic Control* **43**(3), 1480:1484.

Sen, S. & Head, K. L. (1997), 'Controlled optimization of phases at an intersection', *Transportation science* **31**, 5–17.

Sheffi, Y. (1985), *Urban transportation networks: equilibrium analysis with mathemetical programming methods*, Prentice Hall, Englewood Cliffs, New Jersey.

Shelby, S. (2001), Design and Evaluation of Real-time Adaptive Traffic Signal Control Algorithms, PhD Thesis, University of Arizona, System and Industrial Engineering Department.

Smaragdis, E. & Papageorgiou, M. (2003), 'A series of new local ramp metering strategies', *Transportation Research Board 82nd Annual Meeting* . Washington, D.C.

Smith, M. (1979*a*), 'The existence, uniqueness and stability of traffic equilibria', *Transpn. Res. -B* **13**, 295–304.

Smith, M. (1979*b*), 'Traffic control and route-choice: a simple example', *Transpn. Res. -B* **13**, 289–294.

Smith, M. (1980), 'A local control policy which automatically maximises the overall travel capacity of an urban network', *Traffic Engineering and Control* **22**, 298–302.

Smith, M. (1981), 'Properties of a traffic control policy which ensure the existence of a traffic equilibrium consistent with the policy', *Transpn. Res. -B* **15**, 453–462.

Spall, J. C. (1988), A stochastic approximation algorithm for large-dimensional systems in the kiefer-wolfwitz setting, *in* 'Proceedings of the 27th Conference on Decision and Control', IEEE, Austin, Texas, pp. 1544–1548.

Spall, J. C. (1992), 'Multivariate stochastic approximation using a simultaneous perturbation gradient approximation', *IEEE Transcations on Automatic Control* **37**(3), 332:341.

Spall, J. C. (1998), 'Implementation of the simultaneous perturbation algorithm for stochastic optimization', *IEEE Transcations on Aerospace and Electronic Systems* **34**(3), 817:823.

Spall, J. C. (2003), *Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control*, Wiley Press.

Stephanedes, Y. J. & Chang, K. (1993), 'Optimal control of freeway corridors', *ASCE Journal of Transportation Engineering* **119**(4), 504–514.

Stephanopoulos, G., Michalopoulos, P. G. & Stephanopoulos, G. (1979), 'Modeling and analysis of traffic queue dynamics at signalized intersections', *Transpn. Res. -A* **13**(3), 295–307.

Stoffers, K. E. (2003), 'Scheduling of traffic lights - a new approach', *Transportation Research* **2**, 199–234.

Szeto, W. & Lo, H. K. (2006), 'Transportation network improvement and tolling strategies: The issue of intergeneration equity', *Transpn. Res. -B* **40**, 227–243.

Tan, H., Gershwin, S. & Athans, M. (1979), Hybrid optimization in urban traffic networks, Final Report DOT-TSC-RSPA-79-7, MIT Laboratory for Information and Decision Systems, Cambridge MA 02139.

Tian, Z. Z., Balke, K., Engelbrecht, R. & Rilett, L. (2002), 'Integrated control strategies for surface street and freeway systems', *Transportation Research Record* **1811**, 92–99.

TRB (2000), *Highway Capacity Manual(CD-ROM)*, Transportation Research Board, National Research Council, Washington, D.C.

van Vliet, D. (1982), 'SATURN - a modern assignment model', *Traffic Engineering and Control* **23**, 578–581.

van Vuren, T. (1990), *Signal control and traffic assignment*, Pergamon Press, pp. 468–472. Concise Encyclopedia of traffic and transportation systems.

van Vuren, T. & van Vliet, D. (1992), *Route Choice and Signal Control*, Athenaeum Press Ltd.

van Zuylen, H. J. (2002), ITS for traffic signal control: Design and evaluation, Technical report, Department of Civil Engineering and Geosciences, Technical University of Delft.

van Zuylen, H. J. & Taale, H. (2003), 'Urban networks with ring roads: a tri-level optimization', *Prsented at proceedings of 83th TRB annual meeting, Washington D.C.* .

Vaughan, R. (1985), 'Equity solutions for trip distribution', *Traffic Engineering and Control* **26**(1), 23–25.

Vincent, R., Mitchell, A. & Robertson, D. (1980), User guide to TRANSYT version 8, Technical Report TRRL Laboratory Report 888, TRRL Department of the Environment.

Wallace, C., Courage, K., Hadi, M. & Gan, A. (1998), TRANSYT-7F User's Guide, Technical report, U.S. Federal Highway Administration, Washington D.C.

Wardrop, J. G. (1952), 'Some theoretic aspects of road traffic research', *Proceedings of the Institute of Civil Engineering II* pp. 278–325.

Webster, F. (1958), Traffic signal settings, Technical report 39, Transport Road Research Laboratory.

Wright, A. (1991), *Genetic algorithms for real parameter optimization*, Foundations of Genetic Algorithms, Morgan Kaufmann Publishers, San Mateo, California, USA.

Yang, H. & Bell, M. H. (1998), 'Models and algorithms for road network design: A review and some new developments', *Transport Review* **18**(3), 257–278.

Yang, H. & Yagar, S. (1995), 'Traffic assignment and traffic control in saturated road networks', *Transpn. Res. -B.* **29**(2), 125–139.

Yin, Y., Liu, H. & Benouar, H. (2000), 'A note on equity of ramp metering', *The 7th IEEE International Conference on Intelligent Transportation Systems* .

Yoshino, T., Sasaki, T. & Haegawa, T. (1995), 'The traffic control system on the Hanshin expressway', *Interfaces Magazine* **Jan/Feb**.

Zhang, L. & Levinson, D. (2003), 'Balancing efficiency and equity of ramp meters', *82nd TRB Annual Meeting* . Washington, D.C.

Zhang, L. & Levinson, D. (2004), 'Optimal freeway ramp control without origin-destination information', *Transpn. Res. -B* **38**, 869–887.

Zhang, M. H. (2001), Evaluation of on-ramp control algorithms, Research Report UCB-ITS-PRR-2001-36, University of California, Davis, Irvine.

Zhang, M. H. & Richie, S. (1997), 'Freeway ramp metering using artificial neural networks', *Transportation Research C* **5**(5), 273–286.

Ziliaskopoulos, A. K. (2000), 'A linear programming model for the single destination system optimum dynamic traffic assignment problem', *Transportation Science* **34**, 200–207.

Appendix A

Acronyms

ATIS:	*Advanced Traffic Information System*
ATMS:	*Advanced Traffic Management System*
CTM:	*Cell Transmission Model*
d-ENDP:	*Dynamic Equilibrium Network Design Problem*
DNL:	*Dynamic Network Loading*
DTA:	*Dynamic Traffic Assignment*
DTC:	*Dynamic Traffic Control*
DSO:	*Dynamic System Optimal*
ENDP:	*Equilibrium Network Design Problem*
ICC:	*Integrated Corridor Control*
ITS:	*Intelligent Transport System*
LSC:	*Local Synchronization Control*
NDP:	*Network Design Problem*
SPSA:	*Simultaneous Perturbation Stochastic Approximation*
VMS:	*Variable Message Sign*