Institute of Transportation Studies
University of California at Berkeley

Theory of Highway Traffic Signals

Gordon F. Newell

RESEARCH REPORT
UCB-ITS-RR-89-7

Sponsored by the California Department of Transportation and the Federal Highway Administration, RTA-13945-55E014. The contents of this report reflect the views of the author, who is responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California or the Federal Highway Administration. The report does not constitute a standard, specification, or regulation.

April 1989
ISSN 0192 4085

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA/CA/UCB-ITS-RR-89-7		

4. Title and Subtitle	5. Report Date
Theory of Highway Traffic Signals	October 1988
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
G. F. Newell	UCB-ITS-RR-89-7
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Institute of Transportation Studies University of California Berkeley, California 94720	11. Contract or Grant No. F86T012
	13. Type of Report and Period Covered
12. Sponsoring Agency Name and Address FHWA Calif. Dept. of Transportation US DOT 1120 N Street Washington, D. C. Sacramento, California 95814	Final Report
	14. Sponsoring Agency Code

15. Supplementary Notes

This research was funded jointly by the U. S. Federal Highway Administration and the California Department of Transportation.

16. Abstract

This report gives a comprehensive survey of the theory of highway traffic signals including isolated signals, one-way arterials, two-way arterials, and networks, limited, however, to the common right angle junctions. The emphasis is on the logistics of control strategies rather than recipes for implementation. It is anticipated, however, that the implementation of some of the strategies described here will give substantial reductions in delays as compared with existing procedures.

17. Key Words	18. Distribution Statement
Traffic Signals	No restrictions

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	453	

Implementation statement

Copies of this report will be distributed to some districts of the California Department of Transportation.

Financial disclosure statement

This research was funded jointly by the State of California Department of Transportation and the U. S. Federal Highway Administration. The total ceiling amount of this contract, F86T012, was $13,677.

CONTENTS

PREFACE

This treatise on the theory of highway traffic signals is the culmination of work which began some 35 years ago. I was first exposed to the subject of transportation theory as a result of attending a couple of seminars at Brown University about 1953 given by the late Professor William Prager. He had also assembled a nearly complete bibliography on existing traffic theory, which, however, consisted essentially of just one paper, the classic paper by John Wardrop (1951), which summarized what had been done up to that time, mostly at the Road Research Laboratory in England. This is the paper which contains the famous "Wardrop equilibrium conditions" for traffic assignment, but it also contained some of the preliminary results on delays at a fixed cycle traffic signal, later described in more detail in the famous works of F. V. Webster (Webster's formula).

Queueing theory was also in an early stage of development at that time, but the mathematical techniques available then were not well suited to providing an analytic representation of the queues at a fixed-cycle signal. Webster's formula was derived from a mixture of crude theory supplemented by curve fitting to results of simulations. One of my first challenges was to try to obtain a more "elegant" alternative to Webster's formula.

I did not succeed in solving this problem until about 1964, by which time I had also become involved in other aspects of traffic flow theory, including some aspects of signal coordination. The analytic solution of this fixed cycle signal problem was not, in itself, a major achievement; Webster's formula was quite adequate for practical purposes, but it did represent a significant departure from traditional methods in queueing theory. The introduction of "deterministic" and "diffusion" approximations led to a technique of analysis for queueing problems from which one could attack a wide variety of more complex practical problems including vehicle-actuated signals. The emphasis during this

period, however, was more on mathematical techniques and "cute" results rather than practical applications.

In 1965 I interchanged my hobby (transportation theory) and my profession (applied mathematics) and joined the transportation engineering faculty at the University of California, Berkeley. Although I continued to work on traffic signal problems for a while, I gradually got drawn into the complete spectrum of theoretical problems in transportation, all modes and aspects. I did teach part of a course devoted to traffic signals for several years in the early 1970's, but this course was dropped when enrollment by students of transportation engineering declined in the mid 1970's.

In 1984, Masao Kuwahara complained that there were not enough "advanced" courses in transportation. We managed to assemble a group of about nine students, many of whom were working on various aspects of computer modeling of traffic control, to participate in a special topics course on the theory of highway traffic control, mostly traffic signals but also some freeway ramp control. At that time I had not actively worked on traffic signal problems for about ten years, so it was a unique challenge to look at the subject again from a fresh point of view. In reviewing the literature, however, I was rather disappointed that hardly any advances had been made in the theory during the previous ten years and, indeed, in some respects the subject had gone backwards.

previous ten years and, indeed, in some respects the subject had gone backwards.

I assembled a collection of scribbled lecture notes of what I could put together on short notice, essentially an expansion and reinterpretation of notes which I had used in the early 1970's in a shorter course. Having already invested considerable time to get back into the subject again, I decided to embark on the present project of writing a systematic treatise on just about everything I thought was worth doing on the theory of traffic signals. Even in the 1970's I believed that there was no longer any major theoretical obstacles to the analysis of traffic signal systems. It was simply a matter

of systematically describing the logical consequences of existing theory. I expected it to be tedious but straightforward. As it turns out, maybe it wasn't quite as straightforward as I thought, however.

I had a sabbatical leave for the 1985-6 academic year and started the project in the summer of 1985. I spent the fall semester of 1985 at Rensselaer Polytechnic Institute and the spring semester 1986 "in residence" at Berkeley. The plan was first to put aside almost everything that had been done before and start over. I completed a preliminary draft of the first three chapters by the end of the summer 1986 and circulated it among some friends and colleagues for comments.

Progress slowed considerably after that when I went back to full time teaching, but I managed to finish a draft of Chapter 4 by the end of the summer 1987. In the fall, 1987, there were again enough new graduate students interested in traffic signals to give a special topics course, but this time I had the completed draft of Chapters 1 to 4. I was able then, while teaching from the first draft, to start a second draft of Chapters 1 to 3. The project was finally completed in summer of 1988.

It is a pleasure to acknowledge the assistance I have received during the last three years from people who made comments on earlier drafts and who answered my endless questions about what practitioners do.

First, I appreciate the patience of two classes of students who endured swered my endless questions about what practitioners do.

First, I appreciate the patience of two classes of students who endured being told sometimes the opposite of what they had been told by others. RPI, and particularly Professor Pitu Mirchandani, offered a very pleasant atmosphere for me to do my work during the fall 1985.

Very helpful comments on early drafts were obtained from Van Hurdle, Chan Wirasinghe, Harold Garfield (Caltrans), Frederick Rooney (Caltrans), Robert Shanteau, Dennis Robertson (TRRL, England), and Paul Ross (Federal Highway Administration). Needless to say, some of these people were less

than enthusiastic about what I had done and do not necessarily subscribe to all the views expressed here. Particular thanks, however, go to Van Hurdle, Robert Shanteau, I. Jeeva, and Alex Skabardonis who spent many hours sharing with me their experiences in the real world applications, and explaining to me current practice and what is in some of the computer programs for traffic control.

Caltrans gave some financial assistance mostly for secretarial help, and Frederick Rooney has given me considerable encouragement. The one person who never argues is Phyllis De Fabio, who patiently types whatever I give her.

The purpose of this project was not just to carry out an academic mathematical exercise. The purpose was, in part, to bridge the gap between the theory (much of which was, indeed, rather academic), and the real world of traffic control (much of which is based upon misguided logic). The style is certainly not that of a Traffic Engineering Handbook or the Highway Capacity Manual. It is not a recipe book.

Unfortunately, most traffic engineers are trained to interpret formulas as something in which one substitutes numbers and are, therefore, a clumsy substitute for a computer program which can carry out calculations at the push of a button. To me, however, a formula is a "short hand" notation which describes a cause and effect relation among physical observables. Formulas are just part of the text which tries to explain something.

In the future, I plan to concentrate on some specific situations in which the present theory conflicts seriously with what is described in handbooks or computer programs, translate the formulas into numerical examples, show where I believe that current practices are deficient, and give some alternative recipes. At the present stage, however, my goal was to investigate the possible consequences of just about anything which may or may not be relevant to traffic control, limited, however, to the common right angle two or four

directional intersection for one-way or two-way streets (although, perhaps, T junctions could be considered as a special case). Before I start to challenge the profession, I wanted to be sure that no one could say: but you did not consider this or that effect. I believe that I have looked at just about everything which might be relevant to the theory. That is why this treatise is so long; it includes an analysis of many things which, in fact, are typically not important in order to show why they are not (or when they might be) important.

It is not possible to summarize here all the things which are discussed in this report. I will confine my comments to those things in the text which are at variance with current trends.

Chapter 1 gives a brief review of traffic flow theory and queueing theory. The main purpose here is to identify those aspects of the theory which are relevant to traffic control and (more important) which are not. Much of the theory of signal control is independent of the details of how people drive. This is fortunate, because existing theories of traffic flow are notoriously unreliable. About the only information one needs to know accurately is how the number of vehicles leaving an intersection per cycle in any direction (at saturation) depends on the phase interval. This one can observe directly in the field, or one can use some empirical recipes to relate this to lane width, percent of trucks, etc., as described in the Highway Capacity Manual.

pends on the phase interval. This one can observe directly in the field, or one can use some empirical recipes to relate this to lane width, percent of trucks, etc., as described in the Highway Capacity Manual.

Any computer simulation models which describe the motion of individual vehicles requires all sorts of irrelevant input data and gives estimates of delay which are no more accurate than the accuracy with which the model can reproduce the actual output from a signal. The actual output, however, can be measured directly with greater accuracy and less effort than is required to measure all the parameters in the simulation models. For a coordinated signal system, one will also need to know some appropriate trip times between

intersections, but, again, it is easier to observe what is relevant than to evaluate it from some questionable model.

Chapter 2 gives a very detailed development of the theory of the four way isolated intersection (uniform stochastic arrivals) with no turning traffic for fixed-cycle (F-C) and vehicle-actuated (V-A) control (devoid, however, of derivations of the queueing formulas). It then goes on to describe extensions of this theory including turning traffic with or without turn signals.

It would be much too tedious to describe in detail how the delays to all vehicles at even a single intersection depend on the arrival and departure rates for all traffic movements (including turning traffic) and on the signal strategy. I have tried, however, to describe enough of the theory to suggest that it is, in principle, straightforward, and that the practical issues are well defined. I do not see any serious problem in describing anything one would wish to know about how the performance of any isolated intersection depends on the signal strategy and any parameters characterizing the system.

This chapter contains some minor extensions and variations of existing theories for the F-C signal, but there is probably no serious conflict here with current recommended procedures. There is, however, some serious disagreement with common procedures for the operation of V-A signals serving moderately heavy traffic (several vehicles served per phase). The delay for a V-A signal is very sensitive to whether or not a signal phase terminates promptly when the queue vanishes in some appropriate direction. If one does not terminate the phase promptly, the advantage of the V-A signal over a F-C signal may be entirely lost. Indeed most signal phases may run to the maximum extension so that the signal actually behaves like a F-C signal.

To terminate a phase promptly, the location and type of vehicle detectors is very important. Unfortunately, the traffic engineering literature regarding the location of detectors and the strategy for terminating signal

phases is completely chaotic and devoid of much logic, particularly for multi-lane approaches. My expectation here is that a properly designed V-A signal system could typically reduce the delays by perhaps 1/2 as compared with current practice (with simple equipment, no microprocessors, artificial intelligence, or whatever).

Chapter 3, dealing with coordination of signals on a one-way arterial, is also rather long because it gives the background for much of the analysis for two-way arterials and networks in Chapters 4 and 5.

For a pretimed strategy, one should first identify the most critical intersection and choose the cycle time C and splits more or less as one would if this intersection were an isolated intersection. Other signals may operate on a cycle time C or $C/2$, but, in any case, signals downstream of the critical intersection should be set so that a vehicle leaving the critical intersection at the end of the green interval will barely clear the downstream intersections during the green interval, i.e., the "off-sets" should be based on the termination of the green interval (not the start). The length of the green intervals downstream should then be chosen so as to accommodate any vehicle which can pass the critical intersection, with appropriate adjustments for changes in flow due to vehicles turning onto or off the arterial.

Similarly, signals upstream of the critcal intersection should be set so to vehicles turning onto or off the arterial.

Similarly, signals upstream of the critcal intersection should be set so that a vehicle leaving an upstream intersection at the end of a green interval will arrive at the critical intersection at the end of the green. The duration of the green intervals should be barely enough so as virtually to guarantee that the critical intersection is kept busy whenever there is a stochastic queue at any upstream intersection.

This strategy is perhaps not much different from what a traffic engineer might devise as a result of "fine-tuning" some scheme in the field, but it is not necessarily consistent with what would result from some computer programs.

Most computer programs grossly overestimate the amount of stochastic queueing and perhaps also the "platoon spreading." They do not even recognize the very important fact that the fluctuations in the number of vehicles which pass one intersection are highly correlated with the number which pass neighboring intersections. Estimated reductions in delays evaluated from these computer programs may be due mostly to a reduction in a fictitious stochastic delay at noncritical intersections.

It is possible to reduce considerably the (actual) stochastic delay of a pretimed strategy by means of traffic responsive strategies, but by a scheme which is almost the opposite of what some people have proposed. To do so one should start the arterial green interval according to a pretimed plan, promptly terminate the arterial green (particularly at the critical intersections) as soon as the platoon passes (or some preset maximum time expires), and give any excess time to the cross street. Most other proposed schemes terminate the cross street green when the queue vanishes on the cross street and gives the excess time to the arterial.

Chapter 4 treats a single two-way arterial and begins with a discussion of conditions under which one can provide through bands in both directions wide enough to accommodate specified flows in the two directions (which is usually not possible). Most of the analysis, however, is directed toward wide enough to accommodate specified flows in the two directions (which is usually not possible). Most of the analysis, however, is directed toward strategies in which one provides a progression in one direction, as for the one-way arterial, and then partitions the remaining time between the cross street and the opposing arterial direction in some advantageous ways. Much of the theory is, therefore, an extension of that described in Chapter 3.

Chapter 5, dealing with networks, discusses the constraints induced by "loop conditions," and certain idealized geometries of square or rectangular street grids in which one can have a good progression on all streets or at least some subnetwork. Most of the discussion is descriptive rather than

analytical, however, and emphasizes the view that the distribution of traffic over the network will depend on the signal coordination scheme. One should choose a scheme so as to induce the traffic to disperse and utilize all the facilities provided for it. I do not subscribe to the view that one should try to minimize the delays (or other objective) subject to given values of the flows.

1. PRELIMINARIES

1.1. Introduction

Ideally, a scientific approach to the theory of highway traffic signal control should start from some "fundamental equations" describing the interaction between cars and their dependence on any external control system. Given some possible objective function, one would then try to determine signal settings which minimize some "cost" of travel. Unfortunately we do not have any "fundamental equations" for traffic behavior. We have a vast collection of empirical data, much of which is difficult to interpret. Neither do we have a well-defined "objective function." It is not clear what society wants, in particular how to balance one person's gain against another person's loss.

The system exists, however, and, despite the fact that there is no possibility of formulating a highly precise logical framework, the traffic engineer is expected to make the system work in a socially acceptable way.

There are certain causes and effects, and there are at least qualitative goals and issues. Fortunately these goals are often rather insensitive to the detailed behavior of individual cars, which is very difficult to describe anyway. Much of the art of traffic signal theory involves identifying only those aspects of traffic behavior which are relevant to the questions being asked, and disregarding those aspects which are not relevant.

The purpose of any "theory" is to predict an outcome of some experiment asked, and disregarding those aspects which are not relevant.

The purpose of any "theory" is to predict an outcome of some experiment which has not been done. In the present context we might wish to describe how traffic would move through some signalized network which is being designed but has not been built. More likely, however, we would be concerned with how the traffic behavior in an existing system would change if we should modify the control strategy. The latter type of problem is typically much easier than the former because one can make direct observations on the existing system, measuring transit times between intersections, rates of acceleration, etc;

anything that might be relevant but, in particular, certain aspects of the traffic behavior which we expect do not depend upon the control strategy. One need then theorize only about those aspects of the traffic behavior which do change with the control strategy. One typically does not need a "comprehensive theory of traffic flow" for this purpose.

1.2. Time-Space Diagrams

If one wished to analyze the behavior of traffic in an existing network, most aspects of the system could be recorded by taking a moving picture from an aircraft. Although such a record may not show the height of vehicles or other measurements in a direction perpendicular to the two-dimensional surface of the road network (which we assume to be irrelevant to any question which we will pose), it would include a vast amount of irrelevant or redundant information.

Presumably, during any period of observations, the roads and buildings do not move. Anything that does not change with time can be observed, measured, identified, etc., from a single picture frame and erased from any subsequent pictures. If a vehicle appears on many successive picture frames, any physical features of the vehicle (it is a car, truck, bus, etc.) can also be identified from a single frame. In any subsequent pictures, it suffices to identify this vehicle by some number j and to identify its "position" by the location of from a single frame. In any subsequent pictures, it suffices to identify this vehicle by some number j and to identify its "position" by the location of some identifiable reference point (the middle of its front bumper, for example). If one introduces a rectangular coordinate system (x, y) in the two-dimensional plane, the location of the jth vehicle at time t can be represented by a point $(x_j(t), y_j(t))$.

If one introduces a three-dimensional space (t, x, y), one can describe the motion of the jth car by drawing a "trajectory," i.e., a curve $(t, x_j(t), y_j(t))$. One can easily construct such a three-dimensional curve by

separating each picture frame of the movie picture and stacking them on top of each other with a sheet of glass between each frame. The trajectory of vehicle j is now drawn by passing a smooth curve through the identifying point of the jth vehicle in each picture frame.

Traffic signals do not physically move; their actual locations can be identified from a single picture frame. On any subsequent frames it suffices merely to identify any relevant aspects of the signal (red, yellow. green) by means of any code; for example, one could put a point on the film at some reference location if the signal is red but no point if it is green. It makes no difference where one writes the code on the film as long as one knows what the code means.

In the (t, x, y) space it is possible to draw the trajectories of all vehicles $j = 1, 2, \ldots$ and the time-dependent aspects of all signals. This, along with any information about the geometry of roads, physical characteristics of the vehicles, etc., which can be recorded separately or seen on a single film frame, will give a complete description of everything that is happening.

There are certain situations in which it may be useful to sketch some trajectories in this three-dimensional (t, x, y) space, particularly if one has some complicated turning movements at intersections or one wishes to illustrate some network synchronization patterns. Most of the time, however, has some complicated turning movements at intersections or one wishes to illustrate some network synchronization patterns. Most of the time, however, we will deal with two-dimensional trajectory curves. Although vehicles do move on a two-dimensional physical surface, they are usually constrained to move along certain channels (highway lanes). If one wished to illustrate how vehicles stay in their lanes, the details of how they change lanes, or various conflicts for turning movements, then one would want to analyze movements transverse to the direction of the channel. If one is not concerned about these things, however, one could introduce a separate coordinate system for each

channel with one of the coordinates measured along the center line of the channel. The other coordinate would be transverse to this direction but, if we are not concerned about movements in this direction, the "location" of a vehicle can be identified by a single coordinate, its longitudinal position.

If one is concerned with the movement of vehicles in some network of roads but not with the details of how vehicles stay in lanes or switch lanes (including turning movements), one could observe from a single photograph the geometry of the network, label (number) whatever one decides to interpret as "channels," and identify where and which movements are permitted between channels (lane changes or turning movements). These are presumably all time-independent (except that the times at which turns are allowed will depend on the signal phase). Instead of identifying the position of a jth vehicle at time t by two continuous variables $x_j(t)$ and $y_j(t)$, we now identify it with only one continuous variable, which we will also label as $x_j(t)$, although this symbol will now mean a coordinate along the channel, and a discrete variable $\ell_j(t)$ identifying which channel the vehicle is in at time t.

Instead of drawing trajectories of vehicles $(t, x_j(t), y_j(t))$ in a three-dimensional space (t, x, y) we now draw them in a space (t, x, ℓ) of two continuous variables (t, x) and one discrete variable ℓ. The geometric meaning of x depends on ℓ so there is no reason to think of ℓ as a discrete version of the "y". Rather, we interpret the space (t, x, ℓ) simply meaning of x depends on ℓ so there is no reason to think of ℓ as a discrete version of the "y". Rather, we interpret the space (t, x, ℓ) simply as a finite collection of two-dimensional (t, x) spaces. Indeed, if we are concerned about the movement of vehicles in a single channel, we may not explicitly specify the ℓ but assume that its value is understood.

In the analysis of traffic signals we will make extensive use of two-dimensional trajectory curves, so it is important that we be able to "read" and interpret them. One simple way to construct vehicle trajectories for a

straight channel is to project a movie film onto a sheet of graph paper so that the channel projects in the vertical (x) direction of the graph paper as in figure 1.1. Make a dot on the graph paper at the identifying point of

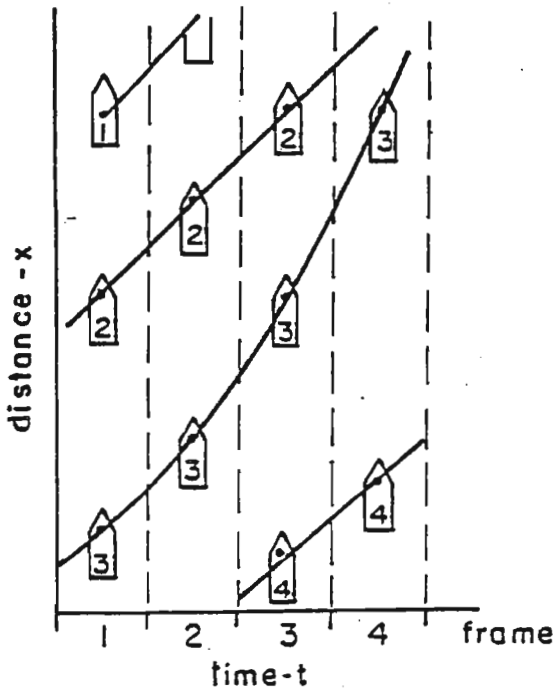


Fig. 1.1 - Construction of trajectories from movie film frames.

each vehicle, i.e., for fixed t (picture frame) one marks the points $x_j(t)$ on the film for each vehicle j . One may wish also to label each point with the vehicle number j . Now translate the graph paper a unit distance (time) to the left, project the next picture frame on the paper and mark the new locations of each vehicle. Repeat this for each film frame and then draw a smooth curve through the points for the same vehicle.

To "see" the movie picture one simply reverses the above construction. Place a sheet of paper with a narrow vertical slit over the graph paper and imagine that each line segment of a trajectory crossing the slit is a vehicle. Now keep the slit stationary and slide the graph paper to the left at a constant

speed (or keep the graph paper fixed and move the slit to the right). The "vehicles" will now appear to move down the road.

It is advantageous that one should learn to read a trajectory plot in this way to take advantage of one's natural familiarity with visual perception. It may even be advantageous that one has eliminated from the trajectory plot all irrelevant information of color, size, etc., so that one sees only the motion of the "vehicles." The main advantage of a trajectory plot, however, as compared with the original movie picture is that one can measure the speed of the vehicle.

$$v_j(t) = dx_j(t)/dt$$

at any time t directly from the graph as the slope of the trajectory. Thus one can also "see" the speed.

For multi-lane highways one can use the same coordinate x to measure the longitudinal position in all lanes of the highway simultaneously (even for two-directional traffic). One typically will want to identify which lane a vehicle is in and could do this by drawing separate trajectory curves for each lane. If a vehicle switches lanes, its trajectory will suddenly terminate in one (t, x) graph, but it then must continue in a second (t, x) graph. If one superimposes the graphs for the two lanes the superimposed trajectory pieces will form a continuous curve, the same curve as one would obtain if one thought one superimposes the graphs for the two lanes the superimposed trajectory pieces will form a continuous curve, the same curve as one would obtain if one thought of the two lanes as a single lane. Alternatively one can identify lane numbers on a single graph by using solid or broken lines, color codes, etc., for the trajectory segments. Vehicles traveling in opposite directions can be identified by the fact that their trajectories will have positive or negative slopes. Typically for heavy traffic it will be advantageous to draw the trajectories separately for different lanes because lane changing will be relatively rare and the dominant interaction between vehicles is between successive vehicles

in the same lane. For light traffic, most traffic will be in the outer lanes and it would typically be convenient to show passing vehicles in the same (t, x) graph. If one draws graphs for each lane on transparent paper, one can, of course, overlay the graphs or not as one pleases.

Figures 1.2 and 1.3 show some hypothetical trajectory plots. In figure 1.2 there are traffic signals at (or near) lane coordinates y_1 and y_2. The time-dependence of these signals is indicated by drawing a horizontal line segment when the signal is red and nothing when it is green (a yellow interval could be represented by a dotted line segment if it is relevant). That a vehicle cannot pass the signal when it is red means that the vehicle trajectory cannot intersect the horizontal line segments at y_1, y_2 ... . In this figure vehicle 1 approaches the signal at y_1 during a red interval and comes to a complete stop (the trajectory has a horizontal segment). After the signal turns green, this vehicle

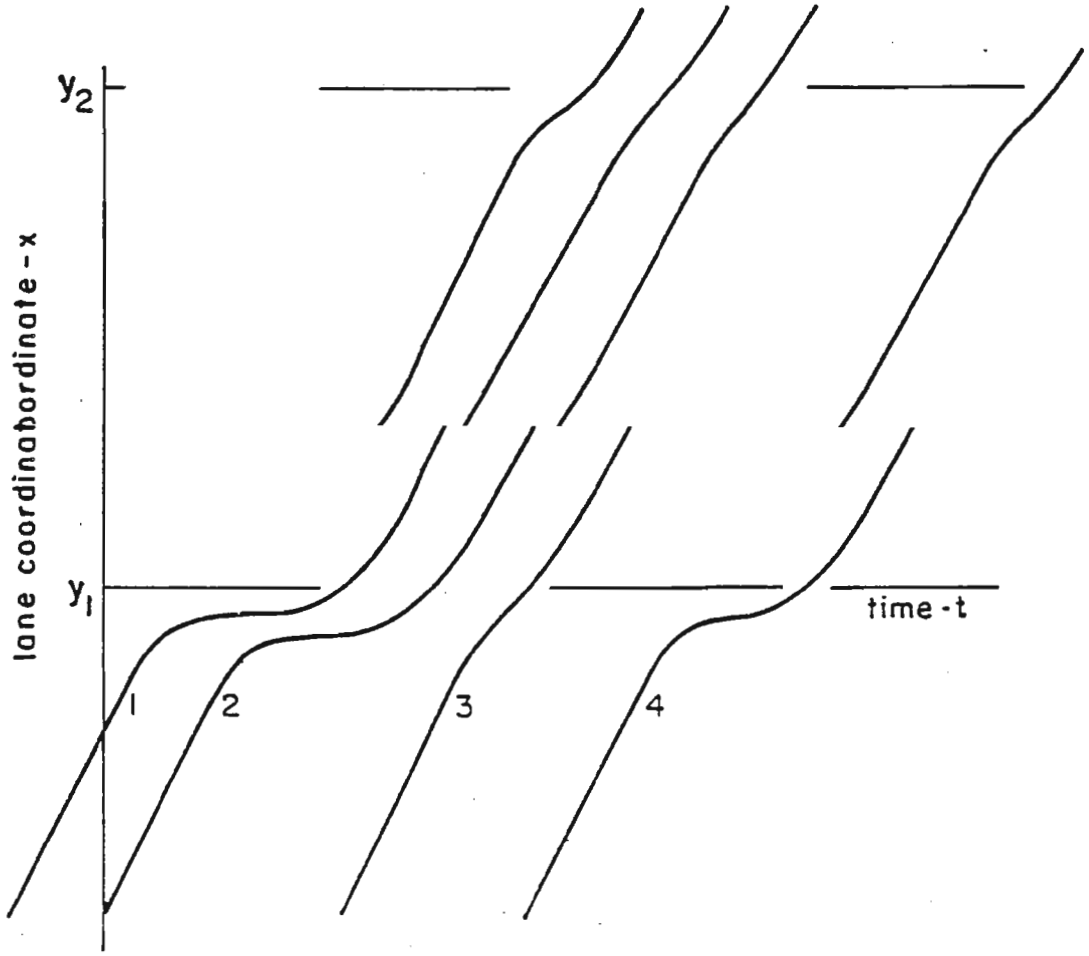


Fig. 1.2 - Trajectories for vehicles passing through traffic signals.

accelerates (the trajectory slope increases) and moves toward the signal at y_2. At the second signal the vehicle decelerates but does not come to a complete stop. As shown, vehicle 2 also stops behind the first signal but at a location slightly upstream of the first vehicle. Vehicle 3 decelerates as it approaches the stopped vehicles but does not come to a complete stop. This trajectory plot is typical of vehicles following each other in a single channel with no passing.

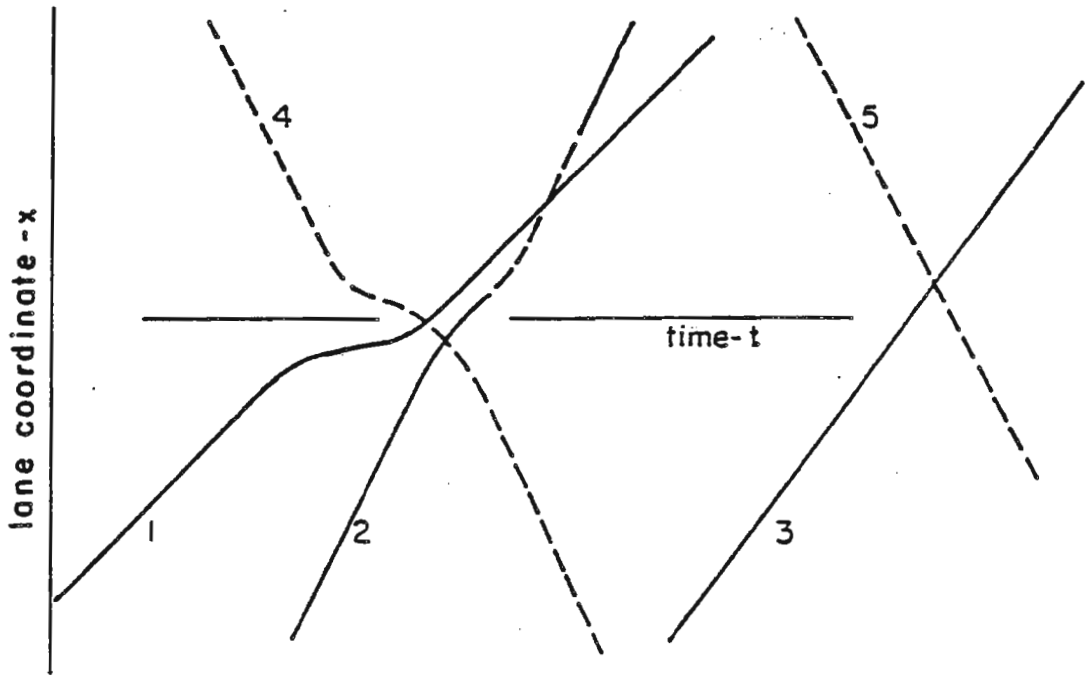


Fig. 1.3 - Typical trajectories for light traffic.

Figure 1.3 shows a typical pattern for light traffic. Vehicle 1 is stopped by the signal. Vehicle 2 is traveling faster and overtakes vehicle 1 but cannot

Figure 1.3 shows a typical pattern for light traffic. Vehicle 1 is stopped by the signal. Vehicle 2 is traveling faster and overtakes vehicle 1 but cannot pass in the same lane. The broken curve segment of the trajectory for vehicle 2 means that it switched lanes to pass vehicle 1 and then returned to the original lane. Vehicle 3 passes the signal with no interference from the signal or other vehicles (it has a constant slope). Vehicle 4 is moving in the reverse direction (negative slope) and in a separate lane. If there are only two lanes, vehicle 2 would be using the same lane to pass vehicle 1 as is used by vehicle 4. Since

vehicles 2 and 4 cannot be in the same place at the same time, the broken line segments for these two trajectories cannot intersect.

Although these trajectory plots are representative of what one might see in a real traffic pattern and illustrate some of the rules of traffic behavior, we do not interpret this as a "theory." It is meant as an illustration of what one might observe in an existing pattern rather than a prediction of what will happen in some proposed system. The rules of what one can or cannot do, however, are sufficiently strict that one can see many qualitative features of any proposed signal system even if one cannot predict in detail the trajectories of individual vehicles.

1.3. Point Observations, Cumulative Counts

In the previous section we assumed that one observed traffic by taking a moving picture, each picture describing the location of vehicles at a fixed time; a vertical slice in the (t, x) space. Since it is often difficult to find a place where one can take pictures and tedious to analyze them if one can take them, traffic engineers are more likely to observe traffic by recording events over time at one or more fixed locations, a horizontal slice in the (t, x) plane. Potentially, one could record, at any location x, the time $t_j(x)$ at which vehicle j passes x, and any relevant physical characteristics of the vehicle.

One particularly convenient way of representing this data graphically at some location x is to draw a curve of

$$n(t, x) = \text{cumulative number of vehicles to pass } x \text{ by time } t \tag{1.3.1}$$

as in figure 1.4. If we number the vehicles consecutively so that

$$0 \leq t_1(x) \leq t_2(x) \leq \dots , \tag{1.3.2}$$

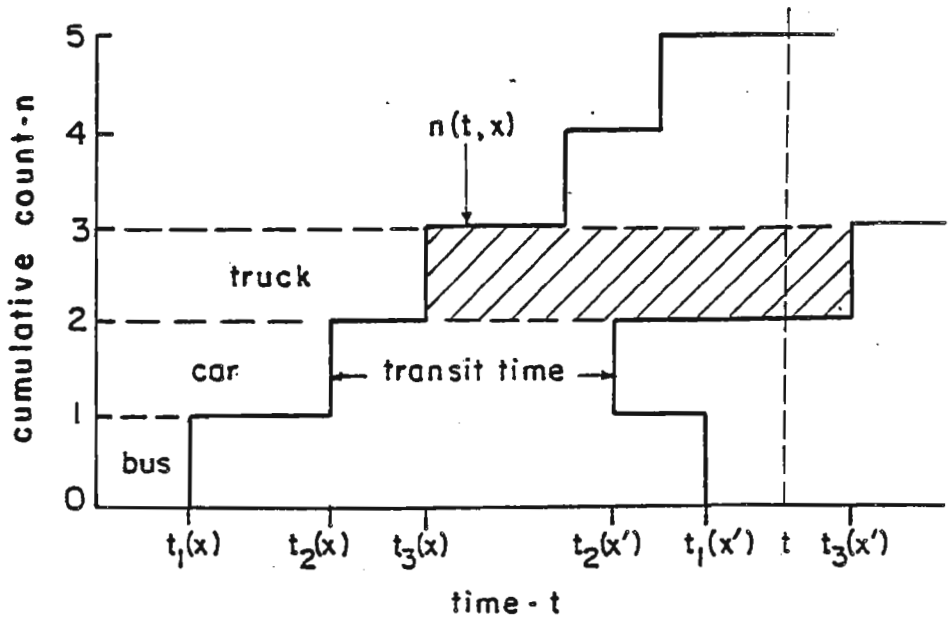


Fig. 1.4 - Graphical representation of events at two locations.

$n(t, x)$ is a step function which increases by one unit at each time $t_j(x)$ when a vehicle passes x. On a trajectory plot, the $t_j(x)$ are the times at which trajectories cross a horizontal line at height x and $n(t, x)$ is the total number of crossings from some time 0 to time t.

Although we will sometimes find it convenient to think of the function $n(t, x)$ for fixed x as if n were the "dependent variable" and t the "independent variable," at other times it will be convenient to think of the "inverse function" relating t to n. Essentially in figure 1.4 we identify a horizontal strip of unit height with a particular vehicle. We can even write on this strip any identifying features besides the counting label which is already assigned to it; for example, "car", or "bus," etc. On this strip we insert a vertical line segment at the time $t_j(x)$ when that vehicle passed x.

Actually this inverse relation is perhaps more basic than (1.3.1), because we could potentially number the vehicles in any order whatsoever and

still draw the inverse graph showing the times $t_j(x)$ at which the jth vehicle passes x as a function of the "count" j (actually the label). The special feature of the graph identified with the ordering (1.3.2) is that $n(t, x)$ is a monotone nondecreasing function of t.

Suppose now that we have two observers located at positions x and x', $x' > x$ observing vehicles traveling in the direction from x to x'. It is possible that vehicles could pass each other in traveling from x to x' or that some vehicles could leave or enter the road section between x and x'. If there are no exits or entrances between x and x', then any vehicle which passes the observer at x must also pass the observer at x' at a later time. If the first observer places an identifiable label on each vehicle as it passes him (the number j, for example), then the second observer could note the time $t_j(x')$ when this same vehicle passes him and draw a vertical line segment at time $t_j(x')$ in the jth strip of figure 1.4.

We now see one of the advantages of drawing a graph such as figure 1.4. The transit time of the vehicle from x to x' is $t_j(x') - t_j(x)$. This subtraction can be done graphically and is identified with the horizontal "distance" between the vertical segments at times $t_j(x)$ and $t_j(x')$.

One can also readily evaluate the count of vehicles between the observers at x and x' at time t from figure 1.4. If $t_j(x) < t < t_j(x')$ (and no vehicles can enter or leave between x and x'), this means that the vehicle at x and x' at time t from figure 1.4. If $t_j(x) < t < t_j(x')$ (and no vehicles can enter or leave between x and x'), this means that the vehicle j has passed x by time t, but not x'. It must, therefore, be between the two observers. Geometrically, if we shade in the strip between times $t_j(x)$ and $t_j(x')$, and a vertical line at time t (broken line) crosses the shaded area, then vehicle j is between the two observers. One can also add line segments graphically. If we shade in all horizontal strips between the $t_j(x)$ and $t_j(x')$, the total number of vehicles between the observers is the total

height of shaded area cut by the vertical line at time t.

If the first observer numbers the vehicles in order as in (1.3.2) and vehicles do not pass each other, they will also pass the second observer in the same order. The times $t_j(x')$ will then define a monotone nondecreasing curve $n(t, x')$ and the number of vehicles between the two observers at time t will be

$$\text{number in road section} = n(t, x) - n(t, x') \tag{1.3.3}$$

which can be identified as the vertical distance between the curves $n(t, x)$ and $n(t, x')$ at time t.

Some of the graphical interpretations above can be generalized for situations in which vehicles may leave or enter the road section. The problem here is that we were trying to number the vehicles in some logical order, but this is rather difficult if some vehicles disappear (from our road section). We could number, in any arbitrary order, all vehicles which may at some time appear in our road section and assign a horizontal strip as in figure 1.4 to each vehicle. We can also mark vertical line segments at times $t_j(x)$ and/or $t_j(x')$ if the jth vehicle passes both or either observer. If a vehicle does not pass either observer, we just have an empty strip, which is no problem. If a vehicle passes both observers, we can deal with it as before. But if a vehicle passes only one observer, we obviously cannot evaluate a transit time $t_j(x') - t_j(x)$. If we knew when the vehicle entered or left the section, we could at least shade in a time strip when the vehicle is in the section and keep account of the number of vehicles in the section.

If we had observers located at arbitrarily close spacing (essentially at every x), we could record everything that happens to every vehicle at all times. We could construct a complete trajectory plot in essentially the same way as was done from a movie picture except that the role of x and t would be reversed.

1.4. Theory of Traffic Flow

Over the last 30 years or so many attempts have been made to develop "theories of traffic flow." In principle, it should be possible to develop a theory which would predict the time-dependent behavior of every vehicle if one knew when each vehicle entered the network, the type of vehicle, and certain behavior characteristics of each driver. Since different drivers behave in different ways (and individual drivers behave differently at different times), one would not care to describe in detail how the motion depends on which drivers are selected. In any such theory, one would try to predict only certain average system behavior, averaged over some random selection of drivers and times when they entered the network.

If one had such a theory, it would obviously be computationally quite tedious to calculate the (average) motion of the vehicles, but that is not the problem. The problem is that there are so many different types of maneuvers (passing, merging, following, etc.) which drivers perform that any comprehensive theory designed to describe all effects which might be relevant in all situations would require so much "input data" (parameters which must be observed) that it would be virtually impossible to apply. It is particularly important in analyzing traffic in signalized networks that one does not try to describe anything more than is necessary to answer specific questions. We will introduce various types of specific models for vehicle motion as they are needed in the context of more than is necessary to answer specific questions. We will introduce various types of specific models for vehicle motion as they are needed in the context of specific problems, but we can make a few general comments here.

If some vehicles are stopped at an intersection as in figure 1.2, let t_1, t_2, ... represent the times at which the vehicles pass this intersection relative to the start of the green. These times will depend upon the vehicle types and characteristics of the driver (which we will typically model by sampling at random from some population), but it is reasonable to postulate that the times t_1, t_2 for cars which are stopped during the red time and pass during the

subsequent green time do not depend on

a) the duration of the red interval,

b) the duration of the subsequent green interval, or

c) the shape of the trajectories of these vehicles upstream of the signal.

This refers only to those vehicles which are stopped. Of course, the number of vehicles which are stopped depends on the duration of the red interval, and whether or not they will pass the intersection during the subsequent green interval depends on the duration of the green interval. Even if a vehicle does not come to a complete stop but must decelerate in order to follow a preceding vehicle, we still expect that the crossing time of the vehicle will be nearly the same as if it had been stopped.

Furthermore, we expect that the trajectories of any of these vehicles downstream of the intersection will be nearly independent of a), b), or c) at least until vehicles start to pass each other or one platoon catches up with another.

One might argue that a driver who has waited a long time may be less (or more) patient than one who has waited a short time but we do not expect that this is a serious issue. One might also need to make some qualifications for right-turning vehicles which can "turn-on-red after stop" or left-turning ve-
this is a serious issue. One might also need to make some qualifications for right-turning vehicles which can "turn-on-red after stop" or left-turning ve-hicles which may block the intersection if there is no left turn bay.

These assumptions will be very important because we will be mostly concerned here with the consequences of changing the signal timing. If, for example, we should simply delay the start of a green interval, the effect of this on the vehicles which are stopped is that their subsequent trajectories are displaced to a later time, regardless of what the actual shape these trajectories may be, i.e., independent of any theory of vehicle dynamics.

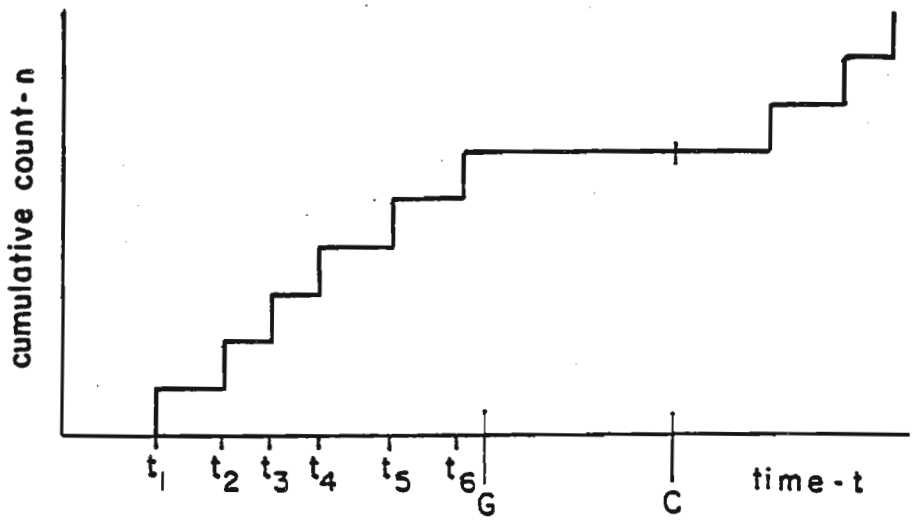


Fig. 1.5 - Typical cumulative counts at a traffic signal.

Suppose now that one were to draw a graph of the cumulative number of vehicles to pass an intersection, $n(t, x)$, with x denoting some reference point in or near a signalized intersection and t measuring the time from the start of the green as in figure 1.5. Suppose, also, that there is no turning traffic and we are observing a single lane. The assumption is that the times $t_1, t_2, \ldots$ of the steps in $n(t, x)$ for vehicles which are stopped may depend on the type of vehicles and drivers, the geometry of intersection, and $t_1, t_2, \ldots$ of the steps in $n(t, x)$ for vehicles which are stopped may depend on the type of vehicles and drivers, the geometry of intersection, and x but not on the past history of the vehicle trajectories.

We could continue the curve $n(t, x)$ over many cycles. If the signal has a constant cycle time C and green interval G, $n(t, x)$ will have a flat portion between times G and C and then some more steps during each subsequent green intervals. Alternatively, we could cut out only that portion of the curve for $0 \leq t \leq C$ and then imagine that we started a new experiment by resetting the clock to 0 at time C and the counter also to 0. Thus, we can generate repeated observations of $n(t, x)$ for $0 < t < C$.

Different signal cycles will contain different vehicles but we might imagine that the vehicles behave as if the vehicles in each cycle were selected at random from some population of possible vehicles. If each cycle contains several (say, at least four or five) stopped vehicles and we evaluate the arithmetic average of the $n(t, x)$ over N cycles, i.e.,

$$\bar{n}(t, x) = \frac{1}{N} \sum_{k=1}^{N} n_k(t, x) \ , \tag{1.4.1}$$

with $n_k(t, x)$ the curve for the kth cycle, then we expect that for sufficiently large N, $\bar{n}(t, x)$ will be a smooth (noninteger) function of t of the form shown in figure 1.6. This curve should be nearly reproducible; if we came back to the same location another day and evaluated $\bar{n}(x, t)$ again, we should obtain nearly the same curve.

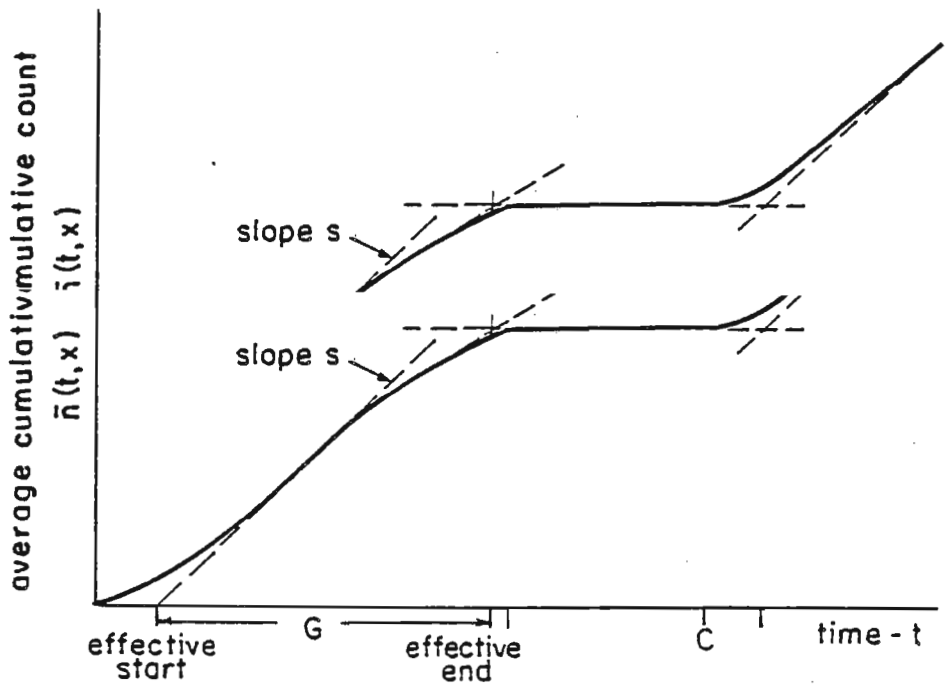


Fig. 1.6 - Average counts at a traffic signal, interpretation of effective green time.

We expect (and should verify) that the slope of the curve rises gradually after the start of the green interval due to differences in the times t_1, t_2 in successive cycles, but after the first few cars have passed there is a nearly linear portion of the curve with slope s while the queue continues to discharge. After the queue vanishes (which may not be at the same time in each cycle), the slope of $\bar{n}(t, x)$ should decrease possibly to another nearly linear segment, but then drop to zero rather sharply at the end of the green time.

The term "flow" will be used here to denote "the number of vehicles passing some point per unit time." If the "flow" is varying with time either due to traffic signals or, on a coarser time scale, to variations in demand, the interpretation of this term is rather tricky. Since counts are integer-valued and also fluctuate due to "random" effects, it is necessary to count a rather large number of vehicles in order to obtain an observation which is reproducible. If, however, one counts vehicles over a time interval long enough to contain many vehicles, the interval may be so long as to average out the time-dependence one is trying to observe.

One can "smooth out" fluctuations or, in effect, increase the size of count by repeating observations under "identical conditions." Conceptually, this is straightforward and this is what is implied in (1.4.1). If one chooses N large enough, the curve $\bar{n}(t, x)$ should be sufficiently reproducible that one can take a subinterval of the cycle time and define a slope of the curve at large enough, the curve $\bar{n}(t, x)$ should be sufficiently reproducible that one can take a subinterval of the cycle time and define a slope of the curve at (or near) some time t. We will interpret an "instantaneous flow" $q(t, x)$ as the slope of $\bar{n}(t, x)$ at time t, i.e.,

$$q(t, x) = \frac{\partial}{\partial t} \bar{n}(t, x) \ . \qquad (1.4.2)$$

The slope s in figure 1.6 will be called the "saturation flow."

If a queue is sufficiently long that all vehicles passing an intersection have been stopped, the linear portion of the curve $\bar{n}(t, x)$ should extend nearly

to the end of the green interval, and the number of vehicles $\bar{n}(G, x)$ to pass during the green interval should be a linearly increasing function of G, with slope s.

The shape of the curve $\bar{n}(t, x)$ near the start or end of the green interval will depend on where one is observing the counts within the intersection, the location of the stop line, geometry of the intersection, etc., but not on signal timing. There may be some important issues relating to where one should locate the stop line, removal of sight restrictions, etc., which could influence "start up" times, but we will not be concerned with such issues. We will be concerned mostly with how many vehicles pass the intersection and the approximate times when they pass, and for this purpose it usually suffices to approximate $\bar{n}(x, t)$ by a piecewise linear curve. If there is a linear portion with slope s, we will extrapolate this backwards in time until it intersects the line $n = 0$ and we will interpret the "effective" start of green to be this time of intersection as shown in figure 1.6. Similarly, at the end of the green, we can make a linear extrapolation of any linear part of $\bar{n}(t, x)$ until it intersects a horizontal line at height $\bar{n}(G, t)$. We will then define the "effective" green interval as the time interval between these intersection points. Unless otherwise specified, we will hereafter use the terms, "start of green," "end of green," or "green interval" to mean these effective times; in particular, the symbol G will here-
we will hereafter use the terms, "start of green," "end of green," or "green interval" to mean these effective times; in particular, the symbol G will here-after refer to the effective green interval. Essentially, by definition now, the average number of vehicles which can pass the intersection during the effective green interval G, if there is a long queue, is sG.

Although the key postulate here, that the curve $\bar{n}(t, x)$ has a linear portion of slope s, is fairly standard, it may be subject to question. Obviously there will be problems if left turning vehicles block the intersection. Also, if there are slow moving vehicles (particularly trucks) one may obtain some gaps in the traffic and the gaps are likely to be longer toward the end of the cycle

than at the beginning, i.e., the (average) flow $q(t, x)$ may decrease with t toward the end of the green even though all vehicles had been stopped.

The Highway Capacity Manual or various handbooks contain extensive empirical data relating the saturation flow (or the flow per hour of green) to lane width, percent of trucks, etc. In fact, these depend also on local habits of driving. They certainly vary from one country to another, but they may even vary within the same city or with the time of day for reasons other than those classified in the Highway Capacity Manual. If one really wants to know the characteristics of some particular intersection, one should measure the relevant properties directly.

Another important property of the $n_k(t, x)$ is the variation of this count from one cycle to the next, particularly the variation in the final count $n_k(G, x)$ when there is a queue during the whole cycle. Typically, one would measure this by the (sample) variance

$$\left(\frac{1}{N-1}\right) \sum_{k=1}^{N} \left[n_k(G, x) - \bar{n}(G, x)\right]^2$$

but actually it is more convenient to measure the ratio of the variance to the mean

$$\text{mean} = \frac{\left(\frac{1}{N-1}\right) \sum_{k=1}^{N} \left[n_k(G, x) - \bar{n}(G, x)\right]^2}{\bar{n}(G, x)} \qquad (1.4.2)$$

$$I_D = \frac{\left(\frac{1}{N-1}\right) \sum_{k=1}^{N} \left[n_k(G, x) - \bar{n}(G, x)\right]^2}{\bar{n}(G, x)} . \qquad (1.4.3)$$

If each driver should choose a headway $t_j(x) - t_{j-1}(x)$ independent of other drivers, the variance in the time for n vehicles to pass the intersection should be nearly proportional to n and the mean time for n vehicles to pass should also be nearly proportional to n. Also, the variance of the count in a given interval G should be nearly proportional to G as is the mean count. If both the numerator and denominator of (1.4.3) are nearly proportional to G,

the ratio should be nearly independent of G. If the probability distribution for the count $n_k(G, x)$ had a Poisson distribution, I_D would be exactly 1, independent of the counting interval.

Of course, vehicles leaving a traffic signal do not have a Poisson distribution. Typically the value of I_D is appreciably less than 1, perhaps comparable with 1/4 to 1/2. There is no theory which would predict a value of I_D but it can be measured. In principle, it should be straightforward to make numerous measurements of the counts during cycles when all vehicles are stopped or counts in fractions of cycle times if only some vehicles are stopped. The value of the I_D, however, is quite sensitive to "outliers": turning cars, trucks, etc. In practice, it typically requires a rather large number N of observations to obtain a reproducible observation of I_D. One does not usually need to know this very accurately, but it is useful to have some estimate. The main conjecture here is that the value of I_D is (nearly) independent of the period of observation (G, for example).

1.5. Objective Functions

Although certain aspects of traffic signal theory depend on how people drive, we saw in the last section that some obviously important aspects of the theory are insensitive to the detailed dynamics of the vehicles. Our goal is to obtain some rational basis for selecting traffic signal settings, but the theory are insensitive to the detailed dynamics of the vehicles. Our goal is to obtain some rational basis for selecting traffic signal settings, but the complexity of any theory and, in particular, its sensitivity to the dynamic behavior of vehicles depends very much on what we choose as an "objective function."

Any individual would like to travel from his origin to his destination with no interference from traffic signals or other vehicles, but he would not like to pay what it would cost to provide this service. In effect, he agrees (through taxes, voting, or whatever) to share the cost and use of some facilities

in which he must take his turn using them. Not all drivers pay the same taxes and some would be willing to pay for special service, but there is no mechanism whereby drivers can be charged for each service they receive. Even if there were, it probably would not be socially acceptable. In the final analysis, one has a reasonable objective function if the use of this objective function leads to strategies which a majority of voters is willing to buy.

One can attach prices to certain things; travel time, stops, fuel consumption, pollution, etc. It is usually assumed that these costs are additive, i.e., the total system cost can be expressed as the sum of costs to individual travelers, or in the case of pollution, or other environmental consequences, as the sum of effects caused by individual travelers. No matter what simple mathematical structure one proposes for a "cost", however one can find some contradiction; something people do for good reason which is not consistent with the proposed minimum cost. The additive cost structure, for example, immediately implies that one unit of cost (such as delay) to each of ten travelers is equivalent to 10 units of cost to one traveler. Society, however, typically favors sharing costs among many people rather than forcing the costs on a few. One might propose some more general cost structure to reflect this preference, but then one must ask how many units of cost to one person is equivalent to one unit of cost to each of 10 people (and all other combinations)? Obviously no one knows how to quantify all these things.

cost to each of 10 people (and all other combinations)? Obviously no one knows how to quantify all these things.

Another very unpleasant aspect of any "optimization" scheme is that any improvement in travel along some route will generally cause an increase in demand for that route. Some of this increase may be due to diversion of traffic from other routes, but some may be newly generated traffic. One might be able to attach some numerical value to the benefits for those travelers who switched routes, but whether or not one wishes to generate new trips or longer trips is highly controversial or, in any case, its value is difficult to quantify.

We will, for the most part, try to avoid any issues related to "elasticity of demand." By implication, at least, we will assume that any strategy which reduces the cost of travel for a fixed demand will generally lead to a net benefit to society even if the demand does change. This obviously is not always true; but if people shift routes or make new trips, they do so because they receive some benefit (possibly at someone else's expense). We are more concerned with developing strategies which are likely to be "beneficial" than with trying to attach some numerical value to the benefit.

We will not explicitly consider fuel consumption, pollution, or other environmental effects because to include these in an objective function without considering elasticities of demand would imply that a reduction of these effects for a fixed demand would also yield a reduction with an elastic demand. Actually the final result is likely to be the opposite. If people are willing to spend a certain fraction of their time traveling, then the faster they can travel, the farther they will travel and the more fuel they will consume. In effect, we will assume that people would like to travel with fewer delays and fewer stops despite possible adverse consequences.

We will, for the most part, be concerned here with strategies which minimize total travel time (delay) and/or the number of stops for a fixed demand, but, at the same time, being conscious of possible conflicts with other objectives. "Delay" is, by definition, any excess travel time relative to some idealized (zero delay) travel time, so travel time and delay are, in effect, equivalent. A "stop" is not as well defined since drivers who will be "delayed" due to a signal or other vehicles may have an option of just traveling slower or coming to a complete stop. We will typically just count the sum of the number of vehicles which are delayed by all signals, whether the vehicles actually stop or not. This is somewhat arbitrary, but we do not necessarily attach any economic value to this. We are merely identifying some possible

measure of performance.

Total travel time is a particularly convenient quantity because, on the one hand, it is a meaningful measure of performance while, on the other hand, it is highly insensitive to the details of the vehicle dynamics, so much so, in fact, that the signal settings which minimize total travel time are often not unique. Thus, among strategies which minimize total travel time, it is often possible to specify also a secondary objective.

Whether or not a vehicle is stopped (or delayed) by a signal is also insensitive to details of the vehicle trajectories. The total number of stops is a meaningful performance measure in that it is highly correlated with fuel consumption, accidents, and driver irritation. One should not, however, use the number of stops as a primary objective, because minimization of the number of stops usually leads to enormous delays to a few drivers so that other drivers have no delay.

We could consider some appropriate linear combination of travel time and stops as an objective function. Indeed the total "cost" of travel including time, fuel, inconvenience, etc., could typically be approximated by such a linear combination. If, however, one is concerned with how the "optimal solution" will vary with the price of fuel, for example, one must evaluate both the travel time and stops separately as a function of the strategy (which is what we propose to do).

travel time and stops separately as a function of the strategy (which is what we propose to do).

To see why the total travel time of all drivers is quite insensitive to driver behavior or to certain strategies of control, suppose we are concerned with the travel time of vehicles along a single highway with no turning traffic. Vehicles enter the highway by passing one traffic signal and they exit at another traffic signal, but there might be any number of traffic signals in between. Suppose we station observers at the two ends of the highway, at locations x and x', and they observe the times $t_j(x)$, $t_j(x')$ when each

vehicle passes as described in section 1.3. The total travel time for a specified number n of vehicles can be written as

$$\text{Total travel time} = \sum_{j=1}^{n} [t_j(x') - t_j(x)]$$

$$= \sum_{j=1}^{n} t_j(x') - \sum_{j=1}^{n} t_j(x) . \qquad (1.5.1)$$

If the first observer numbers the vehicles in order as they pass him, he can draw a graph of $n(t, x)$ as in figure 1.4. The quantity

$$\sum_{j=1}^{n} t_j(x)$$

can be identified with the area on this graph enclosed by the curve $n(t, x)$, the vertical line $t = 0$ and two horizontal lines at height 0 and n. The curve $n(t, x)$, however, will have a shape as in figure 1.5. If we were to repeat the observations over many days and take the average, $\bar{n}(t, x)$, it would appear as in figure 1.6 for each signal cycle.

The time $t_j(x')$ is the time when the vehicle labeled as the jth vehicle at x passes location x'. If vehicles can pass each other enroute from x to x' the $t_j(x')$ will not necessarily be in order. The quantity

$$\sum_{j=1}^{n} t_j(x') , \qquad (1.5.2)$$

$$\sum_{j=1}^{n} t_j(x') , \qquad (1.5.2)$$

however, does not depend on the order in which the $t_j(x')$ are numbered. This would have the same value if the observer at x' renumbered the $t_j(x')$ in order of increasing time. If he constructed a graph of $n(t, x')$, (1.5.2) would again have the interpretation as an area to the left of the curve $n(t, x')$, provided, of course, that the second observer saw the same n vehicles. The total travel time (1.5.1) would then be the area between $n(t, x)$ and $n(t, x')$ and two horizontal lines at height 0 and n.

There could be traffic signals between x and x' and we could station other observers at intermediate locations, and draw curves for $n(t, y)$ at any locations, $x < y < x'$. The trip time from x to x' could be written as the trip time from x to y plus that from y to x' and the total trip time of all vehicles from x to x' could be written on the sum of that from x to y and from y to x', i.e., the area between $n(t, x)$ and $n(t, x')$ is the sum of the areas between $n(t, x)$ and $n(t, y)$, and between $n(t, y)$ and $n(t, x')$ (all between heights 0 and n).

Obviously the trip time of vehicles and consequently (1.5.1) will depend upon any signals that may exist between x and x' but the expression (1.5.1) does not explicitly contain the observations at intermediate locations. Furthermore the shape of the curve $\bar{n}(t, x')$ relative to the start of some green time at x' should be similar to that of figure 1.6 and insensitive to the details of how vehicles arrive at this intersection. This insensitivity of $\bar{n}(t, x')$ to arrival times at x' also implies an insensitivity of (1.5.1) to certain types of strategies at intermediate points. If a driver has an "appointment" to leave x' at a certain time which requires that he must wait somewhere, it makes no difference where he waits as long as he arrives at x' in time for his appointment.

The use of the total travel time as an "objective function" has some obvious advantages in terms of mathematical simplicity, but its insensitivity to

The use of the total travel time as an "objective function" has some obvious advantages in terms of mathematical simplicity, but its insensitivity to certain strategies sometimes implies some equivalences which are not consistent with accepted preferences. For example, (1.5.1) is insensitive to interchange of the ordering of vehicles. If two vehicles which suffer one unit of delay each were to swap positions, it might result in having one vehicle suffer no delay but the other vehicle two units of delay with no change in (1.5.1). Society, however, obviously prefers the former not only because people prefer that everyone be treated equally but also because they do not believe that two

units of delay is worth twice as much as one unit of delay. One might also question whether a person who will be delayed 10 minutes is equally willing to wait ten minutes at the same location or one minute at each of ten locations.

2. ISOLATED INTERSECTIONS (UNIFORM ARRIVALS)

2.1. Introduction

ns.

A platoon of vehicles leaving a signalized intersection will gradually spread as vehicles pass each other or adjust to more cautious headways. Eventually the platoons from successive signal cycles will spread enough as to overlap and form a steady traffic stream. It is likely to require many miles of travel for this to occur, but, if this traffic stream should now approach a second traffic signal, the behavior of the traffic at the second signal should be independent of the timing of the first traffic signal.

Despite the fact that this is not the situation at most traffic signals, we will begin our more detailed analysis of traffic intersections by considering the behavior of traffic at "isolated intersections," which, by definition, means intersections at which the approaching traffic flow is (nearly) uniform over a cycle time. This will allow us to introduce a number of issues, which may also apply to networks, but which involve relatively few parameters.

When vehicles approach an intersection, a queue will form during a red interval and this queue will propagate upstream. The assumption that the approaching traffic stream is time-independent means that an observer located sufficiently far upstream that the back of the queue does not reach him will observe a steady flow q, i.e., a curve $\bar{n}(t, x)$, appropriately smoothed or averaged over many repetitions. will have a constant slope [illegible] a steady flow q, i.e., a curve $\bar{n}(t, x)$, appropriately smoothed or averaged over many repetitions, will have a constant slope q.

An idealized trajectory plot might appear as in figure 2.1a. We have drawn this as if vehicles passed some point, x, at time intervals of (exactly) $1/q$ and all had the same approach velocity v but actually it suffices if only the <u>average</u> time interval between vehicles is $1/q$. The velocities need not be exactly equal, but we do not expect there to be much passing near an intersection. The corresponding averaged curve $\bar{n}(t, x)$ is as shown in figure 2.1b. If $x = 0$ denotes the location of the intersection, the curve $\bar{n}(t,0)$

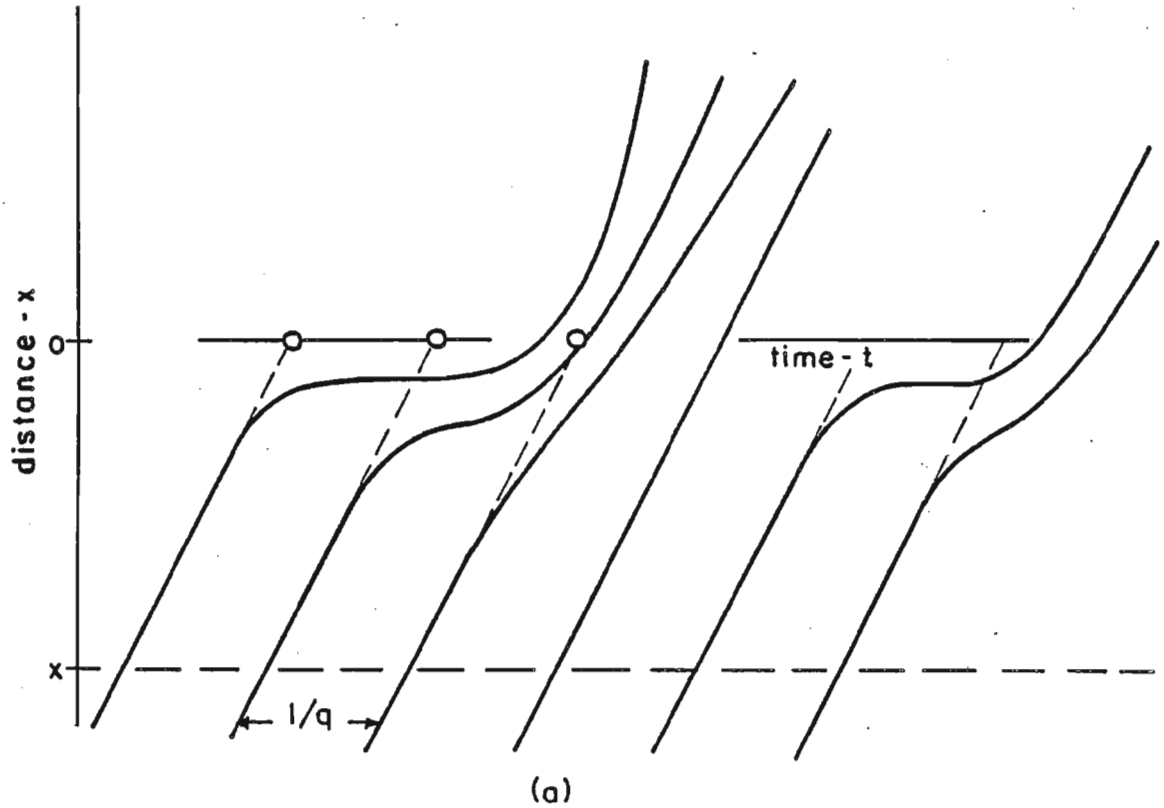


Fig. 2.1a - Interpretation of expected time of arrival.

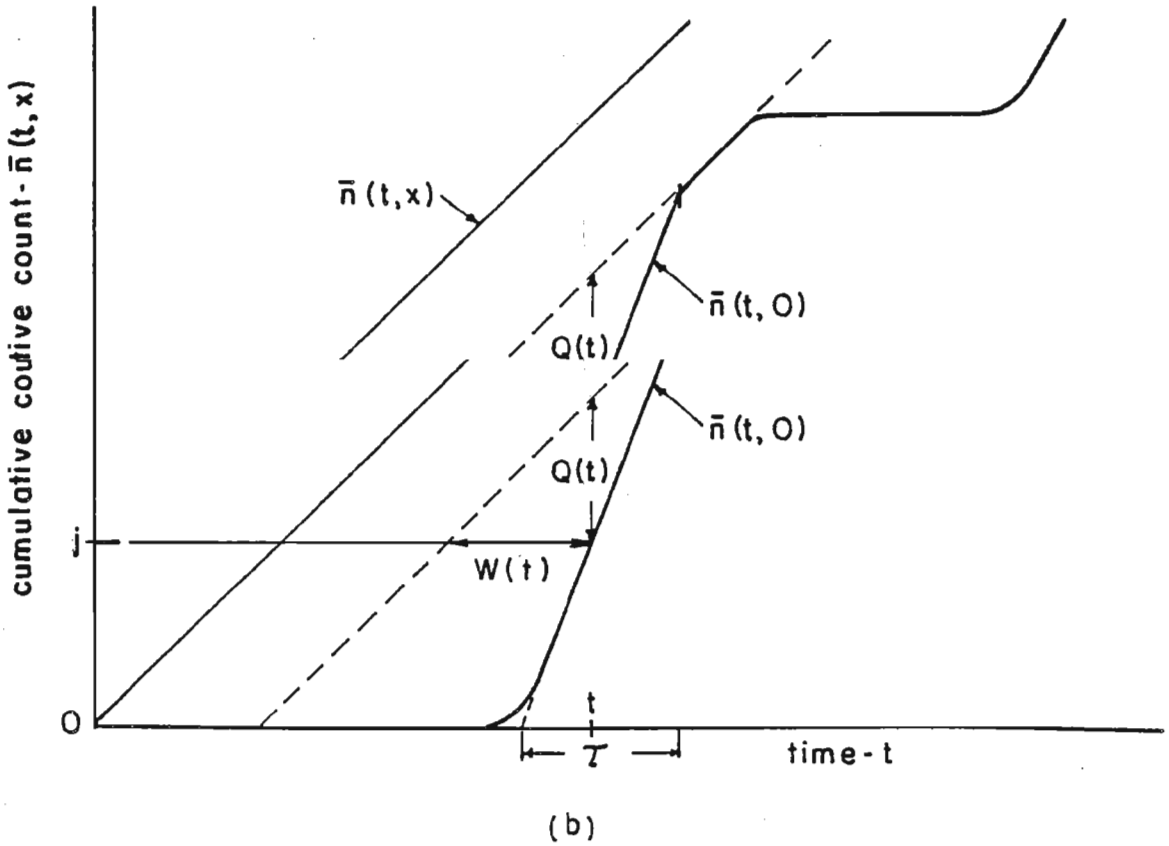


Fig. 2.1b - Graphical interpretation of queue length and delay.

is drawn as in figure 1.6. As described in section 1.2, the horizontal distance between $\bar{n}(t, x)$ and $\bar{n}(t, 0)$ represents the average transit time from x to (past) zero for some jth vehicle and the vertical distance between the curves at time t represents the number of vehicles between locations x and 0 at time t.

The trip time from location x to 0 for any jth vehicle and the number of vehicles in this section of road at time t depend on the locaton x. If the distance $|x|$ is larger than the space occupied by the queue and the approaching traffic is stationary, both the trip time and the number of vehicles in the road section should increase linearly with the distance $|x|$. Specifically, the former will increase proportional to $v_j|x|$ if v_j is the (time-average) speed of the jth vehicle, and the latter will increase proportional to $k|x|$, if k is the spatial density of vehicles (the number of vehicles per unit length of road). If different vehicles have different velocities, we can define an average velocity $\bar{v}$ as the length of a road section divided by the average trip time of all vehicles traversing that road section. The q, k, and $\bar{v}$ will then by related by

$$q = k\bar{v} . \tag{2.1.1}$$

By assumption, q, k, and $\bar{v}$ are all independent of time, even though the trip time from x to 0 and the number of vehicles between x and 0 are By assumption, q, k, and v are all independent of time, even though the trip time from x to 0 and the number of vehicles between x and 0 are both time-dependent.

It is customary to subtract a term $\bar{v}|x|$ from the trip time from x to 0 and a term $k|x|$ from the number of vehicles in the road section. We define

$$\begin{aligned} W(t) &= \text{"waiting time" or "delay" of a vehicle which leaves } x = 0 \text{ at time } t \\ &= (\text{trip time of this vehicle from } x \text{ to } 0) - \bar{v}|x| , \end{aligned} \tag{2.1.2}$$

and

$$Q(t) = \text{"queue length" at time } t$$
$$= (\text{number of vehicles between } x \text{ and } 0 \text{ at time } t) - k|x| \ . \qquad (2.1.3)$$

The $W(t)$ and $Q(t)$ should be independent of the shape of the vehicle trajectories near the intersection provided $|x|$ is sufficiently large that the signal does not disturb the motion of the vehicles at x.

Equivalently, we could imagine some hypothetical vehicle trajectories for which vehicles continue their steady motion until they reach $x = 0$. They then stop instantaneously, but leave the intersection at the times described by the actual trajectories, i.e., we replace the trajectories for $x < 0$ by the broken line extrapolation of the constant speed trajectories shown in figure 2.1a. We refer to the times at which these broken line trajectories reach $x = 0$ as the "arrival times" or "expected arrival times." We can also draw in figure 2.1b a curve (broken line) corresponding to the (average) cumulative number of "arrivals" at $x = 0$. The waiting time of a vehicle which passes the intersection at time t is now the horizontal distance between this cumulative arrival curve and the curve $\bar{n}(t, 0)$ for the departures as shown in figure 2.1b.

The "queue length" at time t is the vertical distance between the arrival and departure curves of figure 2.1b. It is actually the number of vehicles which are expected to arrive by time t less those which have actually left. The latter is directly observable but the former must be inferred from observations at locations $x < 0$ upstream of the queue, as described above.

One should not confuse the "queue length" as defined here with what one might observe as a "physical queue," the number of vehicles (upstream of $x = 0$) which appear to be slowed by the signal or are actually stopped. The latter is always larger because it includes those vehicles which have already reached the end of the physical queue but would not yet have "arrived" at $x = 0$ if they could have continued moving. The value of the "physical queue"

can be easily evaluated during the red interval of the signal when (nearly) all vehicles are either moving at their approach speed or are stopped. If k_j (jam density) is the density of stopped vehicles, then

$$\text{physical queue at time } t = Q(t)/(1 - k/k_j), \qquad (2.1.4)$$

which could be larger than $Q(t)$ by as much as a factor of 2.

After the signal turns green, the number of vehicles in the physical queue should decrease, but the space occupied by the queue may continue to increase for a while (until a starting wave can reach the end of the physical queue). We may be concerned with the space occupied by a physical queue when we consider the possibility of "blocking" in synchronized signal systems, but otherwise this is irrelevant for the purpose of evaluating delays.

It is common practice among traffic engineers in making "delay studies" at an intersection to collect much more information than is necessary. Some people follow individual vehicles by means of moving pictures, or compare license plate numbers of vehicles as they pass two observation points (at x and 0). Others may record the number of vehicles actually stopped near the intersection or the total time spent by vehicles in some road section during some time period.

It is clear from figure 1.4 or 2.1 that the time spent by vehicles in the road section from x to 0 during the time interval t' to $t' + dt'$ is equal to dt' times the number of vehicles in this section, $n(t', x) - n(t', 0)$, road section from x to 0 during the time interval t' to $t' + dt'$ is equal to dt' times the number of vehicles in this section, $n(t', x) - n(t', 0)$, at time t'. Over some time period, say from time 0 to time t, the total time spent by vehicles in this section is

$$\int_0^t [n(t', x) - n(t', 0)]dt'$$

i.e., the area between the curves $n(t, x)$ and $n(t, 0)$ within the time 0 to t. This is valid independent of how vehicles move from x to 0 or even if vehicles may pass each other in the section.

In figure 2.1 we have separated this area into two parts, the area between

$n(t, x)$ and the broken line, and the area between the broken line and $n(t, 0)$. The former area is interpreted as the time that would be spent by vehicles in this section if there were no signal (which depends on the location x), and the latter

$$\int_0^t Q(t')dt'$$

is interpreted as the total "delay" to all vehicles during the time 0 to t (which is, supposedly, independent of x).

The total time spent by vehicles in x to 0 or the total delay during time 0 to t can also be represented as the sum over all vehicles of the time spent in this section during the time period by each vehicle. Thus, in figure 1.4 it would be represented as the area of all horizontal strips such as the shaded strip in figure 1.4 but between two vertical lines at times 0 and t. For given times at which vehicles pass the intersection, i.e., for a given curve $n(t, 0)$, this total time spent by vehicles in the section depends only on the $n(t, x)$ and $n(t, 0)$, as evaluated by the first method above. Thus, this total time, if evaluated from a figure analogous to figure 1.4, must be independent of whether or not vehicles pass each other or not (in figure 1.4 it would be independent of whether the times $t_2(x')$ and $t_1(x')$ were the times at which vehicles 2 and 1 respectively passed, or the times at which vehicles 1 and 2 passed).

To evaluate the total time spent in the section or the total delay or the average delay per vehicle, it suffices to construct the curves $n(t, x)$ and $n(t, 0)$ (provided that there are no entrances or exits between x and 0 so that any vehicle which passes x must also pass 0). It is not necessary to observe license numbers or follow vehicles on a film so as to determine whether or not vehicles pass each other (unless one is also interested in the distribution of the delays over vehicles). One does not need to know precisely the trip time of each vehicle in order to evaluate the average.

For the purpose of estimating the approximate delay, one does not need to know precisely the times at which each vehicle passes x or 0. We expect that vehicles that pass x or 0 during some five or 10 second interval will be more or less evenly distributed over the time interval. If one observes the count $n(t, x)$ only at certain discrete times (every five or 10 seconds), one could interpolate or draw a smooth curve through the observed points. Besides, one is only interested in certain average properties, averaged over many cycles of the signal or over many vehicles. One is not interested in recording properties of the system which are not reproducible.

If, however, one knows or observes that the arrival rate is independent of time, it suffices to measure just the q, plus perhaps the magnitude of certain stochastic deviations of the actual curve $n(t, x)$ from a straight line of slope q. Similarly for the curve $n(t, 0)$, it suffices to observe the saturation flow s and the effective start of green. One need not observe the actual departure times of each vehicle.

If we disregard for now the possibility that the queue of vehicles may clear the intersection during some green intervals but not in others (due to stochastic effects), we can approximate the actual arrival and departure curves by their averages. We assume that any vehicle which arrives after the queue vanishes will not be delayed.

If the queue vanishes during a time when the curve $\bar{n}(t, 0)$ has slope s, will not be delayed.

If the queue vanishes during a time when the curve $\bar{n}(t, 0)$ has slope s, it will vanish at a time τ after the start of the effective green such that the cumulative departures after the start of green $s\tau$ is equal to the cumulative arrivals since the start of the preceding red time $(C - G + \tau)q$. Thus

$$\tau = q(C - G)/(s - q) .$$

The fraction of vehicles which are delayed is $(C - G + \tau)/C$ or

$$\text{fraction of vehicles delayed} = \frac{(1 - G/C)}{(1 - q/s)} . \qquad (2.1.5)$$

A necessary condition for this to be valid is that it be less than 1, i.e.,

$$q/s < G/C \ . \qquad (2.1.6)$$

If several vehicles contribute to the average delay, it will not make much difference if we approximate the departure curve by a straight line even for the first one or two vehicles. The delay to succeeding vehicles will decrease at a constant rate from a maximum of (approximately) $C - G$ to zero with an average value of $(C - G)/2$ for each vehicle which is delayed. The average delay for all vehicles is this multiplied by the fraction delayed.

$$\text{average delay per vehicle} = \overline{W} = \frac{1}{2}\left(1 - \frac{G}{C}\right)^2 \frac{C}{(1 - q/s)} \ ; \qquad (2.1.7)$$

provided that (2.1.6) is true.

Formulas similar to (2.1.5) and (2.1.7) have been in existence almost as long as traffic signals have existed. They appear even in works of McClintock[1], Matson[2], and others as early as 1925, but in some of the earlier works people worried about the wrong things. They tried to relate the delays to the details of accelerations of vehicles during the start-up time, for example. The use here of an "effective green interval" eliminates much of this complexity at the expense of only a small error. This error is of little consequence since delay is only a crude measure of performance anyway.

There are two more serious deficiencies of these formulas. One is that it neglects stochastic effects, which we will discuss in more detail later. The

There are two more serious deficiencies of these formulas. One is that it neglects stochastic effects, which we will discuss in more detail later. The

other is that it describes the delay only upstream of the signal but disregards any effects downstream of the intersection.

There is no satisfactory way of estimating delays to vehicles downstream from an isolated signal. If a jth vehicle approaches a traffic signal with a nearly constant velocity v_j and eventually returns to the same velocity far downstream of the intersection, one could, in principle, extrapolate the (nearly) linear trajectory of the vehicle downstream of the intersection back

to the intersection in a manner analogous to what we did in figure 2.1a upstream of the intersection. We could thus define an effective departure time of this vehicle from the intersection and interpret the total delay as the difference between the arrival and departure times. This is what one would like to do since the total delay should certainly be interpreted as the time displacement of the final trajectory with velocity v_j from the extrapolation of the initial trajectory (presumably the hypothetical trajectory that the vehicle would follow if there were no signal).

If the flow q were sufficiently small that each signal cycle would stop at most one or two vehicles ($qC < 1$), one could observe the actual trajectories of a sample of vehicles and measure directly the time lost by a vehicle during acceleration downstream of the intersection until it reaches a steady velocity v_j (provided the velocity did actually return to its initial value upstream of the intersection). For any vehicle which comes to a complete stop, the delay downstream of the intersection should, presumably, be independent of how long the vehicle was stopped. The total delay of all vehicles downstream of the intersection should, therefore, be nearly proportional to the number of vehicles stopped or the fraction (2.1.5) of vehicles stopped. In some economic cost objective function the contribution of this delay would have a similar dependence on the signal parameters as fuel consumption.

objective function the contribution of this delay would have a similar dependence on the signal parameters as fuel consumption.

If a signal stops two or three vehicles per cycle, the second or third vehicle may need to pass the first vehicle in order to achieve its final velocity. The existence of the signal does not in itself increase the number of passing maneuvers. If the second vehicle wishes to pass the first vehicle, it would have done so someplace even if the signal were not there, but if passing is prohibited in the vicinity of the intersection, a faster vehicle will have to follow a slower car for a longer time before it can pass. The evaluation of such delays is rather tedious. One might expect, however, that they would

be relatively small for light traffic.

There are some "car-following" and "fluid theories"[3, 4] of vehicle dynamics designed to describe vehicle motion when traffic is so dense that there is negligible passing (or lane-changing). These theories are of questionable accuracy even under these conditions. One might use them to describe the motion of vehicles near the intersection, but we are concerned with the total delay until some final equilibrium. The ranges of steady flow q which can pass an intersection (satisfy (2.1.6)) are typically well below the capacity of the highway itself and, in the final equilibrium downstream of the intersection, would be considered as "moderately light."

Actually, calculations have been made of the total delay[5] due to a traffic signal based on the fluid theory of Lighthill and Whitham. This theory would, in effect, imply that vehicles do not pass each other. The final equilibrium would obtain only after platoons had spread so much as to have joined those from adjacent signal cycles and all irregularities of the flow generated by the signal had been smoothed out (after a distance which would certainly be many miles). If vehicles approach the signal at a constant flow q, eventually again achieve a constant flow q at $x \to \infty$, and cannot pass, then all vehicles will experience the same total delay. The final trajectories are simply a uniform translation from the initial trajectories. The value of this displacement was found to be the same total delay. The final trajectories are simply a uniform translation from the initial trajectories. The value of this displacement was found to be $C - G$ (the effective red interval), independent of q or s (provided that (2.1.6) is valid)!

If G were equal to $C/2$, this fluid theory would predict a total delay per vehicle of $C/2$ but the same theory would also be compatible with (2.1.7) and predict that the delay upstream of the intersection would vary between $C/8$ and $C/4$ as q/s increases from 0 to $1/2$. We are thus led to the strange conclusion, from this theory, that the delay downstream of the intersection is

a decreasing function of q/s, varying between 3/4 and 1/2 of the total delay.

For small values of q/s this conclusion is unacceptable. It is a strange consequence of a hypothesis that vehicles cannot pass each other. If one vehicle is delayed, so is every other vehicle delayed (by the same amount). But in a more realistic theory one would conclude that, for sufficiently small q/s, the vehicles do not interact with each other and there would be no further delay downstream of the intersection for vehicles which experience no delay upstream.

For values of q/s closer to the saturation level, the fluid theory might be more reasonable. It does describe a potentially important fact that vehicles which manage to clear the intersection after the queue has vanished will subsequently overtake the platoon of vehicles which had been stopped. These vehicles, which suffer no delay upstream of the intersection will suffer delays downstream of the intersection when they are slowed to the speed of the platoon. It is also true that the lead car of a platoon moves into a (temporarily) empty highway and may travel faster than it could while approaching the intersection at a flow q. This vehicle may, therefore, recover some of its loss (until it overtakes the platoon from the previous cycle).

Despite the above uncertainty about delays downstream of the signal, we will use the delay upstream of the signal or some linear combination of this and the number of stopped vehicles as an "objective function" for isolated signals although we do not claim that this is necessarily a true measure of total

and the number of stopped vehicles as an "objective function" for isolated signals, although we do not claim that this is necessarily a true measure of total
"cost."

2.2. Deterministic Approximations for a Four-Way Intersection

The reason for installing a traffic signal at an intersection is that one has a certain section of pavement which must be shared by two or more traffic streams which cannot use it simultaneously. There must be some strategy for sequencing users to avoid collision and/or excessive delays.

One might use a traffic signal primarily to allow pedestrians sufficient time to cross the road. If one uses it primarily to control vehicles, however, the reason for using a traffic signal rather than stop or yield signs is that one can typically achieve higher average flows from serving a sequence of vehicles traveling in the same direction than if one serves them in some alternating sequence. Thus, one uses a traffic signal primarily in situations where there is some potential congestion.

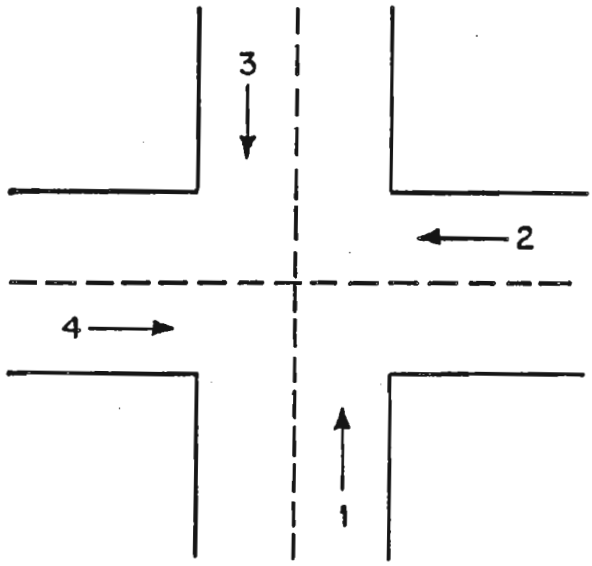


Fig. 2.2 - A four-way intersection.

Fig. 2.2 - A four-way intersection.

To illustrate some of the issues, we consider first the typical four-way intersection shown in figure 2.2 with no turning traffic. We number the lanes as shown with odd numbers in one direction, even numbers in the cross direction.

Each traffic stream reacts to the same traffic signal, but otherwise there is no intersection between the four traffic streams. If there are multiple lanes in any direction, but no turning traffic, we will assume that vehicles distribute themselves among the lanes in such a way that the queue in adjacent lanes vanish at (nearly) the same time. We will treat the multiple lanes as if there were a

single lane with a saturation flow s_i , $i = 1, 2, 3, 4$ equal to the combined flow of all lanes for that direction. This is an "isolated" signal, so we assume that the approach flows are time-independent in all directions. Let q_i , $i = 1, 2, 3, 4$ be the combined flow in all approach lanes for direction i.

For each direction i , we can define an (effective) green interval G_i as described in section 1.4, and evaluate the fraction of vehicles delayed and the average delay per vehicle in this direction by inserting a subscript i on the q, s, and G in equations (2.1.5) and (2.1.7). The cycle time C is, of course, the same for all directions; and if all legs are undersaturated, it is necessary that

$$q_i/s_i < G_i/C \quad \text{for} \quad i = 1, 2, 3, 4 . \tag{2.2.1}$$

Presumably traffic directions 1 and 3 (or 2 and 4) see the same actual green interval (for no left-turning traffic with "leading" or "lagging" turn phases). It is reasonable to assume that they also have (nearly) the same effective green intervals, i.e., $G_1 = G_3$ and $G_2 = G_4$. Since it is arbitrary which of the directions we label as 1 or 3, we will define direction 1 so that

$$q_3/s_3 < q_1/s_1 . \tag{2.2.2}$$

Similarly, we will define direction 2 so that

$$q_4/s_4 < q_2/s_2 . \tag{2.2.3}$$

Similarly, we will define ...

$$q_4/s_4 < q_2/s_2 . \tag{2.2.3}$$

Directions 1 and 2 are the more critical directions in the sense that if (2.2.1) is valid for $i = 1$ and 2 , it is also valid for $i = 3$ and 4 .

One can define both (effective) green intervals G_1 and G_2 from the cumulative departure curves for directions 1 and 2 respectively, as illustrated in figure 1.6. The sum of these, $G_1 + G_2$, will, of course, be less than C . We can, therefore, define an (effective) lost time per cycle as

$$L = C - G_1 - G_2 > 0 . \tag{2.2.4}$$

One could decompose the L into various components. It includes time when the signal may be yellow (for two signal changes) and some loss due to "start-up" times for both directions. The value of L depends on the geometry of the intersection, vehicle types in the traffic streams, etc., but it is not likely to change if one changes the (actual or effective) green intervals. The value of L is likely to be comparable with 10 seconds. To analyze the properties of any particular intersection, however, it would be desirable actually to construct some cumulative departure curves for directions 1 and 2 over several cycles, evaluate some (average) values of G_1, G_2, and calculate the L from (2.2.4); also evaluate the s_1 and s_2. One could then assume or verify experimentally that the value of L (also s_1 and s_2) stay approximately the same if one changes the cycle times, the (actual) green times, or the arrival rates q_1, q_2. The implication here is that this would be true.

Various formulas for delays, stops, etc., will depend on the value of L but they will not depend upon any decomposition of L into parts identified with the switch of the signal from direction 1 to 2 and from 2 back to 1. Any change in the partition of L between the two signal switches will, in effect, merely change the time origin for directions 1 and 3 relative to 2 and 4 but will not change the relative times between events in the ith traffic direction.

If, for given values of the q_i, s_i and L, one has the option of changing will not change the relative times between events in the ith traffic G...

If, for given values of the q_i, s_i and L, one has the option of changing $G_1 = G_3$ and $G_2 = G_4$ (and the values of s_i, L are independent of G_1, G_2), the admissible ranges of G_1 and G_2 which satify (2.2.4) and (2.2.1) are defined by the inequalities.

$$\begin{aligned} L &< (s_1/q_1 - 1)G_1 - G_2 \\ L &< (s_2/q_2 - 1)G_2 - G_1 \,. \end{aligned} \tag{2.2.5}$$

In a space (G_1, G_2), the region defined by (2.2.5) is bounded by two straight lines as illustrated in figure 2.3 by the shaded region. A necessary

condition for there to be any admissible values of G_1 and G_2 is that the slope $(s_1/q_1 - 1)$ is larger than the slope $(s_2/q_2 - 1)^{-1}$, i.e.,

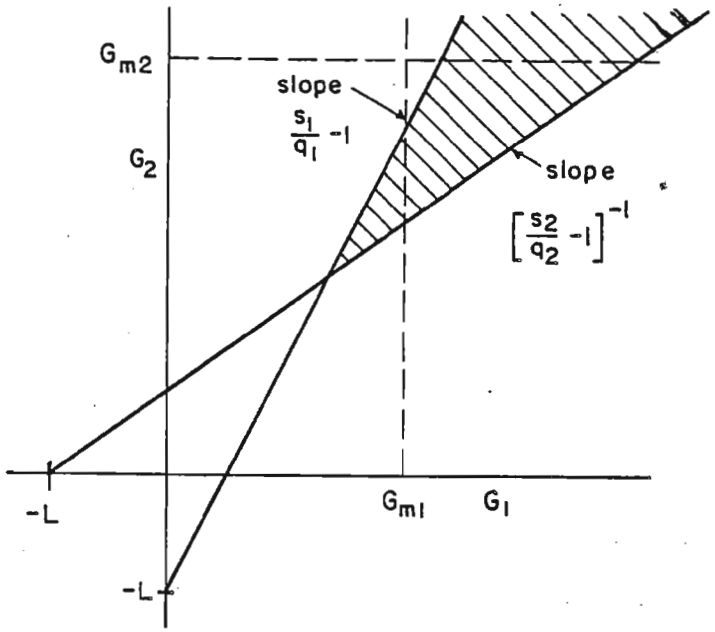


Fig. 2.3 - Admissible ranges of green intervals.

$$\left(\frac{s_1}{q_1} - 1 \right)\left(\frac{s_2}{q_2} - 1 \right) > 1$$

or, equivalently,

$$\frac{q_1}{s_1} + \frac{q_2}{s_2} < 1 \ . \qquad (2.2.6)$$

$$\frac{q_1}{s_1} + \frac{q_2}{s_2} < 1 \ . \qquad (2.2.6)$$

The relation (2.2.6) has a simple interpretation since q_1/s_1 is the fraction of the cycle time needed to serve the traffic in direction 1, and q_2/s_2 the fraction of the cycle time needed to serve direction 2. At the intersection of the two lines, equality signs apply in (2.2.5) which means that

$$1 - q_1/s_1 - q_2/s_2 = L/C \ , \qquad (2.2.7)$$

i.e., the fraction of time not needed by either directions 1 or 2 is equal to fraction of time consumed in switching, L/C .

Quite aside from any questions of "optimal strategy," (2.2.7) already describes some important facts. First of all, there is, so far, only one given parameter with dimensions of time, namely L, and any other times such as G_1, G_2, or C could be measured in multiples of L. This immediately suggests that a fluid type of theory applied to rather light traffic, say $q_1/s_1 + q_2/s_2 < 1/2$, is likely to lead to "optimal" cycle times which are so short as to violate the hypothesis that the number of vehicles served during a green interval can be approximated by a continuous variable $s_i G_i$ (the calculated value of $s_i G_i$ could be less than 1).

If, on the other hand, the approach flows are so large that one must use at least a 60-second cycle time to accommodate the traffic with $L = 12$ sec, for example, then one is using less than 1/5th of the cycle time for switching losses. By doubling the cycle time to at least two minutes, one can reduce the loss to 10 percent, which means that one can handle at most 10 percent more traffi This 10 percent may represent a significant improvement; but if one doubles the cycle time again to four minutes (which drivers would probably consider intolerable), one can gain only another five percent. Although it is true that one can increase the "capacity" of a signalized intersection by using a longer cycle time, the gains are typically rather small as compared with possible improvements that one might achieve by increasing the s_1 and/or s_2 (ban parking near the intersection, restripe the lanes, or widen the highway).
provements that one might achieve by increasing the s_1 and/or s_2 (ban parking near the intersection, restripe the lanes, or widen the highway).

Equation (2.2.7) also illustrates another important point. If direction 1 is a multilane highway but direction 2 has only a single lane, the capacity of the intersection will be more sensitive to changes in s_2 than s_1. To accommodate an increase in q_1, it may be more advantageous to ban parking or whatever on the highway in direction 2 so that one can shorten G_2 and increase G_1 than to try to make improvements in s_1.

Figure 2.3 also gives a convenient way of illustrating the effects of constraints due to pedestrian crossing times. The time G_1 is often constrained to be larger than some value G_{m1}, the time allowed for a pedestrian to cross the highway in direction 2,4. Similarly, G_2 must be larger than some number G_{m2}. The G_{m1}, G_{m2} depend on the geometry of the intersection but are independent of each other and independent of the q_i. The existence of such constraints further restricts the allowed region of figure 2.3 to the intersection of the shaded region with the quarterplane above and to the right of the two broken lines $G_1 = G_{m1}$, $G_2 = G_{m2}$.

As one can see from figure 2.3, there are many possible geometries for the intersection of these two regions, depending on the relative values of the G_{m1}, G_{m2}, and L, and the slopes of the two solid lines. For any particular intersection with specified values of the s_i, G_{mi}, and L, one should sketch the lines of figure 2.3 and see how the allowed region varies with typical values of the q_i. There is no causal connection, however, between the ratio G_{m2}/G_{m1} and the slopes of the two lines in figure 2.3. Except for low flows q_i, one would not typically expect that the intersection point (G_{m1}, G_{m2}) of the two broken lines would be inside the shaded area. If, as illustrated in figure 2.3, it lies above the shaded area, then that part of the shaded area with $G_2 > G_{m2}$ already implies that $G_1 > G_{m1}$. Similarly, if the point (G_{m1}, G_{m2}) lies below the shaded area, the shaded area with $G_1 > G_{mi}$ implies that $G_2 > G_{m2}$. Thus, we typically expect that only one or the other of the two constraints $G_1 > G_{m1}$, $G_2 > G_{m2}$ will be binding.

To determine a possible optimal G_1 and G_2 for given q_i, s_i and L, we can write the total delay per unit time as the sum of the delays per unit time in all directions, i.e., (from (2.1.7) and (2.2.4)).

$$\text{total delay per unit time} = q_1\bar{W}_1 + q_2\bar{W}_2 + q_3\bar{W}_3 + q_4\bar{W}_4$$

$$= \frac{1}{2(G_1+G_2+L)}\left\{(G_2+L)^2\left[\frac{q_1}{1-q_1/s_1}+\frac{q_3}{1-q_3/s_3}\right]\right. \tag{2.2.8}$$

$$\left. + (G_1+L)^2\left[\frac{q_2}{1-q_2/s_2}+\frac{q_4}{1-q_4/s_4}\right]\right\}$$

and the total number of stops per unit time as the sum of those on all legs, i.e., (from (2.1.5)).

$$\text{total number of stops per unit time} = \frac{1}{(G_1+G_2+L)}\left\{(G_2+L)\left[\frac{q_1}{1-q_1/s_1}+\frac{q_3}{1-q_3/s_3}\right]\right. \tag{2.2.9}$$

$$\left. + (G_1+L)\left[\frac{q_2}{1-q_2/s_2}+\frac{q_4}{1-q_4/s_4}\right]\right\} .$$

Both (2.2.8) and (2.2.9) are simple rational functions of G_1 and G_2. It is possible to minimize (2.2.8) or any linear combination of (2.2.8) and (2.2.9) with respect to G_1 and G_2, subject to the given constraints illustrated in figure 2.3, but the <u>quantitative</u> results for typical values of L are not very interesting.

If one changes G_1 and G_2 by the same amount, i.e., moves along a 45° direction in figure 2.3, one can show that (2.2.8) is a montone increasing function of G_1, G_2. This means that the minimum delay in the allowed region occurs on a boundary; at least one of the legs of the intersection will operate with its minimum green interval. Actually, the minimum delay usually occurs at a corner with both directions operating at their minimum green.

One can also show, in the absence of pedestrian constraints, that if (2.2.8) has a minimum with direction 2 at saturation but not direction 1, for example, then the value of G_2 will be appreciably less than L. For typical values of L this is, of course, meaningless because the value of s_2G_2 will be too small to justify a continuum approximation. Nevertheless, this formal

result is trying to describe a potentially important issue. If direction 2 is a minor road intersecting a major road and q_2/s_2 is small, then interrupting the major road traffic even for a minimum time of L will cause considerable delay. It may, therefore, be advantageous to continue the green on the major road after the queue has vanished until enough vehicles (at least one) have arrived on the minor road to justify another interruption of the major road traffic.

As frequently happens when one tries to make simple approximations in some optimization problem, the minimization of the approximate objective function drives the solution into a region where the approximations are not valid. If the cycle time is reduced so that there is barely enough green time to serve the average queue, then fluctuations in arrivals and departures will cause the queue frequently not to clear during the green interval. Thus, the above solution is incorrect even if it leads to long enough cycle times that the vehicles can be treated as a continuum. In most cases, however, it leads to cycle times which are so short that the signal is serving only one or two vehicles per cycle (or fractions of a vehicle). The formulas (2.2.8) may, however, give reasonable approximations if the G_i are constrained by the G_{mi}.

If we now consider the number of stops per unit time (2.2.9), we see that, for a fixed cycle time, the number of stops is least if one assigns any excess time all to one traffic direction: namely, to that direction which has the
for a fixed cycle time, the number of stops is least if one assigns any excess time all to one traffic direction; namely, to that direction which has the largest flow. Thus, this objective function would also lead to a signal setting with one of the traffic directions at saturation. In contrast with the objective (2.2.8), however, minimizing the number of stops per unit time favors arbitrarily long (infinite) cycle times. If there is no penalty associated with the duration of delays, only with the number of delays, the goal would be to maximize the fraction of time when there is no queue. The fraction of time needed to serve the traffic at flows s_1 and s_2 is $q_1/s_1 + q_2/s_2$, independent

of C, so to maximize the fraction of "free time" one should minimize the fraction of time spent on switching L/C, i.e., make C infinite. Obviously, one should not use the number of stops alone as an objective function.

If one were to minimize some linear combination of (2.2.8) and (2.2.9), there would be some competition between the two objectives with the former favoring short cycle times and the latter long cycle times. It is immaterial what price one may assign to a unit of delay. In forming a linear combination of (2.2.8) and (2.2.9), one need only assign a relative weight to these two quantities. The two quantities, however, have different physical units; (2.2.8) is dimensionless (time per unit time), but (2.2.9) has dimensions of $(\text{time})^{-1}$. If we add α times (2.2.9) to (2.2.8), the α must have dimensions of time. Specifically, one would interpret α to be the amount of delay which would be considered as the equivalent of one stop.

One could include in α some typical value for the delay downstream of the intersection caused by a stop and the time equivalent of the fuel consumption and pollution caused by a stop. The latter, of course, depend on the value of people's time and the price of fuel (or at least their relative values), but a reasonable equivalence would be of the order of 10 seconds. The delay downstream of the intersection might be somewhat less than this but of comparable magnitude. We noted previously that the only time unit in (2.2.8) was the L, downstream of the intersection might be somewhat less than this but of comparable magnitude. We noted previously that the only time unit in (2.2.8) was the L, which is rather small compared with typical cycle times for signals. We now have another time unit α, but it is of comparable size to L.

The optimal cycle time for our new objective function is some rather complicated function of the relative values of q_1/q_2, but for a symmetric intersection with $q_1 = q_2$, $q_3 = q_4$, $s_1 = s_2$, and $s_3 = s_4$ the optimal split would be for $G_1 = G_2$. In this case, the optimal cycle time would be given by

$$C^2 = L(L + 4\alpha) \qquad (2.2.10)$$

independent of q_1 or s_1 provided, however, that this cycle time satisfies the constraints (2.2.5), which in this case implies that

$$C > \frac{L}{1 - 2q_1/s_1} . \qquad (2.2.11)$$

The value of C given by (2.2.10) is not very sensitive to the α and is expected to be less than 30 seconds.

Generally, we will find that the inclusion of stops as part of the objective function does not have a large effect on the optimal strategy for ranges of flow where one needs a traffic signal. The reason is that for such ranges of flow where one cannot afford to spend much time in switching the signal, neither can one afford to operate the signal at a flow of q_i rather than s_i. One can reduce the number of stops only by a very sizeable increase in the cycle time.

2.3. Stochastic Approximations for Fixed-Cycle Signal, Four-way Intersection

We saw in the last section that if one treated the vehicle flow as if it were a continuous deterministic fluid and tried to set the cycle time so as to minimize the total delay, the cycle time would be driven to a value such that the (average) queue barely vanished at the end of a green for at least one direction (in the absence of pedestrian constraints). At this point, however, the formulas for delay are inaccurate because some vehicles will often fail to clear during the green interval. The question we will try to answer now is: how short can one make the cycle time before these "stochastic effects" overpower the other advantages of short cycle times? We will still presuppose, however, that the cycle time is long enough so that many vehicles pass during each signal phase.

If a queue fails to vanish during some green intervals, the observed cumulative arrival and departure curves might appear as in figure 2.4 rather than as in figure 2.1. If the signal is close to saturation, i.e., qC is close

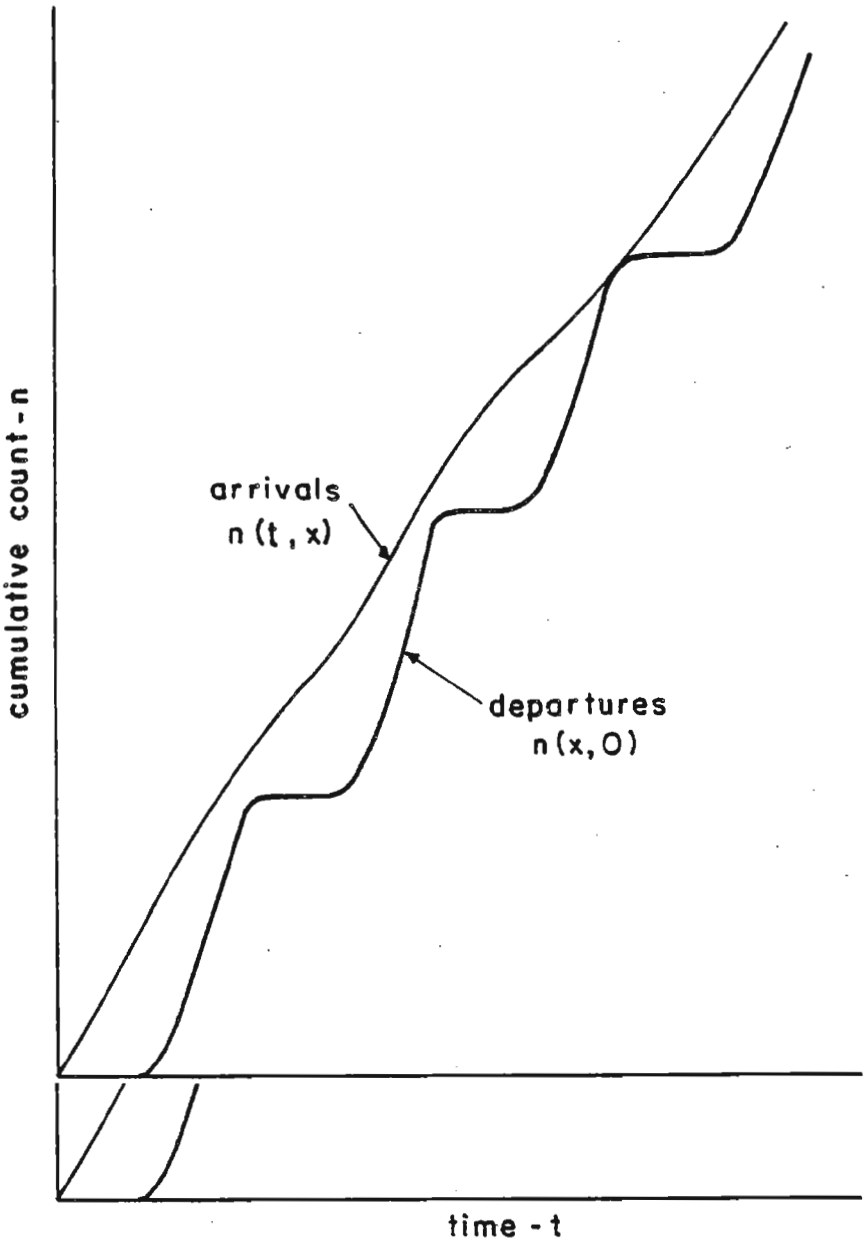


Fig. 2.4 - Typical cumulative curves with stochastic arrivals.

to sG for some traffic direction, there will not be much excess time available during a green interval to accommodate any residual vehicles which may have been left from the previous cycle. Fluctuations in arrivals or departures may cause a residual queue to persist for many consecutive cycles.

If vehicles are served in order of their arrival, a vehicle which fails to clear the intersection in one cycle is likely to pass in the next cycle because it moves toward the head of the line. The total queue length and the total delay to all vehicles, however, are independent of the order in which vehicles are served. For purposes of evaluating queue lengths or delays, it may be convenient to imagine a hypothetical queue behavior in which vehicles are served in the reverse order. Any vehicle which is delayed in one cycle will not be served in the next cycle unless and until there is some extra time left in the next cycle, after all the new arrivals have been served. Since the new arrivals will consume most (if not all) of the green interval, the residual vehicle will almost certainly wait all or nearly all of the next cycle time. The total delay during any cycle can thus be written as approximately the delay to the new arrivals as evaluated in (2.1.7), i.e., $qC\overline{W}$, plus C times the average queue Q at the start of the red. Thus

$$\text{average delay per vehicle} \cong \frac{1}{2}\left(1 - \frac{G}{C}\right)^2 \cdot \frac{C}{(1 - q/s)} + \frac{Q}{q} . \qquad (2.3.1)$$

There is a very extensive literature relating to the evaluation of the

$$\text{average delay per vehicle} \cong \frac{1}{2}\left(1 - \frac{G}{C}\right) \frac{C}{(1 - q/s)} + \frac{Q}{q} . \qquad (2.3.1)$$

There is a very extensive literature relating to the evaluation of the Q. Results of early attempts to evaluate it appear already in the classic paper by Wardrop (1951)[6]. The most commonly quoted work, however, is that of Webster (1958)[7]. There are now several different formulas of comparable accuracy, but it suffices here to describe some of the main features, their interpretations, and limitations.

The original argument of Webster was that, if one looks at the queue $Q(t)$ only at the start of successive red times, $Q(0)$, $Q(C)$, $Q(2C)$, ... , the changes

in these queue lengths from cycle to cycle would be the (actual) number of arrivals less the number of departures during one cycle. As long as $Q(mC)$, $m = 0, 1, \ldots$ remains positive, the number of departures in a cycle is the number which can leave during the whole green interval. If we ignore fluctuations in this from cycle to cycle and equate it to sG, then this number would be the same as if the vehicles departed at equally spaced time intervals of length C/sG throughout the entire cycle time C.

If the arriving vehicles form a Poisson process of rate q and they could leave at constant headways of C/sG, the average queue at the times mC would be the same as for the "well-known" queueing system with Poisson arrivals and constant service times. For such a system

$$Q = \frac{\rho^2}{2(1 - \rho)} , \tag{2.3.2}$$

in which ρ is called the "traffic intensity" in the queueing theory literature or the "degree of saturation" in the traffic signal literature.

$$\rho = qC/sG . \tag{2.3.3}$$

As applied to a fixed-cycle signal with Poisson arrivals and no fluctuations in departures, (2.3.2) overestimates the queue length because it does not account for the fact that if a queue vanishes during the green interval, it stays zero until the end of the green. It does, however, correctly de-

not account for the fact that if a queue vanishes during the green interval, it stays zero until the end of the green. It does, however, correctly de-

scribe the "equilibrium" average queue length if ρ is sufficiently close to 1 and remains (nearly) constant for a sufficiently long time, in which case we could also approximate (2.3.2) by

$$Q \cong \frac{1}{2(1 - \rho)} . \tag{2.3.4}$$

Thus, as the degree of saturation approaches 1, the equilibrium queue becomes infinite. Actually, it takes an infinite time for an infinite queue to form. One can show[8] that if the arrival rate is "slowly" increasing at a

flow $q(t)$ and degree of saturation $\rho(t)$, it would be necessary that

$$\frac{d\rho(t)}{q(t)dt} << [1 - \rho(t)]^3 \tag{2.3.5}$$

in order for the (average) queue at time t to be given by (2.3.4) with ρ equal to the prevailing value of $\rho(t)$.

The left hand side of (2.3.5) represents the change in $\rho(t)$ per inter-arrival time $1/q(t)$. We would certainly expect this to be "small" if $\rho(t)$ is "slowly varying," but "small" would mean small compared with 1. If, however, $\rho(t) = 0.9$, $1 - \rho(t) = 0.1$, then $[1 - \rho(t)]^3 = 10^{-3}$. The condition (2.3.5) would require that the change in ρ be small in the time for 10^3 arrivals (for highway traffic this typically means for a time of the order of one hour). As a practical matter, one does not expect $\rho(t)$ to vary slowly enough for any equilibrium queueing formula to be valid for $1 - \rho(t) \lesssim 0.1$, i.e., for a degree of saturation larger than about 0.9. If during a rush hour $\rho(t)$ should come this close to 1, chances are it would at some time exceed 1, i.e., the signal would be temporarily but predictably oversaturated. We will discuss this situation later but, for now, we will suppose that the $\rho(t)$ stays less than about 0.9. In any case, however, one would certainly not want to reduce the cycle time so as unnecessarily to cause ρ to be equal to 1, as implied by the deterministic fluid theory.

There are two types of modifications one may wish to make in (2.3.2) or ministic fluid theory.

There are two types of modifications one may wish to make in (2.3.2) or (2.3.4). First, one may wish to generalize the formula to situations in which the arrival process is not necessarily a Poisson process and the fluctuations in the departures are not negligible. Second, one may wish to make corrections for the fact that the $Q(t)$ may vanish before the green interval ends.

If the arrival process is stationary but vehicles cannot pass each other freely, vehicles may arrive in "clusters" of two or three at a time (cars which are obvously following each other). Whereas for a Poisson process, the variance

of the number of arrivals in one cycle should be exactly equal to the mean, namely qC, for this more general type of process the ratio of the variance to the mean is often larger than 1 (but nearly independent of the time period of counting). Suppose this is true and we define

$$I_A = \frac{\text{variance of the number of arrivals per cycle}}{qC} . \quad (2.3.6)$$

Suppose, also, that the ratio of the variance to the mean for the number of departures during a fully utilized green interval is nonzero, having a value I_D as in (1.4.3).

If now ρ is sufficiently close to 1, the equilibrium queue length is approximately[8]

$$Q \cong \frac{I}{2(1 - \rho)} , \quad I = I_A + I_D , \quad (2.3.7)$$

i.e., it is larger than in (2.3.4) by a factor $I = I_A + I_D$. If, in particular, one has Poisson arrivals and no fluctuations in departure $I_A = 1$ and $I_D = 0$ and (2.3.7) reduces to (2.3.4). One can measure the I_A and I_D but they are difficult to predict from any theory of driver behavior. One might typically expect I_A to be somewhat larger than 1, but probably less than 2, and I_D to be perhaps 1/4 to 1/2. Thus, (2.3.7) might often be larger than (2.3.4) by about a factor of 2, although no one seems to have made any serious study of this.

(2.3.4) by about a factor of 2, although no one seems to have made any serious study of this.

For any particular intersection, one can easily evaluate the I_A experimentally. If the arrival rate q is stationary, one simply evaluates the sample variance of the number of arrivals during successive cycle times as observed at some location sufficiently far upstream of the intersection that the physical queue does not back up past the observer. Since the I_A should be insensitive to the cycle time, one could also evaluate the I_A from the sample variance of of counts in any sequence of equal time intervals long enough to contain several vehicles in each interval. One must be careful, however, to make sure that the

variation in counts in successive intervals are due to "stochastic" variations, not due to some systematic trends in a time-dependent arrival rate or pulsed due to some neighboring signal. The I_D can be evaluated from a sample variance of the number of departures during fully utilized green intervals. Fortunately, for our purposes, one does not need to know these parameters very accurately (estimates within 20 or 30 percent would be quite adequate).

If ρ is not sufficiently close to 1, all of the above formulas tend to overestimate the queue length. In order for there to be a significant probability that the queue will fail to clear during the green interval, it is necessary that the fluctuations or uncertainty in the difference between the actual number of arrivals and the number of departures during any cycle exceed the average number that could be served during any excess green time. According to the above hypothesis, about the variances, the standard deviation of these fluctuations is $(IsG)^{1/2}$. Thus, the condition for there to be a significant queue (at the start of red) is

$$sG - qC \lesssim (IsG)^{1/2} ,$$

or

$$1 - \rho \lesssim (I/sG)^{1/2} . \qquad (2.3.8)$$

We are assuming here that sG is sufficiently large that the number of vehicles served during a green interval can be treated as a continuous variable,

We are assuming here that sG is sufficiently large that the number of vehicles served during a green interval can be treated as a continuous variable,

i.e., $sG \gg 1$. For a sufficiently large value of sG there would not be a significant probability of the queue failing to clear during the green unless ρ is "very" close to 1. Actually, typical "large" values of sG (say 10 or 20) are such that $(sG)^{1/2}$ is not very large (maybe 3 or 4), so one may typically start to have some overflow of the green interval if ρ exceeds about 0.7.

One can obtain a very accurate estimate of Q (more accurate than one needs, considering the typical errors involved in measurement of the relevant parameters)

if one multiplies (2.3.7) by a correction factor of the form $H(\mu)$ which is a function only of

$$\mu = (1 - \rho)(sG/I)^{1/2} \tag{2.3.9}$$

the ratio of the two sides of (2.3.8). The function $H(\mu)$[9] is analytically quite complicated, but it has been tabulated. It is a function which has the value 1 for $\mu = 0$ (ρ close to 1), decreases smoothly with μ, and decays very rapidly for μ much larger than 1.

The actual numerical values of the delay are not very important in themselves. We are mainly concerned with how the inclusion of the stochastic effects might influence one's choice of "optimal" green intervals.

If the goal is to choose G_1 and G_2 so as to minimize the total delay per unit time in all directions of the intersection, we should add to (2.2.8) the stochastic part of the delay per unit time; namely, $Q_1 + Q_2 + Q_3 + Q_4$, with Q_i the value of Q as described above for the ith direction. To do this minimization accurately leads to some very tedious algebra. Most investigators have resorted to numerical methods to evaluate the minimum, but, since there are so many parameters in the formulas, any numerical method will tend to obscure some of the relevant issues.

If directions 1 and 2 are sufficiently close to saturation that stochastic effects are important, the same is not likely to be true of directions 3 and 4.

If directions 1 and 2 are sufficiently close to saturation that stochastic effects are important, the same is not likely to be true of directions 3 and 4. To simplify the analysis, we also neglect the terms of (2.2.8) involving directions 3 and 4.

In the special case of a symmetric intersection $q_1 = q_2$, $s_1 = s_2$, we would clearly choose $G_1 = G_2$. If we use (2.2.8) along with the approximation (2.3.7) for Q,

$$\text{delay per unit time} \cong \frac{(L + G_1)^2}{(L + 2G_1)} \frac{q_1}{(1 - q_1/s_1)} + \frac{I}{1 - (q_1/s_1)(L + 2G_1)/G_1} \tag{2.3.10}$$

If the traffic signal were close to saturation, i.e., $1 - 2q_1/s_1 \ll 1$, so that one was forced to use a long cycle time with $s_1G_1 \gg 1$, then the fluctuations in the number of arrivals per cycle (of order $(s_1G_1)^{1/2}$) would be small compared with the mean number of arrivals (approximately s_1G_1). One might think that if the fractional fluctuations in the number of arrivals per cycle were small compared with 1, the deterministic approximation would be "accurate." However, the size of the queue is determined by the relative magnitude of the fluctuations as compared with the ability of the system to accommodate these fluctuations, i.e., with the excess capacity. As noted before, increasing the cycle time from a value which is already large compared with L does not increase the capacity very much. Indeed, increasing the cycle time will not necessarily eliminate the stochastic queue.

This competition between the fluctuations and the capacity is illustrated in the denominator of the second term of (2.3.10).

$$1 - \frac{q_1}{s_1G_1}(L + 2G_1) = \left(1 - 2\frac{q_1}{s_1}\right) - \frac{q_1L}{s_1G_1}.$$

The first term, $(1 - 2q_1/s_1)$, is the fraction of time available for serving the fluctuations or switching the signal. It is assumed to be small compared with 1. The second term must be less than the first if the system is undersaturated, but even making G_1 infinite would not eliminate this term.

For $1 - 2q_1/s_1 \ll 1$ we cannot simplify the second term of (2.3.10) very saturated, but even making G_1 infinite would not eliminate this term.

For $1 - 2q_1/s_1 \ll 1$ we cannot simplify the second term of (2.3.10) very much but we can simplify the first term because, under these conditions $L \ll G_1$, and $q_1/s_1 \cong 1/2$. Thus, the first term of (2.3.10) will be approximately q_1G_1. With this approximation, it is now easy to determine a G_1 so as to minimize (2.3.10); namely

$$G_1 \cong \left[\frac{(q_1/s_1)L}{(1 - 2q_1/s_1)}\right]\left[1 + 2\left(\frac{I}{Ls_1}\right)^{1/2}\right], \qquad (2.3.11)$$

which gives a minimum value of (2.3.10) of approximately

$$\text{minimum delay per unit time} \cong q_1 \left[\frac{(q_1/s_1)L}{1 - 2q_1/s_1} \right] \left[1 + 2 \left(\frac{I}{Ls_1} \right)^{1/2} \right]^2 \qquad (2.3.12)$$

We recognize that the first factor of (2.3.11) is the minimum allowed green interval, the optimal green interval in the deterministic approximation. The effect of fluctuations ($I > 0$) is to increase the optimal green interval (and the cycle time) by a numerical factor, the second factor of (2.3.11). The Ls_1 represents the number of vehicles which could be served during a time L. For a single lane road with headways of 2 seconds ($s_1 = 1/2$ sec^{-1}), $L = 12$ seconds, $s_1L = 6$, and $I = 1.5$, the second factor of (2.3.11) is 2, i.e., the optimal green interval is approximately twice the minimum green interval. For multi-lane highways s_1L would likely be larger than 6, but increasing s_1 by a factor of 2 decreases $s_1^{-1/2}$ only by a factor of $2^{-1/2}$. The second factor of (2.3.11) is not very sensitive to changes in the I, L, or s_1.

Webster[7, 10] did extensive numerical calculations with more accurate formulas for Q and also arrived at a simple recipe that the optimal cycle time, with stochastic effects, was approximately <u>twice</u> the minimum cycle time (his calculations were with $I = 1$, however).

One should also take notice here that the goal is not to determine the optimal value of G_1 within some specified error. The goal is to find a nearly minimum total delay. At the minimum delay, however, the delay as a optimal value of G_1 within some specified error. The goal is to find a nearly minimum total delay. At the minimum delay, however, the delay as a function of G_1 has a vanishing derivative. Typically, the fractional deviation of the delay from its minimum is comparable with the square of the fractional deviation of G_1 from the optimal, i.e., a 10 percent error in G_1 will cause only about a one percent increase in delay. It suffices to have only some crude estimates of the optimal G_1.

In (2.3.12), the first two factors represent the minimum delay for the deterministic approximation $I = 0$. The last factor is the square of the

corresponding factor in (2.3.11). Thus, for typical values of the parameters, the inclusion of stochastic effects increases the average delay per vehicle or per unit time by approximately a factor of 4!

The deterministic approximation was not only mathematically inaccurate, it also gave cycle times which were often unrealistically small; for example, a cycle time of about 20 seconds for $2q_1/s_1 = 1/2$. Actually the range of "acceptable" cycle times is not very large. Even without any restrictions due to pedestrian crossing times,one would not likely choose a cycle time less than 30 seconds ,i.e., an (effective) green interval less than about 10 seconds. But (for a two-phase signal) one would not likely use a cycle time much more than two minutes. Thus, we are admitting only about a factor of 4 range of acceptable cycle times. The inclusion of the stochastic effects at least puts the typical "optimal" cycle time into a reasonable range. If, for $2s_1/q_1 = 1/2$, we take twice the minimum cycle time we obtain a 40-second cycle time, which should be acceptable. On the other hand, one would not find it advantageous to use a cycle time greater than two minutes unless $1 - 2q_1/s_1 \lesssim 1/5$ which is rather heavy flow for an undersaturated signal ($\rho \cong 0.9$) .

It is possible to generalize (2.3.11) to nonsymmetric intersections. If we still neglect the flows in directions 3 and 4, one can show that the optimal cycle time is

cycle time is

$$C \cong \left[\frac{L}{(1 - q_1/s_1 - q_2/s_2)}\right]\left\{1 + \left[\frac{I_1 s_2}{Lq_2(s_1 + s_2)}\right]^{1/2} + \left[\frac{I_2 s_1}{Lq_1(s_1 + s_2)}\right]^{1/2}\right\}, \tag{2.3.13}$$

in which I_1 and I_2 are the values of the I for directions 1 and 2, respectively. The first factor is the minimum cycle time. The value of the second factor is typically similar to the corresponding factor in (2.3.11),

although somewhat larger if the values of q_1/s_1 and q_2/s_2 differ by a large amount (a factor of 4 perhaps).

The more interesting question for an asymmetric intersection is how one should partition any excess capacity between the two traffic directions. If $1 - q_1/s_1 - q_2/s_2$ is small, one does not actually have much "free" time to partition. For any cycle time C, one needs a time Cq_1/s_1 to serve the average number of arrivals in direction 1, a time Cq_2/s_2 to serve those in direction 2 and one loses a time L in switching. For a cycle time C given by (2.3.13), the only free time one has to distribute per cycle is

$$C\left(1 - q_1/s_1 - q_2/s_2\right) - L = L\left\{\left[\frac{I_1 s_2}{Lq_2(s_1 + s_2)}\right]^{-1/2} + \left[\frac{I_2 s_1}{Lq_1(s_1 + s_2)}\right]^{-1/2}\right\}, \tag{2.3.14}$$

The quantity on the right hand side is typically comparable with L, i.e., about 10 seconds.

Since the amount of free time one has to distribute for $C/L \gg 1$, is small compared with C one might say that, to a "first" approximation, the optimal value of G_1 is just Cq_1/s_1 (plus a small amount) and that the ratio G_1/G_2 is approximately $(q_1/s_1)/(q_2/s_2)$. If we chose to partition the cycle time <u>exactly</u> in this ratio, this would imply that we were also partitioning the G_1/G_2 is approximately $(q_1/s_1)/(q_2/s_2)$. If we chose to partition the cycle time <u>exactly</u> in this ratio, this would imply that we were also partitioning the free time (2.3.14) in this ratio. This is the usual recipe given in traffic engineering books for partitioning the cycle time. Webster[7, 10] has also shown from some numerical optimization that the optimal ratio of G_1 to G_2 is approximately this.

Actually, the total delay is quite sensitive to how one partitions the free time. This "free time" is not an observable time. It represents (by definition) a hypothetical average time per cycle when the signal would be

completely idle if all vehicles which arrived when the queue was zero could be served at a rate s_j instead of q_j. This "free time" is thus equivalent to an addition to the lost time L. Since the cycle time is some (large) multiple of L, this can be interpreted as the reason why the introduction of a free time of a magnitude comparable with L itself will increase the C by a substantial factor. The reason for introducing this free time was to reduce the stochastic queueing but, of course, one should distribute this free time so as to reduce the queues in each of the two directions by similar amounts.

The two terms on the right hand side of (2.3.14) are in fact the optimal amounts of free time to be assigned to directions 1 and 2 respectively. One should distribute this free time not in the ratio of $(q_1/s_1)/(q_2/s_2)$ but in the ratio

$$\left[\frac{I_1(q_1/s_1)}{I_2(q_2/s_2)}\right]^{1/2} . \qquad (2.3.15)$$

Thus, for $I_1 = I_2$, if the minimum green intervals are in the ratio $(q_1/s_1)/(q_2/s_2)$ of say 2 to 1, one would split the free time in the ratio of $\sqrt{2}$ to 1.

If, in the deterministic approximation, one were to choose the G_1 and G_2 so as to minimize the total delay, subject to the condition that the cycle time C was assigned some value larger than its minimum value, one would certainly give any excess green time all to the same direction; namely, that with the larger q_i. One of the interesting features of the result (2.3.15) is tainly give any excess green time all to the same direction; namely, that with the larger q_j. One of the interesting features of the result (2.3.15) is that it depends on the (q_j/s_j) but not on the q_j's themselves. The reason for this strange result is that the residual queue at the start of red, the Q of (2.3.7), depends on the degree of saturation ρ but is otherwise independent of q and s. Thus, if one had a two-lane approach intersecting a one-lane approach so that $s_1 = 2s_2$, for example, one also had $q_1 = 2q_2$ so that $q_1/s_1 = q_2/s_2$ and one split the cycle time so that $G_1 = G_2$, then both directions would have the same value of ρ and therefore the same value of Q.

Since both queues contribute the same to the total delay, it is equally important that one try to reduce them both.

In the above formulas, we have neglected the effect of any traffic in directions 3 and 4. In the (unlikely) symmetric situation in which $q_1 = q_3$, $s_1 = s_3$, $q_2 = q_4$ and $s_2 = s_4$, the delays would be the same in directions 1 and 3, and in directions 2 and 4, for any signal setting. The total delay would be just twice that for direction 1 and 2 alone, so the optimal signal setting would be the same as described above for just two traffic directions $(q_3 = q_4 = 0)$.

It is interesting to compare the above illustration with $s_1 = 2s_2$, $q_1 = 2q_2$, $q_3 = q_4 = 0$ with a case in which $s_1 = s_2 = s_3$, $q_1 = q_2 = q_3$ and $q_4 = 0$. In the former situation we, in effect, have two lanes of traffic in direction 1, each carrying a flow equal to q_2, with a saturation flow per lane of s_2, but no traffic in direction 3. In the latter situation we have the same flow per lane, but the flows are in opposing directions (1 and 3). The difference between the two situations is that, with two adjacent lanes, vehicles can switch lanes and keep the signal busy as long as there is a queue (in either lane). For two lanes in opposing directions, however, one can have a (residual) queue in direction 1 during some signal cycles when there is none in direction 3, or vice versa.

(residual) queue in direction 1 during some signal cycles when there is none in direction 3, or vice versa.

There is no way that a vehicle traveling in direction 1 can take advantage of any excess green time for direction 3. As the formulas show, the latter case has twice as much stochastic queueing in directions 1 plus 3 as the former has in direction 1. With equal queues in directions 1, 2, and 3 (but not 4) one would, of course, tend to partition more green time to directions 1 and 3 than to directions 2 and 4.

The more likely situation is that q_3/s_3 and q_4/s_4 are smaller than q_1/s_1 and q_2/s_2, respectively, by enough so that the Q_3 and Q_4 are much

smaller than Q_1 and Q_2, because the Q_i are very sensitive to the degree of saturation in the ith direction. To reduce the deterministic part of the queue in directions 3 and 4, however, these traffic directions would favor a shorter cycle time, or, in particular, direction 3 would prefer a smaller G_2 (a shorter red time for direction 3) and direction 4 would prefer a smaller G_1.

In the special case of symmetric flows $q_1 = q_2$, $s_1 = s_2$, $q_3 = q_4$, $s_3 = s_4$ one can easily show that the generalization of (2.3.11) for $q_3 > 0$ gives

$$G_1 \cong \left[\frac{(q_1/s_1)L}{(1 - 2q_1/s_1)}\right]\left[1 + 2\left\{\frac{I}{Ls_1}\frac{q_1}{(q_1 + q_3)}\right\}^{1/2}\right]. \qquad (2.3.16)$$

In effect, one simply replaces the I of (2.3.11) by $Iq_1/(q_1 + q_3)$. Since this factor is inside the square root, and $q_3 < q_1$, q_3 can reduce this term by at most $1/\sqrt{2}$, possibly reducing the second factor of (2.3.11) from a value of 2 to 1.7. This conclusion, however, is based upon an objective in which delays are weighted equally for all vehicles and also linearly in the length of the delay. Actually traffic engineers (and presumably society generally) would tend to assign more weight to long delays than to short delays. They would prefer to reduce the longer delays in directions 1 and 2 than to reduce the delays in directions 3 and 4. They might even disregard the delays in directions 3 and 4 completely.

completely.

For a nonsymmetric intersection, the flows, q_3 and q_4 would together want to reduce the cycle time but if $q_3 > q_4$ the flow q_3 might favor splits giving somewhat less green to directions 2 and 4. If one assigns more weight to the longer delays in directions 1 and 2, however, one may choose to ignore the delays in directions 3 and 4.

As was discussed in the previous section, we could take as the objective function some weighted sum of delays and stops.

The existence of stochastic queueing has no effect on the fraction of the cycle time when there is no queue, i.e., on the fraction of vehicles which are not stopped. The total number of vehicles which are stopped per unit time is, therefore, still given by (2.2.9). However, this formula would assign only one stop to any vehicle which was stopped during one cycle but had to wait several cycles before it could leave. Since a vehicle which fails to clear the intersection during the cycle of arrival will need to start and stop during each cycle as it moves forward in the queue, one may wish also to add a penalty for each of these stops.

The average number of stops per cycle of the latter type is simply the residual queue Q at the start of the red time and the number of such stops per unit time is Q/C (or Q_i/C for direction i). If one wished to assign an equivalent delay time of α' for each such stop, one should add $\alpha' Q_i/C$ to the objective function. The α' is not the same as the α of the last section because the α presumably represents the effective cost of stopping from the approach speed v and accelerating back to this speed, plus any time lost downstream of the intersection. The α' includes only fuel consumption, pollution, etc., needed to stop and go in the queue but not any actual delay.

Since the expression for delay per unit time already contains a term Q_i, a stop and go cost would simply enlarge this term by a factor of $(1 + \alpha'/C)$ with a value of α' probably less than ten seconds. Inclusion of this factor a stop and go cost would simply enlarge this term by a factor of $(1 + \alpha'/C)$ with a value of α' probably less than ten seconds. Inclusion of this factor in the objective function would tend to increase the optimal cycle time (so as to reduce the effects of queueing), but obviously not by very much (typically by a factor of only about $(1 + \alpha'/2C)$, probably less than five percent).

If we add to the objective function a multiple α of the number of (primary) stops per unit time (2.2.9), this would also tend to increase the optimal cycle time, but not very much. Even though stops may represent a significant part of the social cost, nearly all vehicles are stopped no matter what the cycle

time may be if $1 - q_1/s_1 - q_2/s_2$ is small. It might be better to represent this part of the objective function as $\alpha(q_1 + q_2 + q_3 + q_4)$ less α times the number of vehicles per unit time which are not stopped. The fraction of vehicles which are not stopped, however, is, at most, comparable with L/C. Actually, the inclusion of stops will increase the cycle time only by a factor of approximately $(1 + L\alpha/C^2)$ in which both L/C and α/C are expected to be small compared with 1. It will also tend to favor giving a little more of the free time to the traffic direction with the largest q_i, but this is also a small effect.

One of the interesting facts that one should notice about the theory presented so far (for an isolated signal with moderately heavy demand but no turns) is that a rational choice of the green intervals G_1 and G_2 is sensitive to only a relatively small number of parameters which one might use to describe the system. We have actually introduced 15 parameters, q_i, s_i, I_i $(i = 1, \ldots, 4)$, L, α and α' but some of these (particularly the L and I_i) themselves describe only certain relevant aspects of the traffic behavior which might have been described in terms of many more parameters. We have argued, however, that the traffic in directions 3 and 4 and the number of stops typically have little influence on the choice of G_1 and G_2. This reduces the number of relevant parameters to 7, q_i, s_i, I_i $(i = 1, 2)$ and L.

The minimum cycle time depends on q_1/s_1, q_2/s_2, and L whereas the parameters to 7, q_i, s_i, I_i $(i = 1, 2)$ and L.

The minimum cycle time depends on q_1/s_1, q_2/s_2, and L whereas the relevant aspects of the stochastic behavior, (2.3.13) and (2.3.15), depend on these plus I_1, I_2, and $L(s_1 + s_2)$, only six parameters. The formulas do <u>not</u> also depend on s_1/s_2. Actually, the choice is not very sensitive to the I_1, I_2 and $L(s_1 + s_2)$, it depends mostly on just the q_1/s_1, q_2/s_2, and L.

So far, we have been concerned with the minimization of the total delay with no constraints due to pedestrians. If we think of the total delay as some function defined on the space G_1, G_2 of figure 2.3, we do not need to impose explicitly the condition that G_1, G_2 must be in the shaded region of figure 2.3 because the stochastic queueing terms become infinite along the boundary of this region. The analytic minimization automatically gives a solution in the interior of this region. If, however, we further impose a condition that $G_1 > G_{1m}$ and/or $G_2 > G_{2m}$, the unconstrained minimum might not satisfy these conditions. If it does (for sufficiently large q_i) there is no problem, but if it does not, then surely the constrained optimal setting would be on one of the constraint boundaries, $G_1 = G_{1m}$ and/or $G_2 = G_{2m}$.

In the shaded region of figure 2.3, the total delay is most sensitive to the two stochastic queueing terms for directions 1 and 2. Whether we consider the sum of these terms along a line corresponding to a fixed value of C, G_1, or G_2, one is led to nearly the same condition, namely that any excess time allocated to direction i above the minimum needed, $G_i - (q_i/s_i)C$, should be split as in (2.3.15),

$$\frac{G_1 - (q_1/s_1)C}{G_2 - (q_2/s_2)C} \cong \left[\frac{I_1(q_1/s_1)}{I_2(q_2/s_2)}\right]^{1/2} . \qquad (2.3.17)$$

In the G_1, G_2 space, (2.3.17) is the equation of a straight line contained in and passing through the corner of the shaded region of figure 2.3. The unconstrained minimum for the total delay, of course, lies on this line at a cycle time given by (2.3.13), but if the constraints $G_1 > G_{1m}$ and/or $G_2 > G_{2m}$ exclude this point, the constrained minimum delay will be at the intersection of the line (2.3.17) with the boundary of the constrained region.

The customary procedure for choosing splits with pedestrian constraints, presumably is to match the degree of saturation in directions 1 and 2. This

would correspond to replacing the line (2.3.17) by a line of slope $(q_1/s_1)/(q_2/s_2)$ passing through the corner of the shaded region of figure 2.3 and also through the origin.

We will postpone a discussion of the effect of turning movements until the end of this chapter not because they are typically of minor importance, but because the number of potentially important parameters explodes. Not only does one need at least four more parameters to describe the fractions of (left) turning traffic in the four directions, but the delays and strategies may now be sensitive to the q_3/s_3 and q_4/s_4 as well as q_1/s_1 and q_2/s_2.

2.4. Time-dependent Arrivals, Oversaturated Intersections

In the previous section, we imagined that the arrival rates q_i were time-independent, although they never are. We have, however, defined a time-dependent flow $q_i(t)$ in (1.4.2) as the time derivative of a "smoothed" cumulative arrival curve, smoothed by averaging counts over many similar days or by some artificial recipe. Although this smoothing averages out stochastic fluctuations, it will not average out any time-dependence which is reproducible from day to day (or week to week) such as rush hour effects. The typical pattern is that $q_i(t)$ will rise to a maximum during morning and evening peaks giving maximum values of $q_1(t)/s_1 + q_2(t)/s_2$ possibly close to or exceeding 1, but very small values in the middle of the night. We expect, however, that the $q_i(t)/s_i$ will not change very much during any cycle time (a few minutes).

We have previously interpreted a "fixed-cycle" (F-C) signal as one with time-independent values of G_1, G_2 and $C = G_1 + G_2 + L$, but we will now interpret this to mean any signal strategy which is based upon the $q_i(t)/s_i$ (and $I_i(t)$ if the I_i are not constant). We assume that any relevant information about the $q_i(t)$ is obtained from occasional traffic surveys (manual counts or counts from temporarily installed counters) or estimates based on

typical rush hour profiles. The traffic signal operates on the same daily or weekly pattern, possibly for several months or until one has reason to believe that the $q_i(t)$ has changed significantly. The signal is driven by a clock and some predetermined pattern.

In contrast with this, a vehicle-actuated (V-A) signal by definition includes equipment to detect information about current traffic conditions. Presumably the signal system could also retain any relevant information extracted from the vehicle detectors, including historical data. It could evaluate and update estimates of the $q_i(t)$, among other things, for any traffic directions or lanes where there are detectors (although standard equipment does not do this). In principle, a V-A signal could be programmed to out perform a F-C signal according to any specified criteria because it has more complete data with which to operate. A V-A signal, however, costs more to install and maintain than a F-C signal (approximately twice as much). The difference in annual cost (including interest on the investment) is of the order of a few thousand dollars and the savings in delays, etc., does not always justify this expense.

The formulas of the last section have given some possible choices of G_1 and G_2 as a function of the parameters q_i/s_i, etc. If the parameters are slowly varying with time, one could select values $G_1(t)$, $G_2(t)$ based upon current values of the parameters, selecting slight different values in each successive cycle. Most F-C traffic signals are not equipped to follow concurrent values of the parameters, selecting slight different values in each successive cycle. Most F-C traffic signals are not equipped to follow continuously a predetermined 24-hour or weekly program, but such equipment could be built. They can at least switch cycle times a few times each day (either manually or automatically).

The formulas of the last section, however, apply only over a range of values for $q_1/s_1 + q_2/s_2$ from about 0.4 to 0.9. For values below this range, the proposed cycle time is likely to be less than about 30 seconds. One has the option then of either choosing some minimum acceptable cycle time or using a flashing red and yellow, i.e., in effect, converting the signal into a two-way

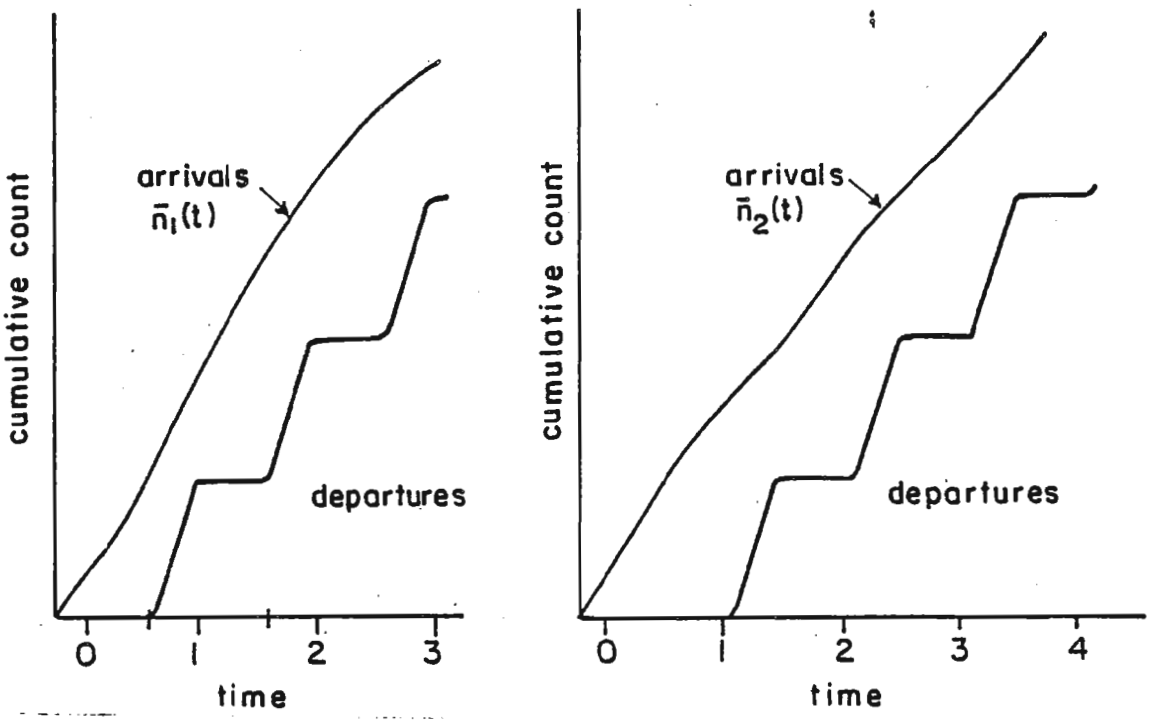


Fig. 2.5 - Cumulative arrivals and departures at an intersection for directions 1 and 2.

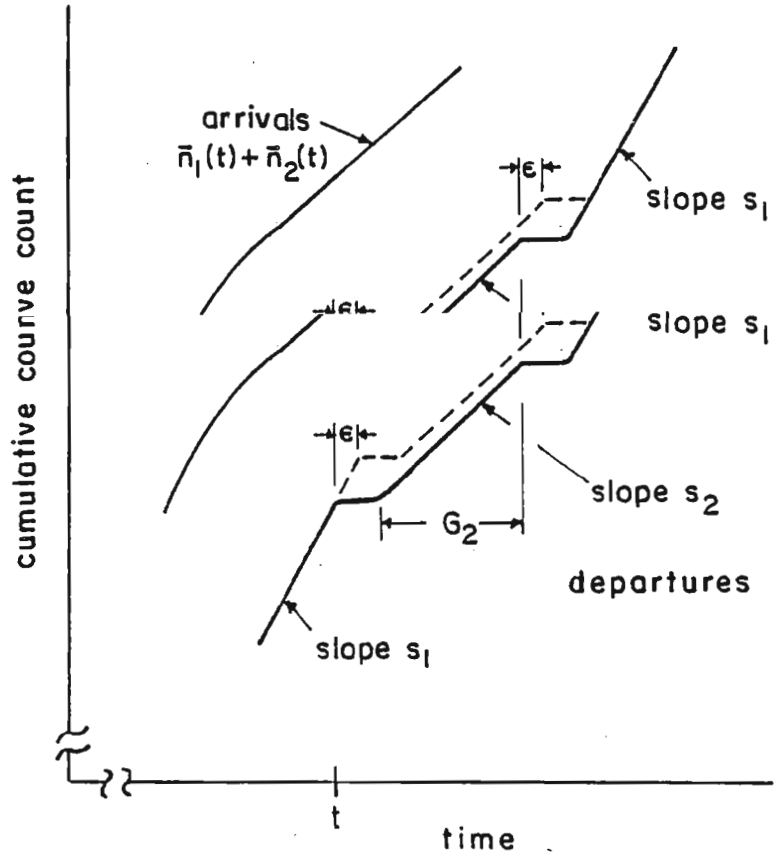


Fig. 2.6 - Combined counts for directions 1 and 2

stop sign. Actually, this choice is seldom made with much consideration of delay to vehicles. If it were, greater use would be made of the latter option.

It is quite common for $q_1/s_1 + q_2/s_2$ to exceed 0.9 or even 1 during peak demand. Indeed, a highway network which has no congestion even during the peak period is overbuilt. Also in this range of demand, commonly used strategies of control are based upon criteria other than minimizing total delays or stops.

If $q_1(t)/s_1 + q_2(t)/s_2$ should actually exceed 1, there is some rather complex queue behavior during the "transition period" when $q_1(t)/s_1 + q_2(t)/s_2$ has a value between about 0.9 and 1. No matter how one varies the cycle time during this transition, the time-dependent queue cannot keep pace with the changing equilibrium queue behavior and when $q_1(t)/s_1 + q_2(t)/s_2$ exceeds 1 there is no equilibrium. The signal is oversaturated and queues will grow until the demands drops below this critical value.

Suppose one can construct from historial data a curve of $\bar{n}_i(t)$, $i = 1, \ldots 4$, the (average) cumulative number of vehicles to arrive in direction i by time t, on a time scale covering the whole rush hour, and they have shapes such as shown in figure 2.5 for $i = 1, 2$. We can also sum the graphs in any combinations but, in particular, we will wish to consider graphs of $\bar{n}_1(t) + \bar{n}_2(t)$ and $\bar{n}_1(t) + \bar{n}_2(t) + \bar{n}_3(t) + \bar{n}_4(t)$.

For any signal strategy we could also draw graphs of the corresponding (average) cumulative number of vehicles to leave in each direction or combina-

For any signal strategy we could also draw graphs of the corresponding (average) cumulative number of vehicles to leave in each direction or combina-tions thereof (or piecewise-linear approximations to such curves). For a fixed-cycle strategy one has the option of switching the signals at any time based upon the properties of these averaged curves (whereas for a V-A signal one chooses the switching times on any day depending on the actual curves $n_i(t)$ for that day).

Prior to the time when the signal becomes nearly saturated, it is possible to switch signals in such a way that the queue vanishes at the end of the green interval during at least some signal cycles, and, in any case, so that the average

queue at the start of red has some (nearly) equilibrium value if the $q_i(t)$ stay (nearly) constant. This would be true for all traffic directions i. When the system becomes oversaturated, however, this is no longer possible, but one can illustrate the consequences of any proposed strategy on a figure like figure 2.5.

When the signal is green in directions 1 and 3, the departure curves in directions 1 and 3 will have slopes s_1 and s_3, respectively, as long as the queue is positive. Similarly, when the signal is green in directions 2 and 4, the departure curves in directions 2 and 4 have slopes s_2 and s_4 respectively. During the (effective) lost time L there are no departures in any direction. The departure curve for directions 1 and 2 will alternate between slopes s_1 and s_2 but with zero slope during a total time L for any switch from one direction to the other and back, provided both queues stay positive. The combined departure curves for directions 3 plus 4 behave similarly except that in these directions it is likely that the queues will vanish in each cycle, if they are undersaturated. The departure curve for the sum of all directions will switch between slopes $s_1 + s_3$ and $s_2 + s_4$ but with s_i replaced by $q_i(t)$ if the queue in direction i vanishes.

If the objective were to minimize the total delay during the whole rush hour, one would want to minimize the area between the curves for the combined arrivals and departures in directions 1 to 4. Since the curves $\bar{n}_i(t)$ are assumed to be given, this objective is equivalent to moving the combined departure curves for $i = 1$ to 4 as high as possible (specifically maximizing the area under the departure curve) through appropriate choice of switching times, but recognizing that the slope of an individual departure curve will drop from s_i to $q_i(t)$ if the queue vanishes in direction i.

If directions 3 and 4 are undersaturated, most of the delay would be in directions 1 and 2 identified by the area between the curves for arrivals and

departures in just directions 1 and 2. Suppose, for now, we disregard the vehicles in directions 3 and 4.

If one has followed some strategy until some time t, the signal is green in direction 1, there is a positive queue in direction 1 (for certain, every day), $s_1 > s_2$, and one wishes to minimize total delay, then would not switch the signal to direction 2 at time t and suffer a lost time in switching the signal to operate at a lower flow s_2. The proof of this intuitively plausible conclusion is illustrated in figure 2.6. If the departure curve for directions 1 plus 2 had slope s_1 at time t, one switched the signal time at time t so as to give a slope s_2 for some period of time G_2 and then switched back to direction 1, the departure curve would be as illustrated by the solid line of figure 2.6. If, however, one could postpone the switching time to time $t + \epsilon$ and maintain the flow s_1, one could give the same green interval G_2 to direction 2 as before and follow the broken line curve of figure 2.6 until it rejoined the solid line curve in the next cycle. If $s_1 > s_2$, the broken line curve is everywhere higher than the solid line and therefore gives less total delay. The conclusion is that, to minimize the delays in directions 1 plus 2, one will always hold the signal green for direction 1 at least until the average queue is so small that there will be a zero queue on a significant fraction of days causing the average departure rate to drop below s_1.
queue is so small that there will be a zero queue on a significant fraction of days causing the average departure rate to drop below s_1.

If one were to hold the signal green in direction 1 after the queue vanished, the departure rate would drop from s_1 to $q_1(t)$ within a short (compared with C) time. From figure 2.6 it appears that there might be a temporary reduction in delay, particularly if $q_1(t) > s_2$, but, unlike figure 2.6, one does not now have a strategy which will return the system to the same state as the solid line in the next cycle. Operating with a flow less than s_1 during one green interval reduces the time during which the signal can operate with a flow s_1 in later cycles. Except in very unusual circumstances, the net result

is an increase in total delay. One should not allow the flow to drop below s_i during any green interval since this is essentially equivalent to an increase in the effective lost time L.

The next question is: having switched the signal to direction 2 (with $s_2 < s_1$), how long should one extend the green time for direction 2? Since, presumably, $q_2(t)/s_2 < 1$, one could extend the green until the queue in direction 2 is nearly zero. Indeed, one could follow a strategy of switching the signal only when the queue is nearly zero, for both directions 1 and 2, throughout the whole rush hour. This is, in fact, the strategy which minimizes the sum of delays in directions 1 and 2 if $s_1 = s_2$, since the argument of figure 2.6 applies for both directions.

If one were to follow such a strategy, successive cycle times would increase until the total demand rate was less than the average service rate. If the $q_i(t)$ varied only slowly with time, the number of arrivals in direction 1 during a cycle time C would be approximately $q_1(t)C$ and the subsequent green interval G_1' needed to serve them would be $G_1' \cong q_1(t)C/s_1$. Similarly for direction 2. The next cycle time would then have a value

$$C' \cong G_1' + G_2' + L \cong \left[\frac{q_1(t)}{s_1} + \frac{q_2(t)}{s_2}\right]C + L \,. \qquad (2.4.1)$$

Thus, if $q_1(t)/s_1 + q_2(t)/s_2 > 1 - L/C$ (the signal is oversaturated), $C' > C$.

Thus, if $q_1(t)/s_1 + q_2(t)/s_2 > 1 - L/C$ (the signal is oversaturated), $C' > C$.

Despite the fact that this is the strategy which minimizes total delay for $s_1 = s_2$ (and $q_3 = q_4 = 0$), this is not the type of strategy commonly used, and it is doubtful that travelers would accept it. A typical intersection which is oversaturated during the rush hour is likely to generate queues causing delays of five or ten minutes (or more) to individual vehicles. With the above strategy, however, this is the magnitude of C. One should also notice that any vehicle which is stopped (namely all of them during this period of oversaturation) will

clear the intersection during the cycle of its arrival. Thus, this strategy also minimizes the number of stop and go movements in the queue (namely, zero).

Maybe if the traffic engineer published a schedule of the signal timing (like a bus schedule), travelers would learn how to adjust to this strategy so as to reduce their uncertainty of travel time. Perhaps travelers should also make reservations. Otherwise, it seems that travelers would prefer to move in the queue every minute or two, so that they have a feeling of making some progress even if this results in increased delay (which the traveler does not appreciate since he cannot control it) and increased fuel consumption.

One could artificially assign a maximum value to C (two minutes, for example) and then devise a strategy to minimize delay (or other cost) subject to this constraint. This is what is commonly proposed but this, in effect, adds a zero penalty for any cycle time less than the maximum and an infinite penalty for a cycle time larger than the maximum. This does not seem to be a very realistic interpretation of what society likes. On the other hand, it is very difficult to determine what travelers would prefer. Since one cannot eliminate the congestion by simply changing the signal timing, travelers will complain no matter what one does. There seems to be no way of formulating and measuring an objective function which describes what society dislikes the least.

and measuring an objective function which [illegible] the above strategy of switching least.

If, however, $s_1 > s_2$ (and $q_3 = q_4 = 0$), the above strategy of switching the signal from direction i only when the queue is (nearly) zero in direction i for i = 1 and 2 is not the strategy which minimizes the total delay. One would still do this for i = 1, but while the signal is green for direction 2, the queue is growing in direction 1. There may come a time (before the queue vanishes in direction 2) when it is advantageous to pay the penalty for switching the signal back to direction 1 to take advantage of the larger service rate s_1 in direction 1.

Suppose that at (approximately) time t when the arrival rates are $q_i(t)$, one chooses a cycle time $C(t)$ but partitions it so that the queue vanishes in direction 1, i.e., so that

$$G_1(t)s_1 = C(t)q_1(t)$$

and

$$G_2(t) = C(t) - G_1(t) - L = C(t)[1 - q_1(t)/s_1] - L \ .$$

The queue in direction 2 will now grow from one cycle to the next by approximately an amount

$$C(t)q_2(t) - G_2(t)s_2 = s_2 \left\{ C(t) \left[\frac{q_1(t)}{s_1} + \frac{q_2(t)}{s_2} - 1 \right] + L \right\}$$

or at an average rate per unit time of this divided by $C(t)$, i.e.,

$$\text{rate of increase of the queue in direction 2} \cong \left\{ \frac{q_1(t)}{s_1} + \frac{q_2(t)}{s_2} - 1 + \frac{L}{C(t)} \right\} s_2 \ . \qquad (2.4.2)$$

The queue in direction 1 clears every cycle but it has an average value of approximately half its maximum,

$$\text{average queue in direction 1} \cong \frac{q_1(t)G_1(t)}{2} = \frac{q_1^2(t)C(t)}{2s_1} \ . \qquad (2.4.3)$$

One still has the option of choosing the cycle time $C(t)$ at various times during the rush hour and one could do this in such a way as to minimize the sum of the delays in directions 1 plus 2 throughout the whole rush hour. The solution of this optimization problem is somewhat complicated and not worth describing in detail, but the qualitative aspects are fairly simple and illustrate some important issues.

The previous strategy in which the queues clear in each direction is a special case of the present strategy in which the cycle time increases fast

enough so as to absorb the increase in the queue (2.4.2), but the implications in the approximate relations above is that $C(t)$ will not increase this rapidly. One will allow a queue to grow in direction 2.

That it is more desirable to let the queue grow in direction 2 than in direction 1 can be seen from (2.4.2). If we had reversed the roles of the two directions and allowed the queue to vanish in direction 2 instead of 1, we would have the same formulas but with the indices 1 and 2 reversed. The first factor of (2.4.2) is independent of the order of the indices 1 and 2 but the factor s_2 would be replaced by s_1. Thus for any specified $C(t)$ and $s_1 > s_2$, the growth of the queue will be less if one assigned the growth to direction 2 than to direction 1. More generally, suppose one were to specify the $C(t)$ but had the option of selecting the partition of it between $G_1(t)$ and $G_2(t)$. If there were queues in both directions, any shift of some green time from direction 2 to direction 1, i.e., $G_1 \to G_1 + \varepsilon$ and $G_2 \to G_2 - \varepsilon$, would give an increase of $(s_1 - s_2)\varepsilon$ in the number of departures during the cycle and a corresponding decrease in the total queue. Clearly, if the objective is to minimize the total delay and a queue is to accumulate somewhere, it should be in the direction with the smaller s_i.

As regards the choice of the cycle time itself, the competition is between the delay in lane 1, (2.4.3), which is proportional to $C(t)$ and the $L/C(t)$ term of (2.4.2). The $L/C(t)$ is a relatively "small" term (one cannot increase the delay in lane 1, (2.4.3), which is proportional to $C(t)$ and the $L/C(t)$ term of (2.4.2). The $L/C(t)$ is a relatively "small" term (one cannot increase the capacity of an intersection very much by increasing C, if C is already large compared with L) but any decrease in the queue length at the start of the rush hour will persist until the queue vanishes at the end of the rush hour. The contribution of this term in (2.4.2) is "weighted" by the duration of the rush hour. If, for example, by increasing $C(t)$ one eliminated one switching time L, one could reduce the queue by $s_2 L$ vehicles (typically about 5) for the remainder of the rush hour.

The optimal $C(t)$ will generally decrease with t but the typical value near the start of the rush hour will be such that $C^2(t)$ is comparable with $[s_1 s_2 L/q_1^2(t)]$ times the duration of the rush hour. If, for example, $L = 1/5$ minutes, $q_1/s_1 = 1/2$, $q_1/s_2 = 1$ and the oversaturated queue lasts for 20 minutes, $C(t)$ would be about $(20 \times 2/5)^{1/2} \sim 3$ minutes. Such a cycle time may be somewhat higher than travelers would tolerate, but, at the optimal $C(t)$, the delay is not very sensitive to the value of $C(t)$. Perhaps the use of a cycle time of only two minutes would not increase the total delay very much. It is worth noting, however, that to reduce the total delay it is more important to take some action early when the queue first starts to grow then to wait until the queue is already large, because any addition to the queue will last for the remainder of the rush hour.

The aspect of the above strategy which is most controversial is that it forces most of the delays onto the travelers in direction 2 (the direction with the smaller s_i, $i = 1, 2$). This comes about because we assume, for example, that one minute of delay to each of ten travelers in direction 1 is equivalent to ten minutes of delay to one traveler in direction 2. Since travelers in direction 1 use their green time more efficiently than those in direction 2, we give them priority.

One could propose a more realistic objective function in which the effective direction 2, we give them priority.

One could propose a more realistic objective function in which the effective "price" of a unit of delay for a traveler who has already waited a time w, $p(w)$, is an increasing function of w. As the delays increase in direction 2, they may reach a state such that $p(w_1)/s_1 = p(w_2)/s_2$ in which w_1 and w_2 are the waiting times of the travelers at the front of the queue in directions 1 and 2 respectively. At this stage both signal phases would reduce the value of time in the queue at the same rate and one would no longer give complete priority to direction 1. Such a cost structure would also increase the penalty for long

cycle times with large oscillations in the waiting times of travelers. For appropriate choice of the $p(w)$ and an objective function equal to the sum of the costs to all travelers, one could probably formulate an "optimal strategy" which is compatible with what society is willing to accept.

The problem with such a theory is that it is mathematically rather complicated but also logically difficult to apply. Since one does not know $p(w)$ and would have difficulty measuring it directly (since travelers themselves do not know what they like and different travelers behave differently), one would probably have to infer an effective value of $p(w)$ by "calibrating" the theory so that it gave results consistent with what one believes is acceptable. But if one knows what is acceptable, one does not need a theory.

Although assigning nearly all of the delays to direction 2 may not be acceptable, one can certainly argue that some preference should be given to travelers who use the intersection more efficiently. Not only would this reduce the total delay for an inelastic demand; but if the demand is elastic, one does not want to attract more travelers who use the system inefficiently and cause long delays to others.

Traffic engineering books suggest that the green intervals be split in the ratio of q_1/s_1 to q_2/s_2. Presumably this recipe is meant for undersaturated signals but, since this is not explicitly specified, traffic engineers would ratio of q_1/s_1 to q_2/s_2. Presumably this recipe is meant for undersaturated signals but, since this is not explicitly specified, traffic engineers would likely use this even if the signal is oversaturated. The consequence of such a strategy is that the queue lengths in directions 1 and 2 grow in the ratio q_1 to q_2 but the waiting times in queue are nearly the same in both directions (independent of the s_i!). This may be very "democratic" but it also means that to achieve this, one is willing to trade a unit of delay to one person in direction 2 for a unit of delay to each of more than one person in direction 1. This makes no sense either.

In the discussion above we have neglected the traffic in directions 3 and 4. If, for the signal setting chosen on the basis of the traffic in directions 1 and 2 only, the queue vanishes in directions 3 and 4 every cycle, the delays in the latter directions will be small compared with those in directions 1 and 2. Although directions 3 and 4 would likely prefer a shorter cycle time, they would typically have little effect on the optimal setting. There are exceptions, however, actually a large number of them, since we have postulated only that $q_1(t)/s_1 < q_3(t)/s_3$ and $q_2(t)/s_2 < q_4(t)/s_4$. It is possible (but unlikely), for example, that $q_1(t)/s_1$ is nearly equal to $q_3(t)/s_3$ or $q_2(t)/s_2$ is nearly equal to $s_4(t)/s_4$. Also, it is possible that $s_3 > s_1$ and/or $s_4 > s_2$.

For any choice of a cycle time $C(T)$ it would be desirable to partition the time so as to minimize (or reduce) the rate of growth of the total queue, i.e., to maximize the number of vehicles served per cycle. The number of vehicles served in direction i per cycle is $s_i G_i(t)$ or an average of $s_i G_i(t)/C(t)$ per unit time, provided there is a queue in this direction throughout the green interval; otherwise the number is $q_i C(t)$ per cycle or q_i per unit time, independent of any changes in the signal setting.

One could have some very complex strategies if the relative ordering of the $q_i(t)/s_i$ values changed during the rush hour. We will assume this does not happen.

the $q_i(t)/s_i$ values changed during the rush hour. We will assume this does not happen.

When the signal first becomes oversaturated, one should give preference to direction 1 if $s_1 > s_2$, but give direction 1 barely enough green time to clear the queue. To give direction 1 more than this, for a fixed $C(t)$, would decrease the rate of service $s_2 G_2(t)/C(t)$ in direction 2 but would not increase the rate q_1 in direction 1. Since $q_3(t)/s_3 < q_1(t)/s_1$, the queue would certainly vanish in direction 3 if it vanishes in direction 1. Since also $q_4(t)s_4 < q_2(t)/s_2$

and $q_2(t)/s_2$ is only slightly larger than $G_2(t)/C(t)$, presumably $q_4(t)s_4$ is, at first, less than $G_2(t)/C(t)$, so that the queue vanishes in direction 4 also.

If we continue to give preference to direction 1 as $q_1(t)/s_1$ increases, $G_2(t)/C(t)$ will decrease. As a result of this, there may come a time when $q_4(t)/s_4 = G_2(t)/C(t)$; direction 4 is at the brink of saturation. If we allow queues to form in direction 1, 2, and 4, the rate of departures in directions 2 plus 4 will be $(s_2 + s_4)G_2(t)/C(t)$ while that in directions 1 plus 3 is $s_1G_1(t)/C(t) + q_3$. If, for fixed $C(t)$, we now should shift an amount ε of green time from direction 1 to direction 2, this would increase the total departure rate by $(s_2 + s_4 - s_1)\varepsilon$. If $s_1 > s_2 + s_4$, we would still give preference to direction 1 but if $s_1 < s_2 + s_4$, we would prefer to give more green time to directions 2 and 4.

In the latter case, the conclusion is that one would give directions 2 and 4 barely enough green time to prevent a queue from growing in direction 4, i.e., choose $G_2(t)$ so that $G_2(t)/C(t) = q_4(t)/s_4$, but allow a queue to form in both directions 1 and 2. One could object to this strategy on the grounds that one is still trading longer delays to one person against shorter delays to many people. Travelers in direction 2 still suffer the most, but they would view direction 1 as their competitors and perhaps would not complain if travelers in direction 1 were also delayed. They may think it somewhat

would view direction 1 as their competitors and perhaps would not complain if travelers in direction 1 were also delayed. They may think it somewhat arbitrary that their delays are dictated by the traffic in direction 4 (if they realize that this is true) but any other strategy would also be "somewhat arbitrary."

If the flows $q_j(t)$ should continue to increase, they might reach such a value that direction 3 and 4 are at the brink of saturation. If a queue forms in all four directions, the departure rates in directions 2 plus 4 and 1 plus 3 would be $(s_2 + s_4)G_2(t)/C(t)$ and $(s_1 + s_3)G_1(t)/C(t)$. One would

now give preference to directions 1 plus 3 if $s_1 + s_3 > s_2 + s_4$, allowing queues to grow only in directions 1, 2, and 4. Otherwise one allows the queue to grow in directions 1, 2, and 3.

As the queues decrease toward the end of the rush hour, one can follow the same sequence of strategies in reverse order. For example, if with $s_1 < s_2 + s_4$ and no queue in direction 3, one had kept $G_2(t)/C(t) = q_4(t)/s_4$ but the queue should vanish in direction 1, one would switch to the strategy of giving direction 1 only enough time to clear the queue, $G_1(t)/C(t) = q_1(t)/s_1$.

2.5. Vehicle Actuated Signals - General Properties

A V-A signal system has vehicle detectors on some or all approach lanes to an intersection. Information from these detectors is transmitted to a control system which, in principle, can switch the signal according to any strategy based upon current or historical data obtained from these detectors plus any other relevant data such as the time of day, day of the week, or special events, and any time-independent information such as the geometry of the intersection, location of detectors, or supplementary data obtained from traffic surveys (for example, turning movements which cannot be observed from the vehicle detectors). A F-C signal could presumably use all of the same data except for the data obtained from the vehicle detectors.

hicle detectors). A F-C signal could presumably use all of the same data except for the data obtained from the vehicle detectors.

The primary purpose of a traffic signal is to avoid collisions between vehicles which are following the legal rules for drivers and even to minimize the chance of collision between vehicles which are not quite following the rules. Vehicles traveling in different directions must be served in some sequence and one typically wishes to do this so as to minimize delays, stops, or some other objective. The purpose of the vehicle detectors is to provide information which may be relevant to reduce the delays or stops (without increasing the risk of collision).

It is possible to collect vast amounts of irrelevant or redundant information. One could place detectors every 20 feet or so in every lane of every approach and even within the intersection to detect turning movements, but detectors are expensive to install and also expensive to maintain. Unfortunately, most vehicle detectors are not as reliable as one would like them to be and are not generally equipped to detect their own failure. The system should certainly be designed to collect only information which is most relevant to a rational decision of when to switch the signal and to collect this information as cheaply as possible. Although microcomputers could be programmed to store information and follow complex procedures, it would also be desirable to keep the logic simple, if it is effective.

Almost any detector (pressure pad, photocell, or induction loop) can be used to record the time at which some part of the vehicle passed some point on the highway. Most of them can also record how long a time a vehicle is over a detector, and induction loop detectors can record if there are any vehicles in some section of the highway. From the shape of pulses one may possibly also infer some information about the type of vehicle (truck, bicycle, car). Not all this information may be relevant, so we will not comment further on the details until we see what information we want.

There is some information that would be useful which one cannot obtain. details until we see what information we want.

There is some information that would be useful which one cannot obtain. A driver who will be instructed to stop at the intersection must be informed of this in time for him to decelerate to a stop, but the decision must be based on whether or not the intersection will be clear of other conflicting vehicles at the time he would arrive if he did not stop. One would thus like to have precise information about where vehicles will be five or ten seconds in the future. If one observes the velocity of a vehicle, one could extrapolate its motion into the future, but one must use a signal strategy which is safe even if the driver does something unusual (but legal). Gathering more information about present

or past behavior will not necessarily improve one's prediction of what could happen in the immediate future.

Another potentially useful type of data is the length of the queue. If, however, the physical queue extends so far back from the signal that the end of the queue fails to cross the detector even during the green interval of the signal (as in an oversaturated condition), the detector will merely record the periodic flow induced by the output from the signal. This flow will be essentially independent of the queue length and consequently give no information about the queue length. The detectors will merely verify whether or not vehicles are leaving the intersection at the expected rate (there are no stalled vehicles, accidents, slow trucks, or whatever).

The result of any strategy is simply a decision to switch a signal or not. A V-A signal can be very effective for the control of light traffic in which the signal would merely act to resolve conflicts between individual vehicles, but the primary purpose is to control moderate to heavy (undersaturated) traffic (a V-A signal is too expensive as a control for only light traffic but offers no special advantage for the control of oversaturated intersections). As compared with a F-C signal, a V-A signal might be able to (a) reduce the effective lost time L in switching and/or (b) respond more effectively to fluctuations in the number of vehicles to arrive from one cycle to the next, which causes stochastic queueing for the F-C signal.

of vehicles to arrive from one cycle to the next, which causes stochastic queueing for the F-C signal.

As regards the former issue, consider some hypothetical vehicle trajectories as shown in figure 2.7 for vehicles passing an intersection during one cycle in a single lane. We are primarily interested here in when to switch the signal from green to yellow and from yellow to red, but the figure shows possible trajectories for the entire cycle because any information which one could obtain about the details of these trajectories might be relevant to the decision of when to switch the signal. The typical pattern is that when the signal turns

red, a queue propagates upstream. Figure 2.7 illustrates a situation in which there is no residual queue initially. The approximate location of the end of the physical queue is indicated by the broken line. After the signal turns green, vehicles accelerate and acceleration waves propagates upstream. The rear of the queue, which had been moving upstream, gradually reverses direction. When the vehicles are moving, however, we would describe them as a "platoon." New vehicles are still overtaking and joining the platoon as the rear end of the platoon passes the intersection and proceeds downstream (if the signal is still green).

For a F-C signal we do not know if there is a residual queue at the start of red (for the optimal signal setting and moderately heavy traffic, there usually will be one) and we do not know whether or not the platoon in the current cycle will clear the intersection. At some time t_y when the signal switches from green to yellow, vehicles may be passing the intersection at a speed v_p typical of vehicles in the platoon (corresponding to a flow of approximately s) or, if the platoon has already passed the intersection, at a speed v typical of the approach speed (corresponding to a flow of approximately q).

At the time t_y there may be vehicles so close to the intersection that they cannot possibly stop, and there may be other vehicles whose drivers are uncertain as to whether or not they can or should stop. For any speed v, one could identify a "minimum stopping distance" and a "legal stopping distance" from the intersection. Anyone beyond the latter distance at time t_y who fails to stop would be considered to be in violation of the law.

For a F-C signal, one does not know where the vehicles may be when the signal turns from green to yellow. The duration of the yellow must therefore be chosen so that a vehicle which is anywhere within the legal stopping distance (in particular the furthest point) and traveling at a safe and legal speed can legally enter the intersection before the signal turns red. If one could

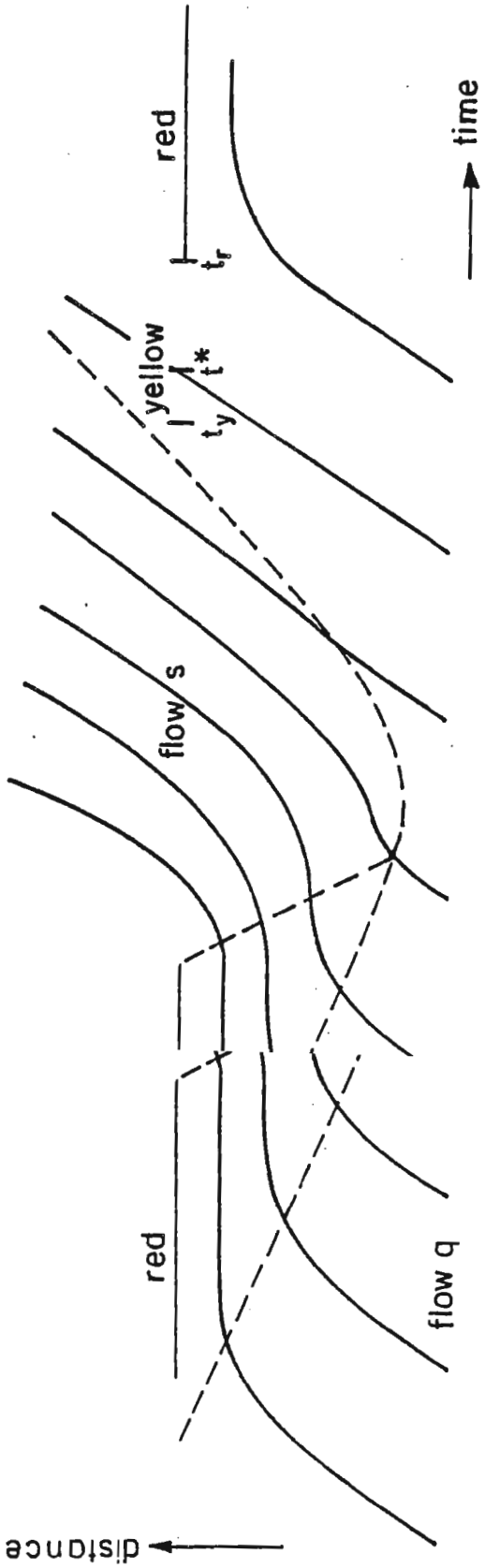


Fig. 2.7 - Some typSome typical trajectories for vehicles passing a signal.

reasonably expect vehicles to decelerate at a rate a and they are traveling at a speed v when the signal turns yellow, the stopping distance would be approximately $v^2/2a$. The yellow interval should be approximately the time to travel this distance without stopping (at velocity v). Thus

$$\text{yellow interval} \sim v/2a \qquad (2.5.1)$$

Actually, the yellow interval should be a bit longer than this. The above formula does not make allowance for the reaction time of the driver or the fact that a vehicle must either stop behind the stop line of the intersection or clear some point within the intersection before the signal turns red.

For $a = 8\text{ft/sec}^2$ and $v = 80$ ft/sec, the time (2.5.1) is about 5 sec (high speed road) but for $v = 50$ ft/sec it would be about 3 sec (urban area). It is an increasing function of v which means that one would need a somewhat longer yellow interval if the signal switched after the platoon had passed the intersection than if the signal switched while the platoon was still crossing the intersection. For a F-C signal, however, one does not know if the platoon will clear the intersection (it certainly will in some cycles, if the intersection is undersaturated) so one must choose the yellow interval on the basis of the higher (approach) speed. Of course, if the drivers knew the duration of the yellow interval and they are in a platoon when the signal switches, they may realize that they can "sneak through" the yellow without violating the law even though interval and they are in a platoon when the signal switches, they may realize that they can "sneak through" the yellow without violating the law even though they are able to stop. There have been extensive experiments done on driver reaction to a yellow signal, but the driver behavior may be different in different locations. Also it is difficult to determine whether one is choosing the yellow interval in response to driver behavior or driver behavior is a reaction to customary choices of the yellow interval.

The lost time L for the F-C signal was defined in terms of "effective green intervals," but it obviously includes parts of two yellow intervals plus

the effects of start-up times. The optimal cycle times and waiting times were, however, all multiples of L. If, for a V-A signal, one could, in effect, reduce the L by even one or two seconds per phase change, this could give a nontrivial reduction in the average delay per vehicle.

Actually for a F-C signal the "effective lost time" may vary somewhat depending on whether or not the queue vanishes during one or both green intervals. In any case, when the signal turns yellow there will typically be vehicles so close to the intersection that they cannot stop. Particularly if the queue does not vanish during the green interval, one can expect that the flow through the intersection will remain at approximately s for part of the yellow interval. This is a "flow" averaged over many cycles because this flow lasts for only about 2 seconds of the yellow interval and therefore represents an average of only about one vehicle (per lane).

Drivers in specified positions should respond to the signal changes in the same way whether the signal is a F-C or V-A signal. The advantage of the V-A signal is that it can potentially choose the times t_y and t_r when the signal switches from green to yellow and from yellow to red, respectively, based upon any prior information from the detectors. We will verify later (at least for the intersection of two one-way streets) that it is usually advantageous to choose the time t_y at approximately the time when the rear of the platoon passes the intersection. For a F-C signal one does not know if or when this happens and, for a V-A signal, this depends on the fluctuations in the cumulative arrivals during the current and previous signal cycles.

For a V-A signal it is still true that if a vehicle is (barely) within a legal stopping distance of the intersection at the time t_y and traveling at speed v, then it should be able to enter the intersection before the signal turns red at some time t_r. Thus, if this is a possibility, one must provide the same yellow interval for a V-A signal as for a F-C signal.

There are two ways in which a V-A signal might achieve a shorter (average) lost time per cycle than a F-C signal. First, one might be able to choose the time t_y so as to give a higher likelihood that the flow through the intersection will maintain a value of approximately s for a short time (about 2 seconds) after the time t_y. To do this, the signal must switch to yellow at least about 2 seconds before the end of the platoon is expected to reach the intersection. This will already be the case for a significant fraction of the cycles for a F-C signal. One might not be able to do much better with a V-A signal than for a F-C signal, but, if one is not careful, the V-A signal might give an appreciably larger lost time than a F-C signal by switching the signal so late that the flow drops below s too soon after the time t_y or (even worse) before the time t_y.

The other possible way of reducing the lost time is to terminate the yellow interval whenever one can be certain that no vehicles could (legally) pass the intersection during the remaining scheduled yellow interval. One might be able to reduce the scheduled yellow interval if one can be certain that any vehicle which might enter the intersection during the yellow would do so at the speed v_p rather than v. Better yet, one might be able to observe the time t^* of figure 2.7 when the last vehicle which could legally pass the intersection actually does so, and switch the signal to red as soon as this vehicle has legally entered the intersection.

The benefit from these (small) savings in time is not just that the vehicles queued in the cross direction can save a little time (which would itself be of little consequence). The main benefit derives from a "chain reaction". If, at the time the signal switches to red, one knows that no vehicle will arrive for a short time interval, then this means that the queue will not start to grow immediately during the next red interval. Also, if the queue on the cross direction is served earlier, the signal in the original direction will switch to green

earlier in the next cycle, and in all subsequent cycles. The effect of a reduction of the lost time for the V-A signal is analogous to the reduction of the scale of time for the F-C signal.

It is fairly clear what one would like to achieve with a V-A signal. The next question is: how can one do it? In particular, one would like to know when the last vehicle in the platoon will pass the intersection, preferably at least 2 seconds before it actually happens.

The simplest type of V-A signal systems would have at most one vehicle detector in each approach direction or lane. It might be either an "impulse" detector which records the passing times of each vehicle or a loop detector which would record the presence or absence of vehicles over some length of roadway. In either case, one might expect that, for moderately heavy traffic, the physical queue would propagate upstream of the detector during the red interval. When this happens, the detector can no longer record the expected arrival times of new vehicles and consequently cannot determine the cumulative number of vehicles to join the queue since the start of the red interval, so as to estimate the approximate amount of green time needed to serve the queue.

After a signal turns yellow, the controller could record the cumulative count of the number of vehicles which are likely to be stopped between the detector and the intersection. After the queue overruns the detector, however, an impulse detector would see a very "long headway" (no vehicles passing the detector and the intersection. After the queue overruns the detector, however, an impulse detector would see a very "long headway" (no vehicles passing the detector) whereas a loop detector would record that the loop is occupied. After the signal turns green, a starting wave propagates upstream and, when it reaches the detector, vehicles start to cross the detector again, possibly with a couple of relatively long headways at first. After a while, however, the headways should approach a mean value of about $1/s$, and vehicles should be traveling at nearly a constant mean speed v_p between the detector and the intersection. In this situation, no information from possible cumulative counts of vehicles

passing the detector or the intersection is of any help in predicting when the end of the queue will pass the detector or the intersection. The detector cannot collect any useful information about the end of the platoon until the end of the platoon has crossed the detector on its way to the intersection.

The detector observes individual vehicles; it does not measure "flows." To measure a flow with any precision one would need to observe several sucessive headways and take their average. If, however, one must wait until several vehicles have passed the detector after the end of the platoon has passed before one can make a measurement of the flow and infer that the flow has indeed dropped to some value q less than s (typically about half of s), the end of the platoon may have long since passed the intersection before the inference has been made. It is necessary, therefore, that one make such an inference as quickly as possible, even at the risk that the inference is incorrect.

The usual method for doing this with an impulse detector is to observe the elapsed time since the last vehicle passed the detector. If vehicles are passing the detector at a flow s, the mean headway between vehicles is $1/s$ (about 2 seconds for one lane) and this elapsed time should seldom be much larger than $1/s$ (at most 3 seconds, perhaps). As soon as the detector has recorded an elapsed time exceeding some specified value β (perhaps 3-1/2 or 4 seconds), it is quite likely that the preceding vehicle was the last vehicle in the platoon and that subsequent vehicles will pass with a mean headway of $1/q$ seconds, it is quite likely that the preceding vehicle was the last vehicle in the platoon and that subsequent vehicles will pass with a mean headway of $1/q$ (perhaps 4 seconds). There is, of course, a danger that the long headway is "accidental," caused by some slow vehicle which failed to keep up with the platoon, but that subsequent headways will again be approximately $1/s$.

Suppose, instead, one were to use a loop detector of a length βv_p, and vehicles were traversing the loop at a constant speed v_p. If some vehicle crossed the outer edge of the loop and no other vehicle followed it within a time β, then the loop would be empty as soon as this vehicle crossed the inner

edge of the loop. Whereas the impulse detector would observe a headway larger than some minimum β, the loop detector would observe a spacing larger than some minimum βv_p. If the velocity of vehicles are indeed constant, the two types of detectors would respond in essentially the same way. The impulse detector has the advantage that one can easily change the β at different times of the day, if this is desirable, but it is difficult to change the length of the loop. The impulse detector has the potential disadvantage that when vehicles are accelerating immediately after the starting wave passes, the impulse detector may record a couple of long headways (which it should ignore). The spacing between such vehicles, however, will likely be shorter than βv_p, so there is less danger that a loop detector would falsely infer that the end of the platoon was passing.

If one uses the above strategy (with either type of detector), it is important that the detectors be sufficiently far from the intersection and that β not be too large. One should not switch the signal from green to yellow until the controller has recognized the end of the platoon, a time at least β after the end of the platoon passes an impulse detector or at the time it passes the inner edge of a loop detector. If the objective is to switch the signal when the end of the platoon is within a minimum stopping distance of the intersection, then the inner edge of the loop detector should be this far from the intersection. The outer edge of the loop detector or an impulse de-

the intersection, then the inner edge of the loop detector should be this far from the intersection. The outer edge of the loop detector or an impulse detector should be a distance

$$\beta v_p + (\text{minimum stopping distance for } v_p) \qquad (2.5.2)$$

from the intersection. For β = 3-1/2 seconds and v_p = 30 ft/sec (for an urban intersection), βv_p is about 100 ft (by present standards a rather long loop) and the stopping distance is about 60 ft, the impulse detector

should be about 160 ft from the intersection.

Choosing a β which is larger than necessary is inefficient for two reasons. First, for an impulse detector, it would require that one place the detector further from the intersection. Second, and more important, the controller might permit one or more vehicles to pass the intersection after the end of the platoon with headways larger than $1/s$ but less than β, causing an average "flow" less than s immediately after the signal switches and possibly even before it switches. Even if the β were chosen as $1/q$, for example, and the controller failed to respond to a headway slightly less than $1/q$, but then the next headway was also slightly less than $1/q$, one should conclude that one has probably already made a mistake; the platoon has already passed the detector but one has not yet seen a headway larger than β. For $\beta > 1/q$, of course, this situation would occur quite frequently. With a detector as far away from the intersection as (2.5.2), however, one might argue that, if the control allowed just <u>one</u> vehicle to pass the detector with a headway of about $1/q$, this vehicle might be able to catch up with the end of the platoon before the platoon passes the intersection.

We next turn to the question as to whether or not one might be able to terminate the yellow interval short of the usual yellow interval for the velocity v, again under the situation in which the queue has propagated upstream of the terminate the yellow interval short of the usual yellow interval for the velocity v, again under the situation in which the queue has propagated upstream of the detector during the preceding red interval. For this, one must distinguish between two situations depending upon whether the distance (2.5.2) is larger or less than the legal stopping distance for velocity v. At an urban intersection these two distances are likely to be similar but, for a high speed rural road, the latter distance would be larger.

Suppose first that the legal stopping distance for velocity v is less than the distance (2.5.2) from the detector to the intersection. Figure 2.8 illustrates some possible trajectories for vehicles near the end of the platoon

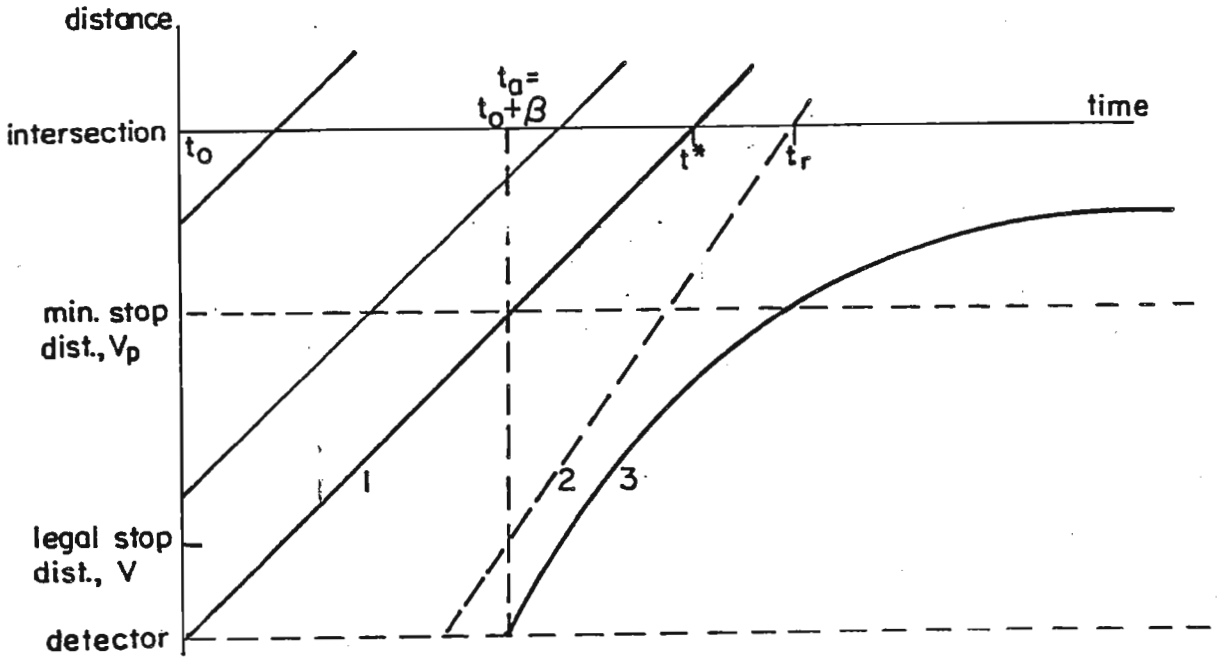


Fig. 2.8 - Vehicle trajectories near a signal at the termination of green, detector farther away than the legal stop distance.

as they approach the intersection. At time t_0 vehicle 1 , the last vehicle in the platoon, passes the detector at speed approximately v_p , but the detector does not recognize that this is the last vehicle until time $t_0 + \beta$. If the detector is located as in (2.5.2), vehicle 1 reaches its minimum stopping distance at this time and the signal switches to yellow at time $t_y = t_0 + \beta$. If a vehicle 3 were to pass the detector anytime after time t_y and the legal stopping distance were closer to the intersection than the detector, this vehicle would be required to stop.

The usual yellow interval would be defined by a hypothetical trajectory 2 which passes the legal stopping distance at the time t_y and velocity v . But if the signal switches because no vehicle crossed the detector between times t_0 and $t_0 + \beta$, then there is no such vehicle. The signal could safety switch to red at time t^* after vehicle 1 passes the intersection.

In the above situation one cannot predict exactly when vehicle 1 will reach the intersection because the detector is an appreciable distance from the intersection. One does know, however, that vehicle 1 was following another vehicle with a headway less than β and therefore was traveling at some speed v_p less than the approach speed v. Also the time t_y was chosen so that this vehicle should be so close to the intersection at time t_y that it could not stop. If, for some reason, this vehicle was traveling slower than expected and was far enough away from the intersection at time t_y that it could stop, it would be necessary to provide at most a safe yellow interval for a speed v_p and switch the signal to red at a time $t_r' < t_r$.

This strategy would already guarantee that the flow remains at a value of approximately s for part of the yellow interval and also has a yellow interval for speed v_p rather than v. One could do even better than this if one had a second detector closer to the intersection, which could accurately evaluate the time t^* when vehicle 1 actually passed the intersection. This is where a loop detector would be particular effective. As soon as a loop detector close to the intersection observes that there are no vehicles within a certain distance of the intersection (at most the legal stopping distance for v_p) after time t_y, it could infer that vehicle 1 had passed and the signal could switch to red. Such a strategy would virtually guarantee a flow of approximately s throughout the yellow interval and even save an additional fraction of a headway because the yellow terminates with the actual passage of a vehicle. The main constraint here is that the loop detector should not be so long that, when vehicle 1 leaves the detector, a following vehicle which is decelerating to stop should not yet have reached the outer edge of the detector.

Now suppose that the legal stopping distance for velocity v is larger than the distance (2.5.2) to the detector. Figure 2.9 illustrates some possible trajectories. Again vehicle 1 is the last vehicle in the platoon but it is not

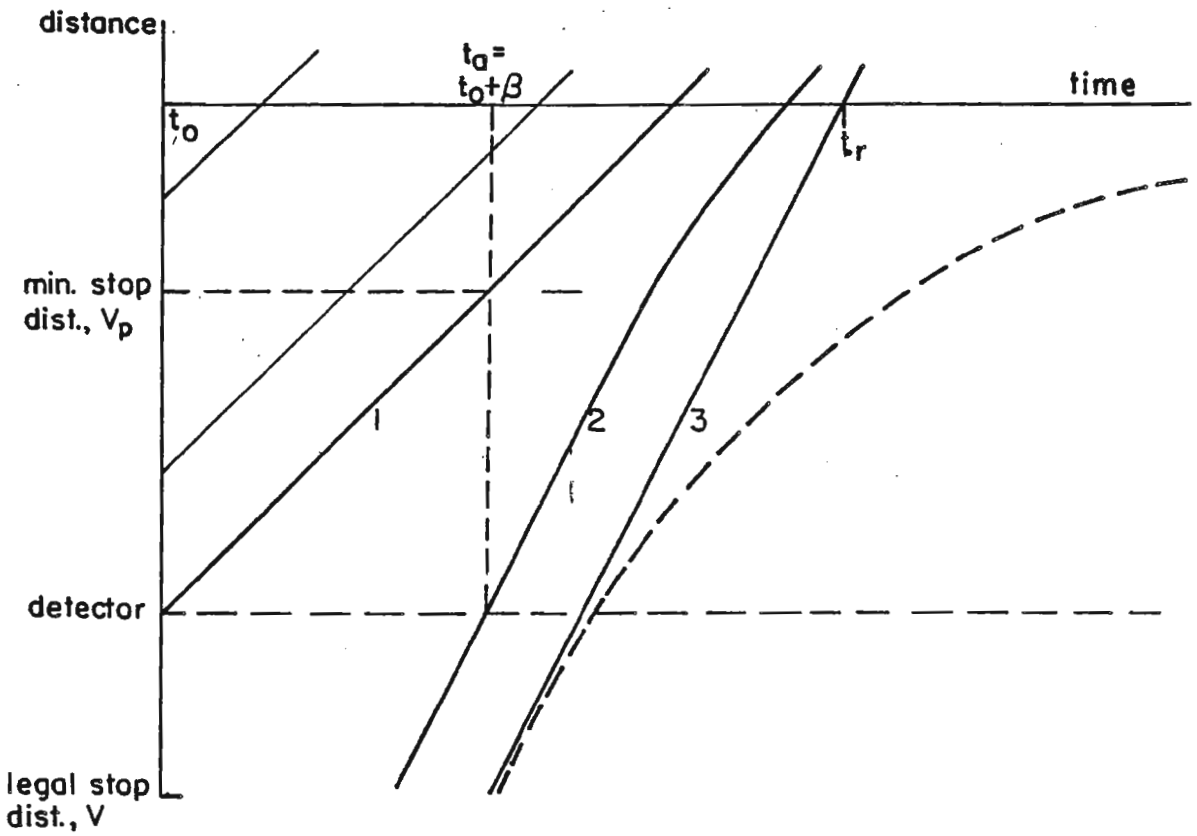


Fig. 2.9 - Vehicle trajectories near a signal at the termination of green, detector closer than the legal stop distance.

recognized as such until time $t_0 + \beta$, at which time the signal switches to yellow. Now, however, a vehicle which crosses the detector (immediately) after time $t_0 + \beta$ (such as vehicle 2) is not legally required to stop, but it might overtake the platoon and be slowed down. Indeed a vehicle 3 which has barely reached the legal stopping distance at time t_y must also be allowed to continue. Note that trajectories 2 and 3 are limiting individual trajectories. As drawn, the trajectories are too close together for there to be vehicles following both trajectories simultaneously.

In contrast with figure 2.8, the fact that no vehicle passed the detector between times t_0 and $t_0 + \beta$ does not itself guarantee that one can reduce the yellow interval below the usual value for velocity v. If, perchance, the

detectors observed that no vehicles passed it during the longer time interval until after trajectory 3 of figure 2.9 passed, then the signal could switch to red at this time or at the time trajectory 1 passes the intersection, whichever is later (one cannot use information until after one obtains it). Otherwise, it will be difficult to save any yellow time based on information from this single detector alone. Any vehicle between trajectories 2 and 3 of figure 2.9 should easily pass the intersection before time t_r (or stop), but it is difficult to estimate when the last one reaches the intersection since some vehicles may be decelerating as they approach the rear of the platoon. One would need another detector closer to the intersection to gather more reliable information.

Figure 2.9 illustrates one of the key problems. If one tries to choose the time t_y so as virtually to guarantee that the last vehicle in the platoon will clear the intersection (it is within a minimum stopping distance for speed v_p at time t_y), then one also risks the possibility that a vehicle of velocity close to v will arrive at such time as to force a delay in the switch to red (it is within a legal stopping distance for speed v at time t_y). If the arrival rates q_i of vehicles are sufficiently close to (but less than) saturation, however, it may be more important to reduce the time lost in switching the signal than to hold the signal until the platoon has completely passed. If one switches the signal while the platoon is still crossing the intersection, the signal than to hold the signal until the platoon has completely passed. If one switches the signal while the platoon is still crossing the intersection, the velocity of the vehicles would be approximately v_p and one could presumably set the yellow interval at approximately $v_p/2a$ based on a legal stopping distance for the velocity v_p rather than v. One would still typically want to allow most or all of the platoon to clear the intersection, but perhaps the criteria should be that the last vehicle to pass is traveling at velocity v_p rather than that the last vehicle of the platoon is certain to clear the intersection.

If one wishes to switch the signal to yellow earlier than in figure 2.9, one must detect the end of the platoon earlier, which means that the detectors must be still further away from the intersection. Suppose, as in figure 2.10, we place the detectors either at the distance (2.5.2) as in figure 2.8 or at the legal stopping distance for velocity v as in figure 2.9, whichever is further, and we switch the signal to yellow as soon as we detect the last platoon vehicle to pass the detector at times $t_0 + \beta$ of figure 2.10. Now, as in figure 2.8, the first vehicle to pass the detector after the platoon passes, is legally required to stop (vehicle 2 in figure 2.10), even though this vehicle might have been able to overtake the platoon before it reached the intersection and might, therefore, have been able to maintain a flow of approximately s for a little longer.

If the detector is further from the intersection than (2.5.2), then, at time t_y, vehicle 1 will certainly be further from the intersection than the minimum stopping distance for v_p. Depending on the various parameters v, v_p, β, etc., the vehicle might also be further away than the <u>legal</u> stopping distances for v_p, in which case one can expect this vehicle to stop also. Any vehicles located at time t_y between the minimum and legal stopping distances would have an option of stopping or continuing, but in any case the maximum yellow interval $t_r' - t_a$ can be chosen as the yellow interval for velocity v_n.

mum yellow interval $t_r' - t_a$ can be chosen as the yellow interval for

velocity v_p.

Information from the single detector located as in figure 2.10 is not sufficient to predict accurately when the last continuing vehicle will pass the intersection. It cannot even predict which vehicles will continue. A second detector located close to the intersection could provide more accurate information. As described previously, the purpose of the near detector would be to switch the signal to red as soon as possible after the last continuing vehicle passes the intersection, and thus reduce the yellow interval even more.

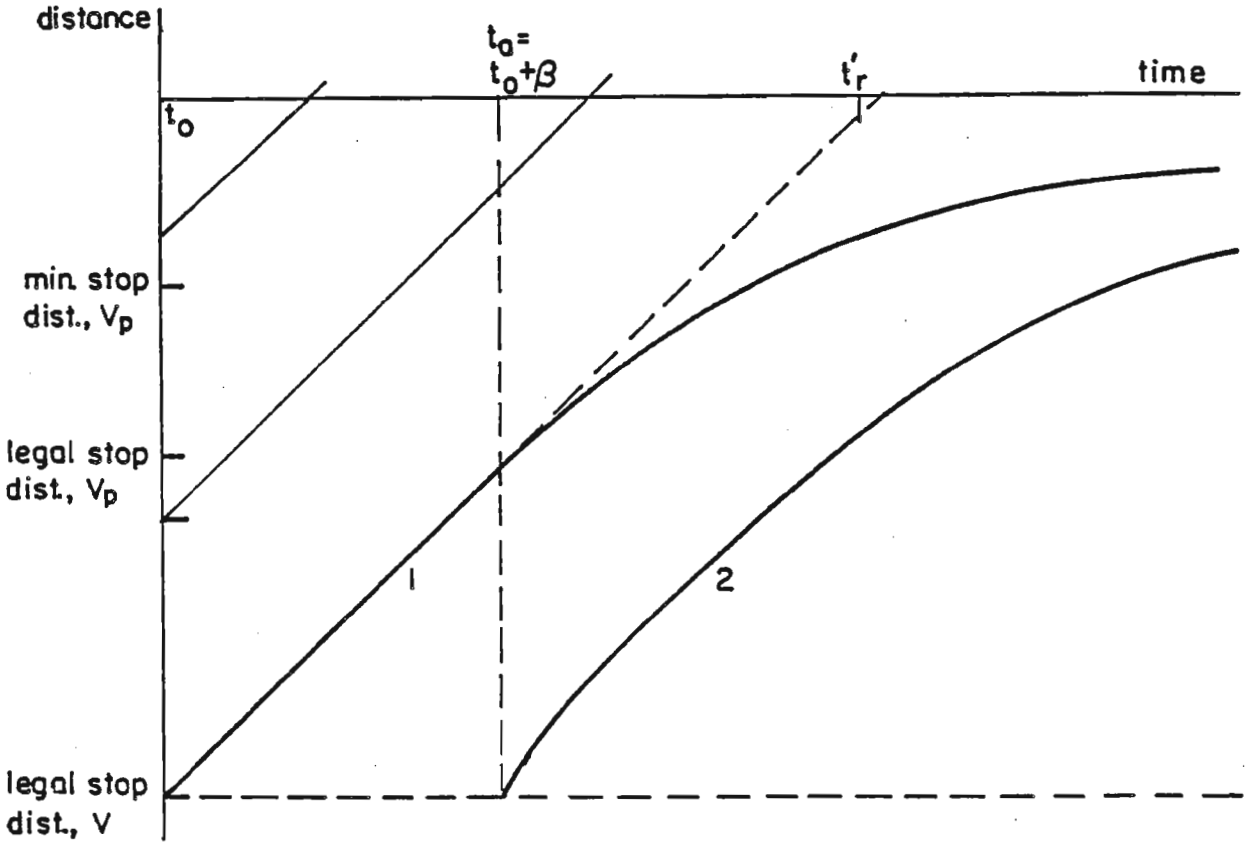


Fig. 2.10 - Vehicle trajectories near a signal at the termination of green, detector at the legal stop distance.

Fig. 2.10 - Vehicle trajectories near a signal at the termination of green, detector at the legal stop distance.

So far, the discussion has been limited to strategies for reducing the lost time (for a single lane of traffic), when it is known that the queue overruns the detector during the red time. With an impulse detector as far away from the intersection as proposed here, however, the queue for an efficiently designed V-A signal would likely overrun the detector only if the system is quite close to saturation (there would typically be room for 6 or 7 stopped vehicles between an impulse detector or the outer edge of a loop detector and the intersection.

Some of the earlier types of V-A controllers were designed to give a certain minimum green G_m and then to extend the minimum green only as long as vehicles kept passing an impulse detector with a headway less than β, or as long as a loop detector was occupied (provided the green interval did not exceed some preset maximum G_M). With an impulse detector and no mechanism for counting vehicles between the detector and the intersection there is a serious problem. If the queue did overrun the detector, one would need to delay looking for a minimum gap in the traffic until after at least a couple of vehicles had crossed the detector following the start of green, so that the controller would not falsely interpret a long start-up headway as the end of the queue. But having chosen a G_m large enough to provide this protection, one is forced to use the same G_m even if there is a queue of only 1 or 2 vehicles.

With a loop detector one does not have as serious a problem because the

to use the same G_m even if there is a queue of only 1 or 2 vehicles.

With a loop detector one does not have as serious a problem because the loop would stay occupied as long as any queued vehicles were on the loop. The distance between the inner edge of the loop and the intersection would store only a couple of vehicles at most. One would need to choose a G_m only large enough to serve one or two vehicles (which one would likely do anyway). Actually, the usual practice was to place the inner edge of the loop detector near or at the stop line, but this virtually guarantees that no vehicles pass the intersection during the early part of the yellow interval. The impulse detectors

were also typically located too close to the intersection to be effective in limiting the lost time in switching.

Current designs of controllers (NEMA controllers) are capable of counting the number of vehicles which pass an impulse detector during the yellow plus red intervals and are programmed to provide at least enough green time ("initial green") for these vehicles to pass the intersection. (This initital green may also have a preset maximum which presumably should be set equal to the time needed to clear a queue that extends from the intersection to the detector. This maximum initial green, however, is redundant since the counter would count at most only as many vehicles as can be stored between the intersections and the detector). These controllers were not specifically designed to do what is proposed here; namely, to terminate the green interval about two seconds before the end of the platoon reaches the intersection, but (with detectors in appropriate locations) they could (possibly) be set to do approximately this.

If no new vehicles pass the detector after the start of the green, the green will terminate after the initial green. Whether one provides a bare minimum or an ample minimum initial green time to serve the vehicles which arrived during the yellow plus red intervals and are known to be in the queue at the start of green is not important, because in this situation the traffic must be so light that one would not care. If, however, there are several vehicles in the queue at the start of the green, but the queue has not overrun the detector, it is almost certain that one or more vehicles will pass the detector during the initial green. Present NEMA controllers would not add these vehicles to the count and extend the initial green, but neither would they terminate the initial green early if these vehicles passed the detector with a headway larger than β. The controller will, however, terminate the green when a headway larger than β occurs after the initial green.

There are two problems that may arise with this scheme. First, if the

the initial green, the controller may not provide adequate time to serve these vehicles. The detector may find a gap immediately after the initial green and terminate the green before these vehicles can be served. Second, if the queue does overrun the detector (possibly after the green has already started), the initial green may terminate before a start-up wave can reach the detector and the flow across the detector increases to a value comparable with s. The detector may misinterpret a long start-up headway as the end of the platoon.

In addition, if either of the above happens and some vehicle fails to clear the intersection, the controller will not know this. In the next cycle, the controller counts only the new arrivals during the yellow plus red intervals. The overflow vehicles could, therefore, accumulate from cycle to cycle. The system could even become oversaturated unnecessarily.

There are ways to overcome these problems. If one uses only impulse detectors, the controller should count not just the vehicles which cross the detector during the yellow plus red. It should continue counting the vehicles which pass the detector after the green starts and keep extending the initial green until the controller infers that there is no vehicle which, if traveling at speed v from the detector to the intersection, could reach the intersection by approximately the end of the extended initial green interval. The controller could then start looking for a gap of duration at least β and terminate the green when one is found. Theoretically, this should eliminate both of the problems described above, but there is still the danger that, if the detector misses a vehicle or the the counter makes an error, some vehicles might fail to clear the intersection and the controller would never know that there is a residual queue.

The most common type of "impulse" detector is actually a small loop detector which is capable of recording the first passage time of a vehicle over

any part of the detector and also whether or not, at any time, some part of a vehicle is over the detector. If the loop is shorter than the length of a vehicle, each vehicle initiates a new pulse so the controller can count vehicles and also check for the presence of vehicles on the detector. The controller, instead of seeking a gap β between the initiation of pulses can look for a minimum time β' that the detector is unoccupied. The β' would be equivalent to the β less the typical time a single vehicle would occupy the detector (approximately v times the length of the vehicle plus the length of the loop). This type of detector gives some protection against having the green terminate before vehicles start moving over the detector, because the signal will not switch if a stopped vehicle is over the detector.

Some systems would also have a second (larger) loop detector close to the intersection. The signal will not terminate a green unless the second detector is unoccupied, and the first detector records a gap of β. This second detector would certainly eliminate the problem of leaving a residual queue at the intersection, but if the inner edge of the loop is too close to the intersection, it will cut off the flow during the yellow interval.

None of the above problems arise if one uses a single long loop detector (of length βv_p) or at least two small loop detectors over the same section of highway. With small loop detectors, the outermost detector could look for a time β' when the detector is unoccupied and also give some protection against terminating the green before vehicles have started to move. The inner loops would give further protection against terminating the green when there was a queue, and also make sure that the vehicle which initiated a minimum gap had actually moved out of the section (and not be stalled by a queue). Instead of a single long loop, one could also use a series of small loops with spacing less than the length of a vehicle so that any vehicle within the section of highway would activate some detector. With more than two small loop detectors, the

system would have some redundancy. The system could be designed so that it would operate well even if one of the detectors fails and also send out an alarm to the operator that something is wrong. With such a system, it is unnecessary to count vehicles which cross the detector during the red interval even though the inner edge of the detector system is spaced about a 2-second trip time from the intersection. One could also install another loop detector closer to the intersection to initiate a possible early switch from yellow to red.

We have considered, so far, only how one would switch a signal from green to yellow and from yellow to red if the signal controlled traffic in only one direction and only in a single lane. In practice, of course, a V-A signal typically would control more than one traffic lane per direction and/or two traffic directions simultaneously. We will postpone considering possible turn phases and imagine that we have only through traffic in directions 1 plus 3 or 2 plus 4.

There are some possible complications if the green interval needed by directions 1 and 3 (or 2 and 4) are nearly equal (we will discuss this in section 2.7) but, for now, let's assume that the queues will (almost always) empty in directions 3 and 4 before they empty in directions 1 and 2 respectively. In this case, we anticipated that one would like to terminate the green for direction 1 when the queue vanishes in direction 1 (again using as much of the yellow interval as possible and/or keeping the yellow interval as short as possible).

If one always provides a usual yellow interval for velocity v in direc-

interval as possible and/or keeping the yellow interval as short as possible).

If one always provides a usual yellow interval for velocity v in direction 3, one can switch the signal from green to yellow at any time, independent of where vehicles may be in direction 3 (as one would do for a F-C signal). In particular, one can terminate the green interval for direction 3 at the optimal time for terminating the green for direction 1. For this strategy one does not need vehicle detectors in direction 3 except that maybe, at another time of day, the role of directions 1 and 3 may be reversed.

This is indeed the proper strategy for terminating the green for direction 1 (except maybe for very light traffic). The complication here is that, if the queue vanishes in direction 3 before it vanishes in direction 1, and one switches the signal to yellow for directions 1 and 3 at time $t_y = t_0 + \beta$ as in figures 2.8 or 2.10, there may be a vehicle traveling in direction 3 at speed v which is located (barely) within the legal stopping distance at time t_y. One does not typically want to extend the green or yellow intervals for such a vehicle in direction 3, but, by the time the detectors in direction 1 have called for a switch to yellow, it is already too late to do anything about the vehicle in direction 3. The yellow time must be extended to allow this vehicle safe passage.

One could try to switch the signal to yellow for direction 3 earlier than in direction 1 so as to stop any vehicle in direction 3 which would cause an extension of the yellow interval beyond that needed for direction 1. To do this, however, one must anticipate when one wishes to switch the signal to yellow in direction 1 in time to switch the signal in direction 3. This would require that one either place the gap detector in direction 1 even further from the intersection, or have another detector further upstream. This is probably not practical for most intersections.

Suppose, however, one had some (presence) detectors in direction 1 to initiate the termination of green and also a loop detector near the intersection which could detect the absence of vehicles near the intersection during the yellow, and one had similar equipment in direction 3. When the outer detectors in direction 1 become unoccupied and the controller switches the signal to yellow for both directions 1 and 3, the outer detectors in direction 3 could observe whether or not there is a vehicle in its section which might require an extension of the yellow interval.

If the outer detectors extend out to a legal stopping distance, one is

reasonably certain that no vehicle in direction 1 will need a full yellow interval, but, in any case, loop detectors near the intersection could be designed so as to terminate the yellow as soon as the last vehicle in either direction 1 or 3 had safely passed some point in the intersection (initiated by the absence of vehicles on the inner loops in both directions 1 and 3). If in doubt, one would provide, at most, the usual yellow interval.

If one has multilane approaches, in direction 1 for example, it would be reasonable to assume that the queue vanishes at nearly the same time in all lanes, because drivers would tend to join the lane with the shortest queue and perhaps even jockey between lanes if one queue is moving faster than the another.

If one had separate detector systems in each lane (preferably presence detectors), one might, in principle, determine the times for each lane when a vehicle is within a minimum stopping distance of the intersection and is followed by at least some minimum time or spatial gap, indicating that the last platoon vehicle in that lane is approaching the intersection. Soon after one has observed such an event in one lane, one should presumably observe it also in the other lanes. If the signal switches to yellow when the first such event occurs in any lane, the signal might cut off one or two vehicles from platoons in other lanes. Newly arriving vehicles in the next cycle, however, should distribute themselves among the lanes so as (nearly) to equalize the queues in other lanes.~~"~~ Newly arriving vehicles in the next cycle, however, should distribute themselves among the lanes so as (nearly) to equalize the queues in all lanes. There is no danger that a residual queue in some lane will accumulate from cycle to cycle (if the detectors are presence detectors).

For nearly saturated conditions, the optimal strategy would be to terminate the green as soon as one expects a drop in the combined flow of all lanes, i.e., when a platoon first passes in any lane. If, however, the end of the platoon passes at nearly the same time in all lanes, this is not of critical importance. Indeed, if there is a possibility of "accidental gaps," it may be advantageous to continue the green until the detectors have observed a gap in a second lane

(not a simultaneous gap in two lanes). On the other hand, the system should operate quite efficiently if there were detectors in only one lane (probably the outer lane).

Most V-A signals on multilane approaches treat the multilane traffic stream as if it were just a single traffic stream. An impulse detector stretched over all lanes records the passage of any vehicle in any lane. A "headway" would then be observed as a time between the passage of a vehicle in some lane and the next vehicle in any lane. Such a system would have all the problems described previously for impulse detectors on a single lane regarding vehicles being caught between the detector and the intersection if the queue does not overrun the detector. Some V-A systems have long loop detectors in two or more lanes (particularly for a double left turn lane), but they typically behave as if the multiple detectors were a single detector covering all lanes. The signal will stay green as long as there is a vehicle on any loop and will switch only if the longitudinal spacing between vehicles in the superimposed traffic stream exceeds the length of the detector.

Whether one uses impulse detectors or loop detectors, to seek a time or spatial gap in the superposition of two (or more) traffic streams is not a very efficient mechanism for determining the end of a platoon. Suppose, for example, that there were two lanes,and vehicles in the queues pass the detector with

efficient mechanism for determining the end of a platoon. Suppose, for example, that there were two lanes,and vehicles in the queues pass the detector with

nearly equal headways of 2 seconds in each lane. Vehicles in the two lanes would not pass in any particular relative position, so the "headway" in the superimposed traffic stream could have any values between 0 and 2 seconds (uniformly distributed) with a mean of 1 second.

The ends of the platoon are likely to pass the detectors at nearly the same time in both lanes, after which the flow will drop to the value q, for example about $s/2$ with a mean "headway" at least of 2 seconds. If one does

not want to risk terminating the green interval while the queue is still passing, one would need to choose a minimum gap, β , of at least 2 seconds even if the headways in a single lane were exactly 2 seconds. If there is some variability in the latter, one would need to choose a β of at least 2-1/2 or 3 seconds. One is trying, however, to maintain a flow s (with a mean "headway" of 1 second) and to respond as quickly as possible when a platoon passes. If one must wait for a "headway" of 2-1/2 or 3 seconds in a stream with a mean "headway" of about 2 seconds, several vehicles will likely pass (with headways between about 1 and 2 seconds) before the controller recognizes that the flow has dropped.

Suppose, on the other hand, that one observed the vehicles in only one of the lanes. While the platoon is passing, the headways in a single lane would not likely be less than 1-1/2 seconds or more than 3 seconds. The mean headway after the platoon passes would probably be at least 4 seconds. If one observers a headway larger than 3-1/2 seconds, it would likely mean that the platoon has passed, but it is not likely that more than one or two vehicles will pass, after the platoon, before one finds a headway of at least 3-1/2 seconds.

Most traffic controllers are designed so that one can reduce the gap β after some specified time. If the objective of the controller is to switch the signal as soon as possible after the platoon passes, this "gap reduction" does not seem to serve any useful purpose.

2.6. Vehicle-actuated Signals - Delays for One-way Streets

not seem to serve any useful purpose.

2.6. Vehicle-actuated Signals - Delays for One-way Streets

We saw in the previous section that a V-A signal could be designed so as to reduce the effective lost time as compared with a F-C signals, but actually most V-A signals do not do this (they may even increase the lost time). The main difference between V-A and F-C signals, however, is the manner in which they respond to fluctuations in the number of vehicles which arrive from one cycle to the next.

If, for a F-C signal, more than an average number of vehicles should arrive during a red interval, the queue may not clear during the subsequent green interval. The overflow carries over into the next cycle which may then cause an overflow in the second, third, etc., cycles until there is some excess green time to serve the queue or there is a fluctuation in the opposite direction (fewer than the average arrivals in some cycle) and some extra time to handle the residual queue. If the system is undersaturated, the queue will eventually clear during some green interval. On the other hand, if one had less than the average number of arrivals during a cycle in which there was no residual queue, the signal would stay green after the queue vanishes even though vehicles may be waiting in the cross direction.

A V-A signal will typically terminate the green (approximately) when the queue vanishes. If more than the average number of vehicles should arrive during some red interval in direction 1, for example, the subsequent green interval would be extended to accommodate the extra vehicles. This would initiate a "chain reaction." The longer green would cause a longer queue to develop in the cross direction (direction 2). Presumably the green for direction 2 would then be extended to accommodate these vehicles which in turn would imply a longer concurrent red interval for direction 1, a longer subsequent green, etc. Thus, the cycle time will automatically increase as if the "arrival rates" q_i had concurrent red interval for direction 1; a longer subsequent green, etc. Thus, the cycle time will automatically increase as if the "arrival rates" q_i had increased temporarily. The cycle time is likely to stay at a higher value than the average until there is a fluctuation of the opposite sign. Correspondingly, a deficiency in arrivals during some cycle may cause the cycle time to drift to a value below the average.

As noted previously in section 2.4, one could allow the cycle time for the F-C signal to vary with time according to prevailing values of the $q_i(t)$ measured by averaging the cumulative arrivals over many days. It would not, however, respond to "fluctuations" in the arrivals on any particular day relative

to the average over many days. The main difference between the V-A and a F-C signal (with time-dependent cycle time) is that a F-C signal accommodates fluctuations in arrivals by keeping a predetermined cycle time but allowing a fluctuating overflow queue. The V-A signal accommodates fluctuations in arrivals by allowing fluctuations in the cycle time but a predetermined size of the overflow queue, namely zero. For an undersaturated V-A signal one, of course, does not need to know the $q_i(t)$ since the average cycle time will automatically adjust to the (average) $q_i(t)$. This gives the V-A signal a big practical advantage over a F-C signal which is restricted to only a few choices of the cycle time in any 24-hour period.

It seems clear, at least for the intersection of two one-way streets, that a V-A signal should give substantially less delay than a F-C signal. Certainly if one had some excess vehicles in direction 1, with $s_1 > s_2$, it would make no difference why one had this excess; if it were due to fluctuations from some average over many days or it were predictable. As long as the output flow from the signal had a value s_1, it would not be advantageous to suffer a lost time to switch the signal to direction 2 where the flow would be at most s_2, if the objective were to decrease the total queue in directions 1 and 2 at the maximum rate. One would not want to switch the signal just because the F-C green interval had expired. Also, if the signal were green in direction 2, the queue had vanished in direction 2, but there was a queue just because the F-C green interval had expired. Also, if the signal were green in direction 2, the queue had vanished in direction 2, but there was a queue in direction 1, nothing can be gained by extending the green in direction 2 and keeping the queue zero while the queue continues to grow in direction 1, just because the F-C green had not yet expired.

As observed in section 2.2, in the deterministic approximation, the optimal F-C signal would operate at its minimum allowed cycle time and consequently the F-C and V-A signals would behave in the same way, if they both had the same lost time L for switching. The difference between the two relates primarily

to the manner in which they respond to fluctuations. For the F-C signal the fluctuations cause stochastic queueing which gives delays proportional to the magnitude of the fluctuations (the stochastic queue in (2.3.10) is proportional to I). For the V-A signal there is an analogous term to be added to the deterministic approximation, also approximately proportional to I . Whereas in the deterministic approximation, the total delay could be described as the area between arrival and departure curves as illustrated in figure 2.1, the area of a set of equal size approximate triangles; for the V-A signal the total delay can be similarly described but the size of the triangles will be unequal. The extra delay due to fluctuations is associated with the fact that the average area of a collection of unequal size triangles is larger than for the same number of equal size triangles over the same total time period. Specifically if the cycle time C is considered as a random variable with expectation (average) E(C), the average waiting time with unequal cycle times would be larger than with equal cycle times by a factor of approximately

$$\frac{E(C^2)}{[E(C)]^2} = 1 + \frac{Var(C)}{[E(C)]^2} , \qquad (2.6.1)$$

in which Var(C) is the variance of C .

The literature on the theory of V-A signals and the related theory of "queues with alternating priorities" is very large. It begins in the early 1950's but much of it was published in the 1960's. Unfortunately very little "queues with alternating priorities" is very large. It begins in the early 1950's but much of it was published in the 1960's. Unfortunately very little of this literature is of direct practical value and much of it is actually "incorrect." The queueing theory literature deals mostly with mathematical techniques for analyzing "exactly" models which are too idealized to be very realistic. On the other hand, most of the literature on optimal control postulates an objective function including delays only over a "finite horizon" namely during the next one or two signal cycles. It fails to recognize that

the consequences of any action will persist for many cycles, particularly if the demand is close to saturation. Even the computer programs for "optimization" are based on questionable models and objectives and shed little light on the issues. Actually the issues are too complex to be described by any "comprehensive" theory which is simple enough to be useful. One can, however, describe some qualitative behavior that should serve to give some guidelines.

Tanner in 1953[13] was the first to apply queueing theory techniques to a model of a V-A signal. He postulated that vehicles arrived as Poisson processes of rates q_1 and q_2 in directions 1 and 2, and left at time intervals of $1/s_i$ during the green intervals. When a queue vanished, the signal could switch but there would be a predetermined lost time associated with each switch. He tried also to consider the fact that if there were no vehicles waiting in direction 2, for example, when the queue vanished in direction 1, the signal would be idle. It would presumably stay green for direction 1 at least until a vehicle arrives in direction 2.

Actually it is not clear what one wishes to do if the signal is temporarily idle. It may be advantageous for the signal to "home" in the direction with the large q_i. Even if the detectors are placed far enough away from the intersection so that a signal which is red will switch to green before a vehicle must start to decelerate, one would probably still prefer to indicate a green as early as possible to as many vehicles as possible.

must start to decelerate, one would probably still prefer to indicate a green as early as possible to as many vehicles as possible.

Another common strategy, however, is to have the signals "rest on red" simultaneously for all directions, so that it can switch to green for whichever traffic direction first actuates a detector. The advantage of this strategy is that there would be no need to insert a yellow interval for direction 1, for example, before the signal turns to green for direction 2 if a vehicle arrived in direction 2 while the signal was idle but resting on green for directions 1 and 3. For this strategy to be efficient, however, one definitely

should have the detectors at least a legal stopping distance from the intersection (for all directions). If,however, there is only one vehicle stopped or decelerating in direction 2 when the queue vanishes in direction 1, but another vehicle in direction 1 has already passed the detector, one may choose to continue the green in direction 1 on the grounds that the vehicle in direction 2 is certain to be stopped no matter what one does, but, at the expense of a slight delay to this vehicle, one may be able to avoid stopping a vehicle in direction 1.

What one does when a signal is idle or there are only two vehicles competing for the signal is not of great importance, and there is no well-defined objective. The signal would not have been installed primarily to control vehicles at flows sufficiently low that this would be an issue anyway. To include such effects in any theory causes a considerable increase in the complexity of any formulas for cycle times, delays, etc., because the formulas would include a term depending upon the probability that no vehicle arrives in direction 2, for example, during the time needed to serve the queue in direction 1. The latter time, however, is one of the "unknowns" in the equations. To evaluate even the average cycle time requires the solution of some very cumbersome equations which also contain many parameters (q_1, q_2, s_1, s_2, and L).

In 1964 Darroch, Newell, and Morris[14] managed to avoid the mathematical complication of the idle signal by assuming that the signal would switch to

In 1964 Darroch, Newell, and Morris[14] managed to avoid the mathematical complication of the idle signal by assuming that the signal would switch to

green for direction 2 after the queue vanishes in direction 1, regardless of whether or not there were any vehicles waiting in direction 2 and vice versa. Except for very light traffic, there would almost certainly be a vehicle waiting in one direction when the queue vanished in the other direction but, in the extreme situation with zero flows, the signal would switch back and forth with a cycle time equal to the total lost time per cycle for switching, L (whereas in Tanner's model the cycle time would be infinite because the signal would

never switch). There was no minimum green inteval G_m or maximum G_M in these models.

It was again assumed that vehicles arrived as Poisson processes of rates q_1 and q_2. During a green interval when a queue was discharging, the times between departures were interpreted as independent identically distributed random variables with means $1/s_i$ (any start-up delays were presumably absorbed in the lost times). After the "queue vanished," i.e., the number of departures during the green interval first become equal to the number of arrivals since the start of the previous red, the green interval in direction i was extended until there was a gap of at least β_i in the Poisson arrival stream. During this extended green interval, vehicles departed with zero delay; a queue would not reform no matter how short the headways in the Poisson arrival stream. The signal would now switch, but there would be a (possibly random) lost time L_i from the time the last vehicle passed in lane i until the start of the (effective) green for the other directions. These lost times were assumed to be "given" but independent of the β_1, β_2.

This model is mathematically "well-defined" but not necessarily a realistic representation of a real V-A signal. For this "fairly general" class of mathematical models, however, it was possible to obtain explicit "exact" formulas for the mean cycle time and the mean waiting time (also higher moments) as a function of the $q_1, q_2, s_1, s_2, \beta_1, \beta_2$ and L_1, L_2 (and higher moments of for the mean cycle time and the mean waiting time (also higher moments) as a function of the $q_1, q_2, s_1, s_2, \beta_1, \beta_2$ and L_1, L_2 (and higher moments of the departure headways and lost times where relevant). Needless to say, the formulas involving this many parameters were rather cumbersome. Although one could make some minor extensions of this class of models, it is about as complex a class as one could expect to evaluate "exactly."

Typically in modeling some complex physical system such as a V-A signal, the more accurately one formulates the model, the more approximations one must make in the mathematical analysis of it. In the present case, if one does not

want to make any mathematical approximations (except possibly numerical approximations of known accuracy in the final solution), then one may be forced to make rather crude models of the physical system. Of course, the ultimate goal is to use the model to make inferences about the real world for which the errors are the combined errors for both the model and the analysis. Exact formulas for idealized models may not be very useful for making direct numerical predictions about the real world, but they may be useful in identifying the relative dependence of the solution on the various parameters (which, with a large number of parameters, is difficult to do by numerical tabulation from more realistic models that can only be evaluated numerically) and for suggesting or illustrating the effects of various mathematical approximations that one might use in the analysis of more complex models.

The main weaknesses of the idealized model described above relate to its representation of the mechanism of switching. The L_1 and L_2 must be interpreted as some appropriate "effective" lost times because the model does not explicitly identify any start-up times or yellow times. In the model, the L_1 was simply described as the time from the last vehicle leaving in direction 1 during the green until the "start of green" for direction 2 (the first vehicle leaves at a headway time later). Obviously the "effective" value of $L_1 + L_2$ will depend on some of the more detailed issues of switching strategy discussed in the last section.

Another weakness relates to the assumption of Poisson arrivals. This may be reasonable for some purposes but it is not very realistic for the purpose of evaluating individual headways immediately after a platoon passes the intersection or for selecting values for the β_i. For a Poisson process the headways have an exponential distribution which admits arbitrarily short headways.

Obviously one should extend the green interval if one has a headway less than $1/s_i$. In this model it is possible to choose the β_i so that, with non-zero probability, there is a flow greater than s_i after the "queue vanishes." There is actually an "optimal" β_i for this model, but it is dictated mostly by this unrealistic property of the model. Indeed the optimal value of β_i is only slightly larger than $1/s_i$.

Despite these obvious weaknesses, the model does illustrate some important facts which we expect to be true also for more realistic models. For example, the average cycle time has a form

$$E(C) = \frac{L^*}{1 - q_1/s_1 - q_2/s_2} \tag{2.6.2}$$

in which L^* is some effective average lost time (depending on the β_i) but typically comparable with or perhaps less than that for a F-C signal. Thus, the mean cycle time is comparable with the minimum cycle time for a F-C signal, independent of the magnitude of the stochastic fluctuations in the arrivals and departures.

A formula of this type would be valid for essentially any type of V-A signal (with only two directions of traffic) which switches almost immediately whenever a queue vanishes (or before it vanishes). This follows simply from the fact that, over a long period of time, the signal will spend approximately whenever a queue vanishes (or before it vanishes). This follows simply from the fact that, over a long period of time, the signal will spend approximately a fraction q_i/s_i of its time serving vehicles in direction i but will spend all the rest of its time, a fraction $L^*/E(C)$, in switching. The individual cycle times will, of course, fluctuate around this average (actually with the magnitude of the fluctuation comparable with the mean). This is valid, however, only under the assumption that there is no minimum green interval, G_m, and no pedestrian constraints, which might cause the flow from the signal to drop from s_i to q_i if the queue vanished before the end of the green

and also the signal switches from direction 1 to 2, for example, even if there is no vehicle waiting in direction 2.

For the F-C signal we concluded that a cycle time of this magnitude would be unacceptable for moderately light traffic, say for $q_1/s_1 + q_2/s_2 < 1/2$, because it would lead to cycle times less than about $2L$, typically about 20 or 30 seconds with effective green intervals of maybe five seconds. At this level of flow the signal is serving only one or two vehicles per signal phase. Under these conditions the assumption in the theory that successive departure headways are nearly equal (except for a single start-up time absorbed in the L^*) is not very accurate, but the conclusion that one would serve only one or two vehicles per signal phase is still valid. Aside from pedestrian constraints, the objection to having a F-C signal operate at this short a cycle time is perhaps associated mostly with the fact that if the F-C signal had a green interval only long enough for two vehicles to leave but three were waiting, the third driver would be rather irritated if he was stopped even if it were virtually certain that he would be served in the next cycle and the queue would also clear in the next cycle. For a V-A signal there is no obvious reason why the signal should continue a green interval in direction 2, for example, after the queue has cleared if there are vehicles waiting in direction 1 (independent of how many vehicles had been served during the green interval in direction 2). The queue-responsive feature is reflected here in the behavior of the system for flows

of how many vehicles had been served during the green interval in direction 2).

We are mostly interested here in the behavior of the system for flows close to but below saturation. In this case (2.6.2) is quite accurate (provided one knows the L^*), and it demonstrates the sensitivity of the average cycle time (and thus also the waiting times) to L^*. The average cycle time can be reduced by the same percentage by which one can reduce the L^*. Note that one cannot increase the "capacity" of an intersection by reducing L^*, at least not in this model for which the saturation flows s_i are maintained no matter how long it takes to discharge the queue. The "capacity" is still

dictated by the condition that $q_1/s_1 + q_2/s_2 < 1$. Reducing L^* simply allows one to distribute the excess capacity into more lost times and thus reduce the mean cycle time.

Since the mean waiting time depends upon the second moments of C, the "exact" formula for the average waiting time from the above model was a very complex expression involving the variances of the departure headways, the times to find a gap, and L_1, L_2, plus all the first moment variables, q_i, s_i, $E\{L_i\}$, and β_i. For moderately light traffic any of these variables could have a significant effect on Var(C) and the mean waiting time but for nearly saturated conditions, with E(C) much larger than L^*, relatively few combinations of these variables are important.

The length of a red interval in direction 1 corresponds to the sum of the times in direction 2 needed to serve the queue, wait for a gap, and switch the signal. The variance of this red interval, given the previous history of the signal behavior, is essentially the sum of the variances of these three components. The variance of the time needed to serve the queue, however, is proportional to the mean number of vehicles served and, effectively, to the coefficient I of section 2.3, whereas the other two variance terms are independent of the number of vehicles in the queue. For sufficiently long mean cycle times, the first of these terms will dominate the others. Indeed the variance of the times needed to find a gap and the variances of L_1, L_2 contribute very little to the variance of C. The variance of C is, however, proportional to the square of the first moment, i.e., even for arbitrarily long mean cycle times, the magnitude of the fluctuations in C (as measured, for example, by the standard deviation) are comparable with E(C) itself. The rather formal mathematical tools (involving generating functions) used to evaluate the moments for this model, however, did not provide a very clear explanation of why this is so. Another deficiency of "exact solutions" is

that the techniques of solution are often so complex that it is difficult to understand qualitatively why the solution comes out the way it does.

A less formal but approximate approach to the theory of the V-A signal under nearly saturated conditions was formulated in 1969[15]. It was argued there that, if the average cycle time was large compared with L^*, statistical fluctuations in the number of arrivals or departures during a single cycle would typically be fractionally of order $[E(M)]^{-1/2}$ with M equal to the number of arrivals in a single cycle. This, in turn, would cause the cycle time to vary from one cycle to the next only by this same (small) fractional amount. An increase in the cycle time due to an excess of arrivals during one cycle, however, would cause an increase of comparable size to a large number of subsequent cycle times. If, before the effects of this fluctuation had decayed, one experienced another positive fluctuation in the number of arrivals during a subsequent cycle, this would cause the cycle time to increase still more. The net effect of fluctuations in the number of arrivals over many cycles is that the cycle time "drifts" with an amplitude proportional to the sum of the fractional fluctuations in arrivals over many cycles. Indeed the amplitude of the drift is comparable with the mean value.

The analogous phenomena for a F-C signal predicts that, if the signal is sufficiently close to saturation, the average residual queue at the start of red (the Q in (2.3.1) and (2.3.2)) will be large compared with the ex-

is sufficiently close to saturation, the average residual queue at the start of red (the Q in (2.3.1) and (2.3.2)) will be large compared with the ex-

pected change in the average queue from one cycle to the next. Thus, the residual queue tends to drift "slowly" in the sense that it changes fractionally very little from one cycle to the next, but over a large number of cycles it may drift from values close to zero to values of perhaps two or three times the mean.

It is this slow drift of the cycle time which causes the mean waiting time for the V-A signal to be appreciably larger than if there were no stochastic

effects. The main issue, however, is the comparison of delays for a V-A signal with those for a F-C signal (with an optimal choice of green intervals for the two directions). Even approximate formulas for the delays at a V-A or F-C signal are quite complex functions of the q_i, s_i and I_i, so detailed comparisons of all cases would be quite tedious. One can obtain some typical results, however, by comparing delays for a symmetric intersection with $q_1 = q_2 = q$, $s_1 = s_2 = s$, and $I_1 = I_2 = I$. In this case one can show that the average delay for a F-C signal, with the cycle time chosen to give a minimum delay, is larger than the average delay for a V-A signal by a factor of approximately

$$\frac{2\left[1 + (qL/2I)^{1/2}\right]^2}{1 + (qL^*/I)} \tag{2.6.3}$$

This is a decreasing function of L^* and we have argued that one should be able to control a V-A signal so as to make L^* less than L. Even for $L = L^*$, however, the factor (2.6.3) is (appreciably) larger than 1. For example, if we choose $I = 2$ and $qL = qL^* = 3$, this factor has a value of approximately 3. Typically the delay for an optimal F-C signal will be at least twice that for a V-A signal with $L^* = L$ (for nearly saturated flows on one-way streets, $q_3 = q_4 = 0$).

We saw from (2.6.2) that a reduction in L^* would decrease the mean cycle time by the same fraction by which one reduces the L^*. The fractional reduction in the mean delay is not quite as large since the delay is proportional to $qL^* + I$. This results from the fact that as one reduces L^* the distribution of the cycle times becomes more skewed.

The above theory is based on the hypotheses that there is no minimum green interval G_m ($G_m = 0$) and no maximum green interval G_M ($G_M = \infty$), and the arrival rates q_1 and q_2 are (exactly) independent of time. If one should choose a $G_m > 0$ but sufficiently small and a $G_M < \infty$ sufficiently large

interval would almost always be between G_m and G_M then, of course, the G_m , G_M would have a negligible effect on the behavior of the system.

One may be required to have a nonzero G_m to accommodate pedestrians. For certain types of detector systems one may also need to provide a minimum green interval because the system is not capable of estimating when the queue will vanish, particularly for short queues.

As regards the maximum G_M , it has been stated in some books (but primarily in relation to two-way streets) that the G_M should be chosen comparable with the optimal green interval for the F-C signal. For the intersection of one-way streets with symmetric flows, this is certainly incorrect. Theoretically, the "optimal" G_M in infinite for undersaturated stationary flows, if the objective is to minimize total delay. Since drivers do not like very long cycle times, however, one should choose the G_M as the maximum acceptable value (perhaps one minute).

For asymmetric flows with $s_1 > s_2$ a strategy which minimizes the total delay may be more complex. Certainly if $q_1/s_1 + q_2/s_2 < 1$, an optimal strategy would guarantee that neither traffic direction is oversaturated so as to cause either queue to increase systematically. One would certainly not terminate the green interval in direction 1 before the queue has (nearly) vanished and, if the system is close to saturation, one would not extend the terminate the green interval in direction 1 before the queue has (nearly) vanished and, if the system is close to saturation, one would not extend the green interval after the queue has vanished. In direction 2 one would not extend the green interval after the queue vanishes either, but one may choose to terminate it before the queue vanishes.

Suppose that one imposed a maximum green interval G_{M2} only in direction 2 with G_{M2} possibly dependent on the actual size of the queue at the start of green and on the parameters s_i, q_i, and L^*. If the stochastic behavior of the cycle time causes the cycle time to drift to a large enough value that the queue in direction 2 does not clear in the interval G_{M2} , the green

interval will terminate. The queue in direction 1 when the signal turns green for direction 1 will now be quite close to the value $q_1(L^* + G_{M2})$, and the time needed to serve this queue will be nearly determined by the G_{M2}, to within an error dependent only on the fluctuations in the arrivals and departures within a single cycle. Specifically, the green time G_1 in direction 1 will be approximately such that

$$s_1 G_1 = q_1(L^* + G_{M2} + G_1) ,$$

i.e.,

$$G_1 \cong (q_1/s_1)(L^* + G_{M2})/(1 - q_1/s_1) . \qquad (2.6.4)$$

Thus, by imposing a maximum green interval only in direction 2, one, in effect, also limits the green interval for direction 1. There is no reason to impose a maximum green interval also for direction 1 unless the G_1 of (2.6.4) is unacceptably large.

In principle, one has the option of choosing the G_{M2} dependent on the length of the queue in direction 2 but, unfortunately, the length of the queue is difficult to measure with conventional detector systems. Suppose, therefore, that one chose the G_{M2} dependent only on the q_i, s_i and L^* (i.e., on the time of day). If the queue in direction 2 fails to empty during the interval G_{M2} for several successive cycles, the signal would appear to behave (temporarily) almost like a F-C signal. The green interval for direction 2 would be exactly G_{M2} and the green for direction 1 would be approximately G_1 as in (2.6.4). The advantage over a F-C signal is that there is no stochastic queueing in direction 1. Indeed the delays in direction 1 are essentially as described in sections 2.1 and 2.2. The penalty for this is that the overflow queue in direction 2 must absorb not only the normal fluctuations in the number of arrivals and departures in its own direction but also additional fluctuations in the arrivals per cycle generated by the fluctuations in the green interval for direction 1.

The long time behavior of the above system is that it will operate like the usual V-A system with each queue clearing during each green interval as long as the cycle time stays shorter than approximately $G_1 + G_{M2} + L^*$. The cycle time tends to drift, but any time it reaches and tries to surpass this value, the drift will abruptly cease, and an overflow queue would form in direction 2. The cycle time would stay at this value until the overflow queue vanished and cycle time drifts to lower values.

There will be a finite optimal value of G_{M2} for $s_1 > s_2$ because, for any q_1 and q_2, there is a nonzero probability that the cycle time and the length of queue in direction 1 during the red interval will become so large for $G_{M2} = \infty$ that it is advantageous to sacrifice a lost time due to switching in order to take advantage of the higher flow s_1 in direction 1. The optimal G_{M2} will, however, depend on the q_i, s_i, and L^*.

A detailed theory associated with this strategy has not yet been developed, but this may be somewhat academic. It is certainly advantageous to use a strategy of this type, but one might not care to choose the G_{M2} so as to minimize the total delay for the prevailing values of the $q_i(t)$.

First of all, a theoretical "optimal" G_{M2} would likely be evaluated on the premise that the $q_i(t)$ vary so slowly with t that they can be treated as virtually constant, but the demand is sufficiently close to saturation that the premise that the $q_i(t)$ vary so slowly with t that they can be treated as virtually constant, but the demand is sufficiently close to saturation that the queue lengths can be treated as continuous (noninteger) variables. The implication of this is that if one imposes a suitable maximum green G_{M2} and the queue does not clear during some cycle, and possibly not during several subsequent cycles either, because the actual number of arrivals per cycle is (temporarily) higher than its average value, then eventually there will be a fluctuation in arrivals of the opposite direction (less than the average number) and the queue in direction 2 will vanish without any further control. One is willing to trade some excess delay in direction 2 in order to reduce the delays

in direction 1. Typically this will mean that one is trading longer delays to a few travelers in direction 2 in order to obtain a smaller benefit to each of a large number of vehicles in direction 1. One might not be willing to make such a trade even though it might reduce the total delay. Also, this strategy would cause additional stops to vehicles in direction 2 and to apply such a strategy one would need to know the $q_i(t)$. The "optimal" G_{M2} would vary throughout the day as the $q_i(t)$ change.

A more serious problem is that such a theory would probably not apply to real situations anyway because, by the time one would want to apply such a control when the signal was close to saturation, the $q_i(t)$ are likely to be changing too rapidly for an "equilibrium" theory to apply. If one should observe an increase in the actual number of arrivals during a few successive time intervals (five-minute intervals, for example), one has no way of knowing if this is due to "stochastic causes," i.e., accidental, or to an increasing $q_i(t)$ unless one has observed the traffic over many days so as to measure the $q_i(t)$. The strategies one should use in the two situations, however, are quite diffirent. In the former case one would be more willing to let a queue form in direction 2 by restricting the G_{M2} on the grounds that the queue would probably not last very long; that the number of arrivals in later cycles would likely be lower and the queue would dissipate. In the latter case, however, probably not last very long; that the number of arrivals in later cycles would likely be lower and the queue would dissipate. In the latter case, however, one may anticipate that the signal is or will become oversaturated and that any residual queue would persist for the remainder of the period of oversaturation and be added to what would otherwise be there.

If a V-A signal becomes oversaturated, information from detectors may be useful as a mechanism for minimizing the switching time L^* but, as discussed in section 2.4, any decision as to how one should split the cycle time between directions 1 and 2 would depend on the queue lengths (which the detectors cannot observe) and is nearly independent of any stochastic properties of the arrivals.

Also, one is likely to use some criteria other than minimizing total delays and stops.

Since the objective is rather ill-defined, one might consider some simpler types of strategies which are easier to implement. If, during the rush hour, $q_1(t)/q_2(t)$ remains nearly constant as both vary with time, a reasonable strategy might be to select some fixed values of the maximum greens for both directions, G_{M1} and G_{M2}, with G_{M2}/G_{M1} somewhat less (maybe 10 or 20 percent) than the ratio $(q_2/s_2)/(q_1/s_1)$, and with $G_{M2} + G_{M1}$ at some maximum acceptable value (2 minutes perhaps).

Prior to the rush hour such a signal would behave like a conventional V-A signal with the queues clearing in each direction during their respective green intervals. As the $q_i(t)$ increase during the rush hour, the average amount of green time needed in direction 2 may approach the value G_{M2}, i.e.,

$$\frac{(q_2(t)/s_2)L^*}{1 - q_1(t)/s_1 - q_2(t)/s_2} \cong G_{M2} . \qquad (2.6.5)$$

At first one may experience some stochastic queueing, but then the queue will grow systematically when the left hand side of (2.6.5) exceeds G_{M2}.

With the green interval for direction 2 at G_{M2}, the green interval for direction 1 will be approximately as in (2.6.4) but with q_1 replaced by $q_1(t)$. At first, when (2.6.5) is true, the G_1 in (2.6.4) will have a value of approximately

At first, when (2.6.5) is true, the G_1 in (2.6.4) will have a value of approximately

$$\frac{(q_1(t)/s_1)L^*}{1 - q_1(t)/s_1 - q_2(t)/s_2} ,$$

which is less than the proposed value of G_{M1}. The green interval $G_1(t)$ will, however, increase as $q_1(t)$ increases, thereby reducing the fraction of time allocated to direction 2 and further accelerating the growth of the queue in direction 2. The queue will now grow at an average rate of

$$q_2(t) - \frac{s_2 G_{M2}}{G_{M2} + G_1(t) + L^*} .$$

If there is no maximum green G_{M1}, $G_1(t)$ will increase to whatever value is needed to clear the queue each cycle (since presumably $q_1(t)/s_1 < 1$), but, if one imposes a maximum green, the signal will behave like a F-C signal when $G_1(t) > G_{M1}$.

As the $q_i(t)$ decrease toward the end of the rush hour, the residual queue in direction 1 will eventually vanish. The V-A control will now automatically terminate the green in direction 1 when the queue vanishes with a green interval again given approximately by (2.6.4). As the $G_1(t)$ decreases, the fraction of time allocated to direction 2 will increase and the residual queue in direction 2 will also decrease. The system will eventually return to its original behavior after both queues are gone.

The choice of G_{M2}/G_{M1} is somewhat arbitrary, as is the choice of $G_{M1} + G_{M2}$. If one wished to minimize the total delay one would never terminate a green for direction 1 if there were still a queue, i.e., one should choose $G_{M1} = \infty$. At the other extreme, one should not choose $G_{M2}/G_{M1} = (q_2/s_2)/q_1/s_1)$ (or larger) because this would make the waiting times essentially the same for the two directions, but the number of vehicles in queue would be larger in direction 1. Actually the waiting times in the two directions are quite sensitive to the split G_{M2}/G_{M1}. If one does not wish to penalize the drivers in direction 2 too severely, one would probably only drop the G_{M2}/G_{M1} "somewhat below" $(q_2/s_2)/(q_1/s_1)$.

direction 2 too severely, one would probably only drop the G_{M2}/G_{M1} "somewhat below" $(q_2/s_2)/(q_1/s_1)$.

2.7. Vehicle-actuated Signals - Two-way Streets

For an intersection with traffic arriving in four directions at rates q_1, q_2, q_3, and q_4 (but no turning traffic) any models which might be analyzed exactly would give formulas for delay that are so complicated as to be virtually useless. Simulation is also of questionable value because one could not possibly tabulate the dependence of the delay on all the relevant parameters. There are, however, some important issues that must be considered for

two-way streets which do not arise for one-way streets.

If $q_1/s_1 > q_3/s_3$, $q_2/s_2 > q_4/s_4$ and the intersection is undersaturated, the queues in directions 3 and 4 will usually vanish before those in directions 1 and 2, respectively. When a queue vanishes in direction 4 before that in direction 2, the combined output flow will drop from $s_2 + s_4$ to $s_2 + q_4$. To decrease the total queue in all directions at the maximum rate, it might be temporarily advantageous to switch the signal at this time to directions 1 and 3, if the queues in directions 1 and 3 are sufficiently large and $s_1 + s_3 > s_2 + q_4$. One cannot continue to follow such a strategy indefinitely, however, because it would cause the queue in direction 2 to grow from one cycle to the next, Sooner or later one must serve this queue. One cannot let it grow forever.

In the absence of stochastic effects, it seems clear that, with time-independent arrival rates q_i , an optimal steady-state strategy would follow some repetitive pattern. One can indeed verify what seems intuitively obvious that it is better to discharge the queues in directions 1 and 2 during every cycle than leave some residual queues every cycle or to clear the queues only every second or third cycle.

If we include stochastic effects but assume that q_3 and q_4 are "sufficiently small," then the strategy which minimizes the total delay should be (nearly) the same as that which minimizes the delays in only directions 1 and 2. Since a V-A signal controlled by the vehicles in directions 1 and 2 only gives considerably less delay in these directions than any F-C strategy, the latter would certainly not be the preferred strategy. Any small values of q_3 and q_4 should, at most, cause only a "small" deviation in the strategies for directions 1 and 2, but certainly nothing could be gained by leaving small residual queues in directions 1 or 2. The optimal strategy should be to allow

the queues to vanish in directions 1 and 2, independent of the q_3 and q_4.

As noted in Section 2.5, the presence of traffic in directions 3 and 4 may affect the choice of yellow intervals. If one tries to choose the yellow interval for directions 1 and 2 based on the platoon speed v_p, then one should switch the signal to yellow for directions 3 and 4 earlier than for directions 1 and 2 so as to provide directions 3 and 4 with a yellow interval for the approach speed v before the signals in directions 1 and 3 (or 2 and 4) switch to red simultaneously.

Traffic in directions 3 and 4 does reduce the relative efficiency of a V-A signal as compared with a F-C signal. If the queues in directions 3 and 4 empty every cycle, the delay per cycle in these directions will be nearly proportional to the square of the effective red intervals for these directions as described by the deterministic approximation (2.1.7). The green intervals and the cycle times which are determined by the traffic in directions 1 and 2 will drift, however, and the long-time average delay per vehicle in directions 3 and 4 will be proportional to the average of C^2. These delays will be larger than for a F-C signal with the same mean cycle time by a factor (2.6.1). Certainly the vehicles in directions 3 and 4 would prefer equal cycle times to a variable cycle time with the same mean driven by and for the benefit of the traffic in directions 1 and 2. Except in some very hypothetical situations in which, for example, s_3 and q_3 are much larger than s_1 and q_1 (even

traffic in directions 1 and 2. Except in some very hypothetical situations in which, for example, s_3 and q_3 are much larger than s_1 and q_1 (even

though $q_1/s_1 > q_3/s_3$) the desires of the travelers in direction 3 will not carry enough weight to affect the optimal strategy.

If q_3/s_3 and q_4/s_4 are sufficiently close to (but less than) q_1/s_1 and q_2/s_2, respectively, then occasionally the queues in directions 1 and/or 2 may vanish before those in directions 3 and 4. One could still use a strategy of switching the signal as soon as the queue vanishes in directions 1 and 2, but if the queue has not vanished also in directions 3 and 4, respectively,

there will be a residual queue in this direction. The queue may persist for many cycles, but it will not continue to grow indefinitely. Indeed the queues in directions 3 and 4 behave as if they had degrees of saturation or traffic intensities of $(q_3/s_3)/(q_1/s_1)$ and $(q_4/s_4)/(q_2/s_2)$, respectively. The average values of the residual queues for directions 3 and 4 should be nearly the same as for a F-C signal with the same mean cycle time, i.e., from (2.3.7).

$$Q_3 \cong \frac{I_3}{2(1-\rho_3)} \quad , \quad Q_4 \cong \frac{I_4}{2(1-\rho_4)} \, , \tag{2.7.1}$$

with I_3 and I_4 the sum of the I_A and I_D for directions 3 and 4, respectively, and

$$\rho_3 = (q_3/s_3)/(q_1/s_1) \quad , \quad \rho_4 = (q_4/s_4)/(q_2/s_2) \, . \tag{2.7.2}$$

These formulas are only rather crude estimates, but they serve to illustrate the main qualitative issue. These queues become very large if ρ_3 or ρ_4 are close to 1. The V-A signal controlled by directions 1 and 2 only has reduced the mean cycle time to the minimum necessary to serve the vehicles in these directions, but this may now be so short as to cause excessive stochastic queueing in directions 3 and/or 4.

The "optimal strategy" in such situations is not obvious, but certainly the total delay under any strategy will be quite sensitive to the ρ_3 and ρ_4 if they are close to 1. If the signal is green in directions 1 and 3, it is the total delay under any strategy will be quite sensitive to the ρ_3 and ρ_4 if they are close to 1. If the signal is green in directions 1 and 3, it is unlikely that one would want to switch the signal before either queue has vanished, while the flow is $s_1 + s_3$. One would lose time in switching only to achieve a temporary flow of $s_2 + s_4$. If $q_1/s_1 > q_3/s_3$, the queue in direction 3 will usually vanish before that in direction 1. If so, one would not typically find it advantageous to switch the signal when the queue vanishes in direction 3, despite the fact that the flow drops to $s_1 + q_3$. If one did switch, the residual queue in direction 1 would carry over to the next cycle.

But in the next cycle it is even less likely that the queue in direction 1 will vanish before that in direction 3. If one continues to switch the signal when the queue in direction 3 vanishes (or when the first of the two queues vanishes) the queue will keep growing in direction 1. At some time one must serve the queue while the flow is only $s_1 + q_3$ and one is not likely to find a better time to do it in the future than the present.

The difficult question is what one should do if the queue in direction 1 vanishes before that in direction 3, despite the fact that $q_1/s_1 > q_3/s_3$ (which would be half the time if $q_1/s_1 = q_3/s_3$) . If we consistently switched the signal when the queue vanishes in direction 1, whether or not the queue in direction 3 vanishes, the residual queue in direction 3 can eventually be served during some subsequent cycle while the queue in direction 1 is being served, but it may take many cycles before this happens. This is essentially the situation described by (2.7.1). This will not be satisfactory if ρ_3 is too close to 1, similarly for directions 2 and 4.

A "safe" strategy would be to hold the signal green until both queues have vanished. If the queue in direction 3 almost always vanishes before that in direction 1, these strategies would all be essentially equivalent. Suppose, however, that $q_1 = q_3$ and $s_1 = s_3$. If one waited for the last queue to vanish, then one would be serving the total queue at a rate $s_1 + q_3 = q_1 + s_3$ every cycle between the times when the first and last queue vanished. Relative to the deterministic approximation for which the two queues would vanish simultaneously, having a flow of only $q_1 + s_3$ rather than $s_1 + s_3$ is equivalent to adding an extra "lost time" per cycle of $(s_1 - q_1)/s_1$ times the average duration of this flow. The duration of the reduced flow, however, would be proportional to the standard deviation of the number of arriving vehicles per cycle, which in turn would be proportional to $[E(C)]^{1/2}$. If the flows were

sufficiently close to saturation, the average cycle time could be arbitrarily large and consequently this additional lost time could be much larger than the L* itself. Indeed with this strategy the mean cycle time increases very rapidly as the q_i approach the saturated condition. Each increase in the cycle time increases the effective lost time per cycle which causes the cycle time to become even larger.

For $q_1/s_1 > q_3/s_3$ the magnitude of the above effect is very sensitive to the relative values of q_1/s_1 and q_3/s_3 since it depends primarily on the time spent waiting for the queue to vanish in direction 3 after the queue has already vanished in direction 1; thus, it is more or less proportional also to the fraction of cycles during which this happens. For q_3/s_3 sufficiently close to q_1/s_1 (and/or q_4/s_4 sufficiently close to q_2/s_2) and for flows close to saturation, i.e., $q_1/s_1 + q_2/s_2$ close to 1, the strategy of extending the green until the latter of the two queues vanishes will give average delays even exceeding those of a properly set F-C signal.[16]

Any "optimal" strategy for control of a signal with four traffic directions is likely to be quite sensitive to the q_i/s_i (and any possible time-dependence of them) and be quite difficult to apply, since it would require accurate measurements of the $q_i(t)$. The practical question is whether or not one can find some type of control which is easy to implement and fairly accurate measurements of the $q_i(t)$. The practical question is whether or not one can find some type of control which is easy to implement and fairly efficient under any conditions.

If the signal is serving only about five or less vehicles per cycle in any direction, one would probably prefer to extend each green until both queues had vanished, even though this may not minimize the total delay. To reduce the time lost in switching, one may also wish to switch the signal to yellow a bit earlier for that traffic direction in which the queue vanishes first. This would also be a satisfactory strategy at higher flows if the

queues in directions 1 and 2 almost always vanish before those in directions 3 and 4 respectively (but the cycle time does not exceed some maximum acceptable value). The problem is to find some suitable modification of this strategy which will prevent excessive delays when fluctuations in the arrivals cause the queues in direction 3 and/or 4 to last longer than those in directions 1 and/or 2 during a significant fraction of cycles.

Webster and Cobbe[10] proposed that one select appropriate maximum green intervals G_{M1} and G_{M2}, in particular, that one choose them as the optimal values for the F-C signal. They did some simulations with $q_1/s_1 = q_3/s_3$ and $q_2/s_2 = q_4/s_4$ and verified that these were nearly the optimal values of the G_{M1} and G_{M2}. This was done, however, before anyone realized that the delays were much more sensitive to the relative values of q_1/s_1 and q_3/s_3 (or q_2/s_2 and q_4/s_4) than to the relative values of q_1/s_1 and q_2/s_2. This strategy at least guarantees that the delays for the V-A signal are no larger than those of the optimal F-C signal. But for such choices of the q_i/s_i and nearly saturated flows, the delays would be nearly the same for the two strategies because there would be residual queues in all directions during most cycles and the green intervals would usually run to their maximum values.

If $q_1/s_1 > q_3/s_3$ and $q_2/s_2 > q_4/s_4$ the optimal choices of the G_{M1} and G_{M2} are clearly not the values for the optimal F-C signal. In particular, for small values of ρ_3 and ρ_4 the "optimal" values of G_{M1}, G_{M2} are infinite, subject only to restrictions on the maximum acceptable cycle times. For ρ_3 and ρ_4 close to 1, however, a strict adherence to a maximum green interval will often cause the green to terminate before <u>either</u> of the two queues vanish. It is clearly not advantageous to switch the signal when the output flow is $s_1 + s_3$ or $s_2 + s_4$.

A reasonable alternative to the above class of strategies would be to

keep the signal green until both queues vanish, provided that the resulting green intervals do not exceed some specified values G^*_{M1} and G^*_{M2} (not necessarily the values for the optimal F-C signal). If, however, the green interval reaches G^*_{M1} or G^*_{M2}, it will terminate only if the queue has already vanished in direction 1 or 2 (but not in directions 3 or 4), otherwise the green will be extended until the queue does vanish in directions 1 or 2 or the green interval exceeds some maximum acceptable value. One could expect that the best choices of the G^*_{M1} and G^*_{M2} are at least comparable with those for the optimal F-C signal.

If, with this control strategy, the fluctuations in arrivals caused the cycle time to drift to values below its average value primarily because of a deficiency in arrivals for directions 1 and 2, the signal would still allow the queues in directions 3 and 4 to vanish, thereby preventing large queues to form in these directions just because the traffic in directions 1 and 2 needed less time. If the fluctuation causes the cycle time to drift to values well above the average, however, it would most likely be due to an excess of arrivals in directions 1 and 2 (for ρ_3 and $\rho_4 < 1$) and it is also likely that the queues in directions 3 and 4 would vanish before those in directions 1 and 2. In this case, it is typically advantageous to extend the green until the queues vanish. Otherwise, the residual queues in directions 1 and 2 might persist In this case, it is typically advantageous to extend the green until the queues vanish. Otherwise, the residual queues in directions 1 and 2 might persist for many cycles. But if the cycle time is long, still due most likely to an excess of traffic in directions 1 and 2 over several previous cycles, it is still possible (particularly for ρ_3 and ρ_4 close to 1) that during some cycles the queues in directions 1 and 2 will vanish before those in directions 3 and 4. Now one must make a choice. If one extends the green for the vehicles in direction 3 or 4, this will, in effect, increase the effective lost time. Vehicles will be arriving in the cross directions during this time which initiates a chain reaction. The next green interval for the cross section will

increase, then the subsequent green for the original direction, etc. If one does not extend the green, there will be a residual queue in directions 3 or 4, which might carry over for several subsequent cycles. With ρ_3 and $\rho_4 < 1$ and also a cycle time larger than the average, there will, however, be an average excess of green time in subsequent cycles for direction 3 and 4. The residual queue will dissipate faster than if they were in directions 1 and/or 2. The proposed strategy would select the second option whenever the green intervals were larger than some G^*_{M1} or G^*_{M2}.

If there is a possibility that the intersection will become oversaturated, we have the same issues as discussed in section 2.4 and 2.6. Presumably the control system will also have a second set of maximum green times G_{M1} and G_{M2} besides the G^*_{M1} and G^*_{M2}. The first set would be absolute maxima which would not be exceeded even if the queues have not vanished. One still has the option of choosing $G_{M1} + G_{M2} + L^*$ as the maximum acceptable cycle time and the G_{M1}/G_{M2} so as to give preferential treatment to one direction or the other. One would tend to favor the direction with the larger value of s_1 or s_2, or with the larger value of $s_1 + s_3$ or $s_2 + s_4$.

2.8. Semi-actuated Signals

Most of the theory described in the previous sections is based on the assumption that the flows q_i (or at least q_1 and q_2) are of comparable

Most of the theory described in the previous sections is based on the assumption that the flows q_i (or at least q_1 and q_2) are of comparable magnitude. Whenever we applied a continuous approximation for one traffic direction, we did so also for the other direction or if the signal served only one or two vehicles per cycle in one direction, we implied that the same was true also for the other direction. Sometimes, however, a signal is installed at the intersection of a major road with a minor road particularly if the traffic on the major road is so heavy that it is difficult or dangerous for a vehicle on the minor road to cross. The traffic on the minor road may be so light that vehicles on the minor road must be served only one or two at a

In such cases the signal should stay green for the major road except for occasional interruptions to serve vehicles on the minor road. There is no need to have vehicle detectors on the major road because one is not likely to switch the signal back to the minor road as soon as the queue on the major road, generated by a previous interruption, vanishes. A detector could be used to observe when the queue vanishes, but that is not usually relevant. It could look for gaps in the traffic stream in an attempt to reduce the yellow time, but if the major road has two-way traffic and possibly multiple lanes in each direction, one is not likely to find a gap large enough to be of much use. The vehicles are likely to be traveling at the normal approach speed, and so one has no option but to provide a usual yellow time for that speed whenever the traffic is interrupted.

One will need detectors on the minor road because the signal should not switch unless a vehicle arrives on the minor road. A signal with detectors on only the minor roads is called a semi-actuated signal.

If one were concerned with minimizing stops and delays to the minor road vehicles, one could, theoretically, place the detector far enough away from the intersection that a single vehicle crossing the detector could cause the main street signal to switch to yellow and then to red by the time the minor road vehicle reached the intersection without stopping. One would not, however, street signal to switch to yellow and then to red by the time the minor road vehicle reached the intersection without stopping. One would not, however, wish to encourage minor road vehicles to cross the intersection at full speed just as the signal turns green. The vehicles should stop and be prepared to yield to any main road vehicles which had entered the intersection before the signal turned green for the minor road. Detectors are usually placed rather close to the intersection. The vehicle on the minor road will be stopped anyway and one is not too concerned about the possibility that one might save the vehicle from being delayed a full yellow time if it could be detected one or two

seconds before it reached the stop line. It is advantageous, however, to have the detector at least 50 feet or so from the stop line so that it can detect the possible arrival of a second or third vehicle at the intersection when a first vehicle is already stopped.

The issues regarding semi-actuated signals are fairly straightforward. One certainly should not interrupt the main street traffic while it is still discharging a queue from a previous interruption. Presumably the "cost" of an interruption per unit time is quite large compared with the cost per unit time of delaying a minor street vehicle, but the cost of two overlapping interruptions is appreciably larger than non-overlapping interruptions. Also, if these interruptions occur too frequently, there is a risk that the major road will become oversaturated, i.e., the queue will never disappear.

If a vehicle should arrive on the minor road when there is no queue on the major road, one can either serve this vehicle as soon as possible, or wait. Presumably the cost of interrupting the major road is (nearly) independent of how long it has been since the end of the previous interruption. The only benefit one could realize from postponing service to the minor road is that a second vehicle might arrive on the minor road and one could serve the two vehicles nearly simultaneously instead of serving them with separate interruptions. If the flow q_2 on the minor road is sufficiently low (say an average of one vehicle in ten minutes), however, one could not reasonably expect a vehicle to wait an average time of $1/q_2$ until another vehicle arrives.

If we assume that the price $p(w)$ for a unit of delay to a vehicle which has already waited a time w is independent of w and seek to minimize the total cost to all vehicles, we will be led to some unacceptable conclusions, because we will be trading possibly long delays to a single vehicle against short delays to a large number of vehicles. Suppose, therefore, that the $p(w)$ is an increasing function of w.

If we interrupt the major road for an effective red interval R, the number of vehicles stopped (or delayed) is approximately

$$q_1 R/(1 - q_1/s_1)$$

and the total delay to all vehicles will be approximately

$$q_1 R^2/2(1 - q_1/s_1) .$$

The value of R , which includes the effective lost time, is typically (at least) about 15 seconds. If we assume that $p(w) \cong p(0)$ is nearly constant for w less than R , then the total cost of the interruption will be approximately

$$P = \frac{p(0)q_1 R^2}{2(1 - q_1/s_1)} + \frac{\alpha p(0) q_1 R}{1 - q_1/s_1} , \qquad (2.8.1)$$

with α , as in (2.2.10), equal to the time equivalent of one stop.

The two terms of (2.8.1) are likely to be of comparable size (with R ~ 15 sec. and α ~ 10 sec). For q_1 = 1200 vehicles/hr, $1/q_1$ ~ 3 seconds, and q_1/s_1 small, P/p(0) would be about 3 vehicle-minutes, i.e., the cost is equivalent to a total of about 3 minutes of short delays.

If there is one vehicle on the minor road and it has already been delayed a time w , the cost (by definition) of delaying this vehicle an additional time dw is p(w)dw . The benefit from waiting is that a second vehicle might arrive on the minor road during the time dw (with probability $q_2 dw$) . If we use a strategy in which we will certainly switch the signal if a second vehicle arrives and assume that the cost of an interruption for two vehicles is nearly the same as for one vehicle, then, by waiting for a second vehicle, we would save a cost P to the major street as compared with interrupting the major road for each vehicle individually. The expected benefit from waiting a time dw is, therefore, $Pq_2 dw$. The net benefit is $[Pq_2 - p(w)]dw$, so it is advantageous to postpone serving a minor road vehicle if

$$\frac{p(w)}{p(0)} < \frac{P}{p(0)}\, q_2 \,. \qquad (2.8.2)$$

This formula may not be quite correct because, if one served a single vehicle, one would not be willing to serve a second vehicle if it arrived during the interruption caused by the first vehicle. This is unlikely to happen, however, for the "small" values of q_2 of interest here. Also, the strategy of switching if there are two vehicles waiting on the minor road may not be optimal. But if q_2 is large enough that one would likely wait until three or more vehicles had arrived, one would probably install a fully actuated signal. This formula is accurate enough, however, to illustrate some possible issues.

Since $p(w)/p(0) \geq 1$ and increasing in w, (2.8.2) implies first, that if $1/q_2 > P/p(0)$, then one should not wait at all. For example, with $P/p(0) =$ three minutes, one should serve each arriving vehicle as soon as possible if the mean headway $1/q_2$ is larger than three minutes. If $p(w)$ were independent of w, i.e., $p(w)/p(0) = 1$, and $1/q_2 < P/p(0)$, (2.8.2) would imply, conversely, that one should make the minor road vehicle wait (at least) until a second vehicle arrives, possibly an average of as much as three minutes.

Actually we do not have any realistic data on the function $p(w)$ but the conclusion from (2.8.2) is that, for $1/q_2 < P/p(0)$, one should delay serving a minor road vehicle in the hope that a second vehicle will arrive. One will conclusion from (2.8.2) is that, for $1/q_2 < P/p(0)$, one should delay serving a minor road vehicle in the hope that a second vehicle will arrive. One will not wait indefinitely, however. As w increases, so does $p(w)$ and for some w_0 there will be an equality in (2.8.2). If a second vehicle has not arrived within a time w_0, one will serve the vehicle alone. The fact that one would not typically be willing to delay a vehicle more than one or two minutes implies that the $p(w)/p(0)$ must be rather large for such values of w. On the other hand, if there is a reasonable chance that a second vehicle will arrive within some tolerable waiting time, the vehicle should wait some tolerable time.

The above theory suggests that a semi-actuated signal may be appropriate if the traffic on the major road is sufficiently heavy that it is difficult for a minor road vehicle to cross, but the flow on the minor road is such that an average of less than about one vehicle will arrive during an interruption time, i.e., $1/q_2 \geq R$ (typically $q_2 \leq 250$ vehicles per hour). If $R \leq 1/q_2 \leq P/p_0$ (typically $20 \leq q_2 \leq 250$ vehicles per hour), one should interrupt the major road only after two (or more) vehicles have arrived on the minor road or a single vehicle has waited for some tolerable time (maybe at least 30 seconds). In the higher part of this range of q_2 one might even insist that three or more vehicles arrive before interrupting the major road. For q_2 below this range, however, one should serve the minor road vehicles one at a time, as soon as possible (but not within some minimum time of a previous interruption).

Any rule here is somewhat arbitrary because it necessarily involves trading longer delays to a few vehicles against shorter delays to many. Traffic engineers generally seem to be more generous in their treatment of minor road traffic than suggested here. Perhaps they are too generous.

For higher values of q_2, $1/q_2 < R$, one is likely to use a fully actuated signal (or a F-C signal), but if there is an average of less than about two vehicles arriving per cycle on the minor road (or in one direction of a two-way road), one may wish to modify the strategy described in the previous sections for the V-A signal. Just before the queue vanishes on the major road when the vehicle speed is v_p, one can use various strategies to minimize the lost time in switching the signal (such as using a yellow time for the speed v_p). If at this time, however, there is only one or two vehicles waiting on the minor road, one may choose to extend the green. By doing so, one will typically need to provide a longer yellow time (for speed v) which will lengthen the duration of the interruption. In this case one should consider extending the green until at least two or three (or more) vehicles are waiting on the minor road.

2.9. Right-of-way, priority rules, stop signs

Most traffic conflicts are resolved by legal rules of driver and pedestrian behavior which assign a right-of-way for every possible situation in which two people may wish to use the same facility at the same time. A traffic signal, in effect, alternates the right-of-way rules between different classes of people so that people respond mostly to the signal and to other people in the same class. They do not directly interact with people in other classes, except that, when a signal first turns green, the first drivers are expected to yield to anyone (of any class) already in the intersection, emergency vehicles, or any potential cause of an accident.

In the simple situations described so far, a two- or four-way intersection with no turning traffic, the traffic signal eliminated all potential conflicts. Generally, with possible turning movements and pedestrians, there will be conflicts, even with a traffic signal, and there will be rules to cover all such conflicts. There will also be rules for intersections with no signals and rules for changing lanes, merging, etc. The most basic rule, of course, is that a driver must travel at a safe distance behind another vehicle traveling in the same lane. In effect, he must "yield" to the vehicle ahead of him. In any accident someone is considered to be "at fault."

Many of the theoretical issues associated with turning movements are similar to those for any other type of right-of-way rules for conflicting traffic streams including those for unsignalized intersections. We will, therefore, consider next some general issues of priority service with particular emphasis first on an intersection with no signal and no turning traffic.

There is not much one can say about a four-way stop sign except that it is quite inefficient relative to the usual criteria (but it is inexpensive to install). One might use a four-way stop at an intersection where there are sight restrictions (curves, buildings, etc.) which make it difficult for a

driver, who would normally be expected to yield, from seeing when it would be safe to cross the intersection if vehicles traveled at their normal speed in the cross direction. A four-way stop might be used simply to eliminate the possibility of high speed collisions. Sometimes, in residential areas, it is used intentionally to hinder the flow of vehicles so that they will go elsewhere or in an attempt to reduce speeds.

In some cities of the world, aggressive drivers can make a four-way intersection operate almost like a roundabout with four cars in the intersection at nearly all times. They can achieve rather high flows, probably also quite a few minor accidents. In most cities, however, there would be at most only two cars in the intersection at a time alternating between directions 1 plus 3 and 2 plus 4 when there are queues waiting in all directions. In this case, the output flow is not very high. There will be one vehicle served in each direction in approximately the time it takes for a vehicle to cross the intersection in directions 1 or 3 plus the corresponding time in directions 2 or 4. These crossings, however, are at slow speed starting from a stopped position. The outputs in all directions will be nearly equal, independent of the size of the queues or the input flows q_i .

If the output flow per direction is s^* when there are queues in all directions but there is a steady arrival flow $q_i < s^*$ in direction i , then

If the output flow per direction is s^* when there are queues in all directions but there is a steady arrival flow $q_i < s^*$ in direction i , then there will be essentially no queue in that direction. If there is a queue in any direction, vehicles traveling in any other direction i where there is no queue will, in effect, have "priority" as long as q_i is less than s^* , in the sense that they will be delayed at most a time $1/s^*$.

If there is no queue in either directions 2 or 4 (i.e., $q_2, q_4 < s^*$) but there is a queue in direction 1 and/or 3, the excess capacity not used by directions 2 and 4 is not wasted. A vehicle in direction 1, for example, can be followed by another vehicle in direction 1 if there is no vehicle waiting

in direction 2 or 4 when the second vehicle in direction 1 moves up to the stop line. If, however, drivers strictly obey the law and come to a full stop before following a vehicle in the same lane, the headway between such vehicles is likely to be about four seconds (about twice that for vehicles which are not required to stop). In particular, if q_2 and q_4 are small compared with s^*, the output flow in direction 1 may be only about half the saturation flow s_1 without the stop sign. Of course, if there are no queues in any direction, each vehicle must come to a stop and may also be delayed even though there may be very few potential vehicle conflicts.

The most common method of resolving conflicts between two traffic streams is to give one traffic stream the right-of-way at all times. The second class of vehicles must then seek gaps in the primary stream.

There is an enormous literature dealing with methods of measuring the size of gap needed for various maneuvers, the time intervals between gaps of various sizes, the maximum rate at which secondary vehicles can be accommodated (the capacity) and the queueing of secondary vehicles, as a function of various characteristics of the primary stream. Attempts to develop some fairly comprehensive theory, however, have not been very successful, not so much because the mathematics is too difficult, but because the number of potentially relevant parameters needed to describe the physical situation is so large that no one

mathematics is too difficult; but because the number of potentially relevant parameters needed to describe the physical situation is so large that no one would want to measure all the parameters needed to apply the theory.

One can describe some simple issues qualitatively with simple models, provided one is willing to give up the possibility of using the results for precise numerical estimation. As a practical matter, one is not particularly concerned with how the capacity or queueing depends on the distribution of gap sizes that people want or if they will accept a shorter gap after they have waited a long time, etc., because one does not have much control over these things, anyway.

If one wants to know the capacity of some facility, one can measure it directly easier than one can measure all the relevant parameters. The main practical issue is whether or not one should introduce some other type of control, reverse the priorities, install a traffic signal, or use a multiphase signal.

One objection to giving the right-of-way to one traffic stream is that all the delays are imposed on the secondary stream. If one does not have a traffic signal, however, such is an unavoidable consequence of any simple rule which drivers can be expected to follow. A more serious problem is that if the primary stream has gaps large enough to accommodate secondary vehicles, it also will typically have many more gaps larger than the minimum safe headway between the primary vehicles but not large enough to be used by the secondary vehicles. If there is an insufficient number of gaps large enough to accommodate the secondary flow, one would like to compress some of these shorter headways without changing the primary flow (i.e., the mean headway) and, in effect, amalgamate several short excess headways into a headway large enough to accommodate a secondary vehicle. This is, in essence, what a traffic signal does. It holds up a group of vehicles, creates a large gap and then releases the vehicles at headways close to their minimum value. If the signal is undersaturated, it does this in such a way that the long time output flow is equal to the input flow. One might be able to induce a certain amount of "clustering", however, does this in such a way that the long time output flow is equal to the input flow. One might be able to induce a certain amount of "clustering", however, without a signal. For example, if passing is prohibited over some distance on the approach to an intersection, this will tend to generate some larger gaps ahead of slower cars (and shorter gaps ahead of drivers who would like to drive faster but cannot pass). Of course, a traffic signal at some neighboring intersection will also cause clusters.

Some of the issues can be formulated simply. Suppose the primary traffic is time-independent and that one could measure

$$F(h) = \text{fraction of headways less than } h \,. \tag{2.9.1}$$

This is evaluated by listing consecutive headways over some long time interval and giving "equal weight" to each headway in evaluating the fraction, i.e., it is the total number of headways less than h divided by the total number of all headways. If we smooth any graph of $F(h)$ vs. h, we can also define a probability density of the headways as

$$f(h) = dF(h)/dh \,. \tag{2.9.2}$$

One cannot specify both $F(h)$ and the flow q independently. If one observes a large number n of headways H_j in a total time

$$\sum_{j=1}^{n} H_j \,,$$

the flow will be

$$q = \frac{n}{\sum_{j=1}^{n} H_j} = \frac{1}{\frac{1}{n}\sum_{j=1}^{n} H_j} \cong \frac{1}{\int_0^\infty hf(h)dh} = \frac{1}{\int_0^\infty [1 - F(h)]dh} \tag{2.9.3}$$

the reciprocal of the mean headways.

The arrival rate of headways less than h is equal to the arrival rate of vehicles, q, times the fraction of these with headways less than h, i.e., $qF(h)$. If

$qF(h)$. If

$m(h)$ = average number of secondary vehicles that can be served in a headway h,

then the rate at which secondary vehicles can be accommodated (the capacity) is

$$\text{capacity} = q\int_0^\infty m(h)f(h)dh = \frac{\int_0^\infty m(h)f(h)dh}{\int_0^\infty hf(h)dh} \,. \tag{2.9.4}$$

This formula is deceptively simple. We have said nothing about probabilities

of events, behavior of drivers, or the properties of the traffic stream other than its headway distribution. The possibly complex dependence on driver behavior is absorbed in the (as yet unspecified) $m(h)$. If, for example, very few headways are large enough for two vehicles to use the same gap, then $m(h)$ is either 0 or 1 for most values of h. If some driver is timid and seeks a relatively large gap, he might let several gaps pass which another driver would have accepted. In this case, the $m(h)$, or equivalently, the average number of gaps of size h which are accepted, actually depends on the history of previous gap sizes. For example, if the previous headway was so large that even a timid driver would have accepted it, then one knows that the driver considering the current headway is not a timid driver carried over from the previous headway.

If we make the rather naive assumption that the $m(h)$ depends only on h, then the capacity (2.9.4) depends on the properties of the primary traffic only through the distribution $F(h)$, not on the joint distributions of successive headway. The formula would then apply even if the primary traffic had been manipulated by any type of control strategy, by a traffic signal, for example, which generated any well-defined distribution $F(h)$. One could, therefore, use (2.9.4) to illustrate the effect of any control strategy on the capacity, through its effect on the $F(h)$.

Suppose, for example, that all drivers were, in effect, identical. No driver would accept a headway less than some number h_0 but everyone would accept one greater than h_0. Furthermore, for headways larger than h_0, drivers would pass at intervals of time h_1 until the headway terminated, i.e., the function $m(h)$ is a step function as shown in figure 2.11.

$$m(h) = \begin{cases} j & \text{for } h_0 + (j-1)h_1 \leq h < h_0 + jh_1 \\ 0 & \text{for } h < h_0 \end{cases} . \qquad (2.9.5)$$

The capacity would then be

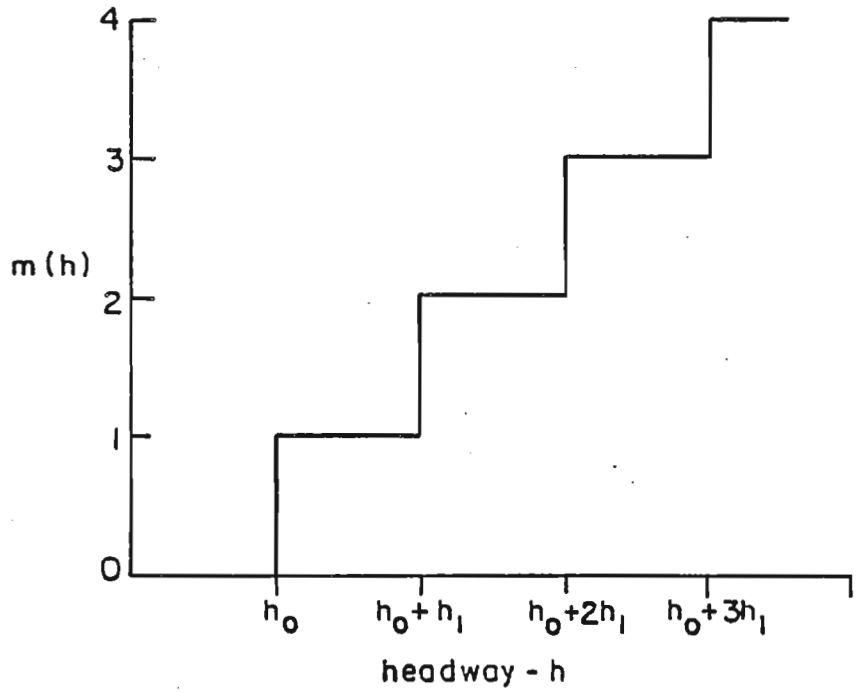


Fig. 2.11 - Number of vehicles which can cross in a headway of duration h.

$$\text{capacity} = q \sum_{j=1}^{\infty} j[F(h_0 + jh_1) - F(h_0 + (j-1)h_1] \tag{2.9.6}$$

$$= q \sum_{j=1}^{\infty} [1 - F(h_0 + (j-1)h_1)] \ .$$

For uncontrolled light (small values of q/s) primary traffic, one might postulate that traffic behaves like a Poisson process of rate q having a headway distribution

$$1 - F(h) = \exp(-qh) \ . \tag{2.9.7}$$

In this case

$$\text{capacity} = q \sum_{j=1}^{\infty} \exp(-qh_0 - q(j-1)h_1) = \frac{q\exp(-qh_0)}{1 - \exp(-qh_1)} \ . \tag{2.9.8}$$

Formulas of this type appear in many traffic engineering books to describe the capacity of two-way stop signs, yield signs, merge sections, etc., but they obviously are not very accurate. To use the formula, one would need to know some "effective" average values of the parameters h_0 and h_1 in (2.9.5). Nevertheless, the formula is useful to describe some qualitative causal relationships.

If in (2.9.8) we assume that q is "small" in the sense that $qh_0 \ll 1$

tionships.

If in (2.9.8) we assume that q is "small" in the sense that $qh_0 \ll 1$ and $qh_1 \ll 1$, we can approximate (2.9.8) by expanding $\exp(-qh_0)$ and $\exp(-qh_1)$ in powers of qh_0, qh_1 to obtain

$$\text{capacity} = \frac{(1 - qh_0 + \ldots)}{h(1 - qh_1/2 + \ldots)} \cong \frac{1}{h_1}[1 - q(h_0 - h_1/2 + \ldots] \tag{2.9.9}$$

At the other extreme if q is "large" so that qh_1 is larger than about 2 or 3, we can expand the denominator of (2.9.8) in powers of the small quantity

$\exp(-qh_1)$ to obtain

$$\text{capacity} \cong q\exp(-qh_0) + q\exp(-q(h_0+h_1)) + q\exp(-q(h_0+2h_1)) + \ldots \quad (2.9.10)$$

The interpretation and generalization of the approximations (2.9.9) and (2.9.10) to arbitrary $F(h)$ and/or $m(h)$ is quite straightforward. For sufficiently large gaps in the primary traffic (small q) and a large queue of secondary traffic, it is plausible to assume that the secondary traffic would establish a steady flow through each gap, after some initial "start-up" time. Indeed the first term of (2.9.9), $1/h_1$, represents this steady flow. For more general $m(h)$, the h_1 should be interpreted as the mean headway between secondary vehicles and $1/h_1$ as the mean steady flow through long gaps. This would not be the "saturation" flow as at a traffic signal. Presumably each secondary vehicle must stop or at least check to make sure that no primary vehicle is approaching the intersection. The h_1 is likely to be closer to the headway at a four-way stop sign (with no cross traffic), i.e., about 4 seconds.

The second term of (2.9.9) is proportional to q, and the $h_0 - h_1/2$ represents the effective loss in time available to the secondary stream caused by each interruption of the flow by a primary vehicle (if the primary flow q is so low that $qh_0 << 1$). Obviously the secondary flow cannot exceed $1/h_1$. For some more general model, the form of (2.9.9) would be the same but with so low that $qh_0 << 1$). Obviously the secondary flow cannot exceed $1/h_1$. For some more general model, the form of (2.9.9) would be the same but with $(h_0 - h_1/2)$ replaced by the appropriate average loss in time to the secondary traffic per interruption by the primary traffic.

Since, for any particular intersection, one would need to make some field observations to infer the appropriate values of h_0 and h_1 anyway, one might as well observe directly the parameters most relevant to the question. If (for $qh_0 << 1$) one were to construct a graph of the cumulative number of departures of secondary vehicles vs. time (while there is a queue of secondary vehicles)

noting on the graph each interruption due to a primary vehicle, the value of $1/h_1$ would be interpreted as the slope of the cumulative curve between interruptions. The "theory" here merely predicts that the actual flow, i.e., the mean slope of the cumulative curve over a long time period, should differ from $1/h_1$ by an amount proportional to q, i.e., the mean number of interruptions per unit time. The "effective value" of the $(h_0 - h_1/2)$ could thus be measured directly.

Of course, this situation with $qh_0 \ll 1$ is quite unlikely to occur at a two-way stop sign, for example, because, if this happened, one would probably interchange the priorities. Analogous situations, however, exist at a four-way intersection where there is a queue in one direction but light flow in the cross direction.

At the other extreme, for $\exp(-qh_1) \ll 1$, the factor $\exp(-qh_0)$ in (2.9.10) can be interpreted as the probability that the headway between primary vehicles is larger than h_0 or equivalently that it is large enough to accommodate at least one secondary vehicle. This multiplied by q is the flow of such gaps. The second term of (2.9.10) similarly represents the flow of gaps large enough to accommodate at least two vehicles, the third term three vehicles, etc. For $\exp(-qh_1) \ll 1$, this series would converge very rapidly and, in most cases, could be approximated by just the first term.

$\exp(-qh_1) \ll 1$, this series would converge very rapidly and, in most cases, could be approximated by just the first term.

For more general headway distributions or more general forms for the $m(h)$, one can still decompose the capacity into a series of the type (2.9.10) with the first term equal to the flow of gaps which can accommodate at least one secondary vehicle, the second term the flow of gaps which can accommodate at least two vehicles, etc. The dependence of the individual terms of this series on the properties of the primary traffic and the secondary traffic, however, may be quite complex as compared with the approximations for $qh_0 \ll 1$. The

generalization of (2.9.9) depended on the secondary traffic only through the effective values of h_1 and $(h_0 - h_1/2)$ and it depended on the primary traffic only through the value of q (independent of the shape of the headway distribution).

If we keep the same form for $m(h)$ as in (2.9.5) but admit an arbitrary headway distribution $F(h)$, the generalization of (2.9.10), i.e., the series (2.9.6) gives

$$\text{capacity} = q[1 - F(h_0)] + q[1 - F(h_0 + h_1)] + q[1 - F(h_0 + 2h_1)] + \cdots \quad (2.9.11)$$

which clearly shows that the capacity depends on the shape of $F(\cdot)$ as well as the parameters h_0 and h_1. If different drivers accept different size gaps, however, the formulas are more complex.

There is another type of approximation for the capacity which is sort of a "mixture" of (2.9.9) and (2.9.11). If the primary traffic arrives in platoons or batches (possibly due to a traffic signal at some neighboring intersection), then it might be reasonable to assume that all headways within the batches are too short to accommodate any secondary vehicles. Indeed, one could define a vehicle to be "in a batch" if it follows another vehicle with a headway less than h_0. The $[1 - F(h_0)]$ in (2.9.11) would then (by definition) be the fraction of primary vehicles not in batches and $q[1 - F(h_0)]$ the "flow of batches."

If we now write (2.9.11) in the form

$$\text{capacity} = q[1 - F(h_0)] \left\{ 1 + \frac{[1 - F(h_0 + h_1)]}{[1 - F(h_0)]} + \frac{[1 - F(h_0 + 2h_1)]}{[1 - F(h_0)]} + \cdots \right\}, \quad (2.9.12)$$

the first factor is the flow of batches and the second factor represents the

average number of vehicles which can be accommodated in a headway which is known to be at least h_0. If most headways larger than h_0 can actually serve several vehicles (as might be the case for traffic pulsed from a neighboring traffic signal), then this second factor would be approximately $(1/h_1)$ times the mean headway of those headways greater than h_0. In any case, the formula (2.9.11) can be used to illustrate the consequences of compressing primary vehicles into platoons.

It is important to recognize here that if the secondary traffic must cross more than one lane of primary traffic, the same type of theory applies with q equal to the combined flow in all primary lanes, but the $1 - F(h_0)$, in particular, represents the fraction of the gaps in the superimposed stream with headway larger than h_0. The h_0, in turn, will also be larger for multiple lanes than for a single lane. Since the factor $\exp(-qh_0)$ in (2.9.8) is such a rapidly decreasing function of qh_0, a simple "rule of thumb" would be that one will have considerably difficulty finding a suitable gap if qh_0 is larger than 3/2 or 2.

There are some very elaborate formulas for waiting times and queue lengths of secondary traffic crossing a stream of primary traffic. The most critical parameter in these formulas, however, is the capacity or actually the traffic intensity or degree of saturation. A simple rule of thumb is that one will have significant queueing (an average queue length of perhaps three or four vehicles) if the degree of saturation exceeds about 3/4.

The "warrants" for installing a traffic signal or, perhaps even more important, the "public pressure" are such that traffic signals are usually installed at intersections (in most U. S. cities) for flows well below those which could be justified on the basis of capacities or delays. Although some of the issues discussed above are conceptually important, one seldom needs to evaluate capacities or delays with any precision because they are not actually used as the basis for making a choice.

Although we have emphasized various "economic" criteria for choices (delays, stops, etc.), it is difficult to put a price on "effort." A significant fraction of drivers, given a choice between two routes crossing a highway, one with a stop sign and the other with a signal, will choose the route with the signal even though the average travel time through the signal is obviously longer. Also, drivers who must wait at a stop sign until four or five cars have passed on the primary road become quite irritated even though the waiting time may be only ten or fifteen seconds. They may become even more irritated if they must wait in queue for other drivers to find gaps. Maybe drivers would behave quite differently if each user of a signal had to pay his "fair share" of the cost of maintaining the signal, or even a fraction of it. Certainly it is not obvious that the public would accept the consequences of a transportation system designed on the basis of measurable economic criteria alone.

2.10. Turning Traffic, Two-phase Signals

It is virtually impossible to analyze all the possible situations which can arise from turning traffic. Again the problem is not in the possibility of developing models or formulas; the techniques of analysis are fairly straightforward. The problem is that even for a standard four-way intersection which already required the introduction of possibly thirteen parameters, s_i , q_i , I_i , L , we must now also specify the fractions of left-turning traffic which already required the introduction of possibly thirteen parameters, s_i , q_i , I_i , L , we must now also specify the fractions of left-turning traffic for each direction, and possibly also the fraction of right-turning traffic. Worse yet, whereas with no turning traffic there was little interference between the various traffic streams and the formulas for capacity and delays contained only a few combinations of these thirteen parameters, we now find that nearly all of the original parameters plus all the new ones are potentially important in some situations. Furthermore, if there are pedestrians, they can interfere with the turning traffic which, in turn, can interfere with the through traffic.

Also the behavior of the traffic is sensitive to the geometry of the intersection, particularly whether or not there is room to store the turning vehicles which may need to wait for opposing traffic or pedestrians. If there is no place to wait, these vehicles may block the through traffic.

Traffic engineering books often describe various combinations of traffic movements one might use for multi-phase signals. The Highway Capacity Manual also contains some empirical recipes for the "average" reduction in capacity caused by turning movements, but this is obviously a very naive description of their effect. It would appear that local traffic engineers or consultants have, for the most part, been left to devise their own special remedies to reduce congestion at troublesome intersections. In most cases they do not have very many options in their choice of geometry of lanes, turn bays, or whatever. If they succeed in eliminating the problem, no one will ask if there is a better solution. There is very little guidance in the transportation literature on the relative advantages of various strategies.

For the intersection of two one-way streets, turning vehicles do not typically affect the capacity or delays very much. Indeed this is perhaps the main advantage of one-way streets. The only possible blocking would be caused by pedestrians interfering with the traffic. This would occur only on two of the four possible pedestrian crossings at the downstream side of the one-way streets. If there is only a small fraction of turning traffic giving at most one or two turning ve-

pedestrian crossings at the downstream side of the one-way streets. If there is only a small fraction of turning traffic giving at most one or two turning ve-

hicles per cycle, one could eliminate some of the interference by moving these pedestrian crossings about fifty feet downstream of the intersection so as to leave space for one or two turning vehicles between the crossing and the intersection where they could wait for pedestrians without blocking the through lanes.

Generally, the complication with turning traffic is that turning vehicles are distributed among the approaching vehicles in each direction. Vehicles cannot pass each other within their lanes, but turning vehicles must typically

yield to other types of traffic. If a turning vehicle must stay in a lane used also by through traffic, it will block the lane to through traffic until it is served. If one transfers the turning vehicle to some storage area (a turn bay), one must provide space for the storage (usually adjacent to the intersection) which might otherwise be used as a lane for through traffic.

The distribution of turning vehicles in each direction can generally be modeled quite accurately. For vehicles approaching the intersection in direction i, there should be a well-defined (observable) fraction p_i of drivers which wish to turn left and another fraction p_i' which wish to turn right. If there are several approach lanes in direction i, the p_i refers to the fraction of all vehicles in all lanes for direction i. As the vehicles approach the intersection, however, the left turning vehicles should all be in the left lane and the right turning vehicles in the right lane (if there is more than one lane). Until these vehicles are possibly caught in queues of different lengths in different lanes and jockey for position, it is reasonable to assume that each vehicle in the combined traffic stream is equally likely to be a turning vehicle, independent of what any other vehicle may do. Thus the behavior of any sequence of consecutive vehicles can be modeled as independent Bernoulli trials with probabilities p_i and $(1 - p_i)$ for a left-turn or not a left turn. (or p_i, p_i', and $1 - p_i - p_i'$ for a left turn, right turn, or

Bernoulli trials with probabilities p_i and $(1 - p_i)$ for a left-turn or not a left turn, (or p_i, p_i', and $1 - p_i - p_i'$ for a left turn, right turn, or no turn).

a. A F-C signal with no turn bays, single lane

To illustrate some of the issues, we consider first one of the worst cases but assume that the p_i are "small." Suppose that each approach has only one lane with no space anywhere for vehicles going straight to pass a turning vehicle which is stopped. The critical directions are assumed to be directions 1 and 2. If the vehicles in these directions can be accommodated, those in directions

3 and 4 will also. We will even disregard turns from directions 3 and 4. The arrival rates in directions 3 and 4 are, however, assumed to be sufficiently large that a driver turning left from directions 1 or 2, who must yield to the traffic in directions 3 and 4, will almost certainly wait until almost the end of the green interval to make his turn. He must wait until the queue vanishes in direction 3 or 4 and can then turn only if he can find a sufficiently large gap in the traffic stream. Even if he can find a gap, most of the green interval is likely to expire before he finds it.

A two-phase F-C signal will provide no response if a left-turning vehicle appears and blocks a lane. Obviously, if we knew that a left-turning vehicle were in the intersection from direction 1, it would be advantageous to do something immediately because no traffic can move in direction 1 until the vehicle is either removed by switching the signal to direction 2 (after clearing the intersection) or interrupting the flow in direction 3 to let the vehicle turn. For a F-C signal, part of the green interval will be wasted whenever there is a left-turning vehicle.

For $p_i = 0$, we saw that the capacity of the intersection could be increased by increasing the cycle time and thus reducing the fraction of time L/C lost in switching. For $p_i > 0$, however, there will be some finite value of the cycle time C which maximizes the capacity. If C is too small, one will spend too much time switching the signal, but if C is too large, the

of the cycle time C which maximizes the capacity. If C is too small, one will spend too much time switching the signal, but if C is too large, the

ith lane will be blocked after the first left-turning vehicle appears during a green interval. Only an average of about $1/p_i$ vehicles will pass before the traffic is blocked (independent of the length of the green interval) and for $C \to \infty$ the long time average output flow will go to zero.

To obtain some estimate of the effect of $p_1, p_2 > 0$ on the capacity, suppose that during a green interval G_1 in direction 1 there are $s_1 G_1$ "slots" in which the vehicles could leave. We will assume that vehicles can leave

in the last two slots whether there are turning vehicles or not; they might squeeze through the yellow, if necessary. In addition, there will be a vehicle served in the first slot if it does not turn. There will be one served in the second slot if neither of the first two vehicles turns, and one in the jth slot if none of the first j vehicles turns. The expected number of vehicles to pass during the green interval would therefore be

$$\begin{aligned} \text{number to leave} &= 2 + (1 - p_1) + (1 - p_1)^2 + \ldots + (1 - p_1)^{s_1G_1-2} \\ &= 1 + \left[1 - (1 - p_1)^{s_1G_1-1}\right]/p_1 \end{aligned} \tag{2.10.1}$$

for $s_1G_1 \geq 2$.

If it is highly likely that there will be at least one turning vehicle among s_1G_1 vehicles, i.e., $p_1s_1G_1 >> 1$, this becomes approximately

$$\text{number to leave} \cong 1 + 1/p_1 \tag{2.10.2}$$

nearly independent of G_1 . If, however, $p_1s_1G_1 << 1$

$$\begin{aligned} \text{number to leave} \cong s_1G_1 &- \frac{(s_1G_1 - 1)(s_1G_1 - 2)p_1}{2} \\ &+ \frac{(s_1G_1 - 1)(s_1G_1 - 2)(s_1G_1-3)p_1^2}{6} . \end{aligned} \tag{2.10.3}$$

The first term of (2.10.3) is the number of vehicles to leave for $p_1 = 0$. In the second term one can interpret $(s_1G_1 - 2)p_1$ as the probability that The first term of (2.10.3) is the number of vehicles to leave for $p_1 = 0$. In the second term one can interpret $(s_1G_1 - 2)p_1$ as the probability that the green interval will have one left-turning vehicle in the first $(s_1G_1 - 2)$ slots. The factor $(s_1G_1 - 1)/2$ is then the average location of this vehicle and the average number of slots blocked by the turning vehicle.

The average number of vehicles which can leave per unit time in directions 1 and 2 is

$$\text{capacity} = \frac{2 + \left[1 - (1 - p_1)^{s_1G_1-1}\right]/p_1 + \left[1 - (1 - p_2)^{s_2G_2-1}\right]/p_2}{G_1 + G_2 + L} . \tag{2.10.4}$$

We could add to this the vehicles in directions 3 and 4. But we are assuming here that if the arrival rates q_1 and q_2 in direction 1 and 2 are near capacity for their directions, then the arrival rates in directions 3 and 4 are below capacity. The outputs in directions 3 and 4 would then be q_3 and q_4 independent of G_1 and G_2.

We could seek values of G_1 and G_2 which maximize (2.10.4), but this would likely give a split with either G_1 and G_2 equal to zero. This is not what one typically wishes to do. Perhaps one would like to maximize this with respect to the cycle time $C = G_1 + G_2 + L$ but with a specified ratio for the outputs in the two directions (equal perhaps to q_1/q_2). It is somewhat easier, however, to vary C with a fixed split. Suppose we let

$$G_1 = \eta(C - L) \ , \ G_2 = (1 - \eta)(C - L) \tag{2.10.5}$$

and consider η as fixed.

To make a preliminary estimate of the optimal C and to illustrate the main issues, it is convenient to use the approximation (2.10.3) keeping only the first two terms, and also to approximate the $s_1G_1 - 1$ or $s_1G_1 - 2$ in the second term by $s_1\eta C$.

$$\begin{aligned} \text{capacity} \cong\ & [s_1\eta + s_2(1 - \eta)] - [s_1\eta + s_2(1 - \eta)]L/C \\ & - [p_1s_1^2\eta^2 + p_2s_2^2(1 - \eta)^2]C/2 \ . \end{aligned} \tag{2.10.6}$$

The first term of (2.10.6) represents the combined output if the signal could operate at the saturation flows s_1 or s_2 at fractions η and $(1 - \eta)$ of the time. The second term corrects this for the fact that the signal is idle a fraction L/C of the time. The third term represents the approximate loss due to turning traffic. To obtain an "optimal" C, one must balance the loss due to switching which is proportional to C^{-1} against the loss due to blocking which is proportional to C. The cycle time C_0 which maximizes (2.10.6) for fixed η is

$$\frac{C_0}{L} \cong \left\{ \frac{2[s_1\eta + s_2(1 - \eta)]}{[p_1 s_1^2 \eta^2 + p_2 s_2^2 (1 - \eta)^2]L} \right\}^{1/2} \tag{2.10.7}$$

giving a maximum capacity of approximately

$$\text{capacity} \cong [s_1\eta + s_2(1 - \eta)] - \left\{2[s_1\eta + s_2(1 - \eta)][p_1 s_1^2 \eta^2 + p_2 s_2^2 (1 - \eta)^2]L\right\}^{1/2} . \tag{2.10.8}$$

As a second approximation for small p_1, p_2, one should add to this an amount of approximately

$$p_1 s_1 \eta[2s_1\eta L + 3]/2 + p_2 s_2 (1 - \eta)[2s_2(1 - \eta)L + 3]/2$$

from other terms of (2.10.3) neglected in (2.10.6).

For a symmetric intersection with $s_1 = s_2 = s$, $p_1 = p_2 = p$, $\eta = 1/2$, this gives

$$\frac{C_0}{L} \cong \frac{2}{(psL)^{1/2}} , \tag{2.10.9}$$

and

$$\text{capacity} \cong s\left[1 - (psL)^{1/2} + \frac{p(sL + 3)}{2} + \ldots\right] . \tag{2.10.10}$$

The sL represents the number of vehicles which would be served during a lost time L, typically about five. The most interesting feature of this, however, The sL represents the number of vehicles which would be served during a lost time L, typically about five. The most interesting feature of this, however, is that C_0/L is proportional to $p^{-1/2}$ and the (minimum) loss of capacity due to blockage and switching is nearly proportional to $p^{1/2}$. These are very sensitive to small values of p. With only five percent turning vehicles and $sL = 5$, C_0/L would be only about four, giving a value for C_0 of perhaps fifty seconds, but already about a thirty percent (minimum) loss in capacity!. Actually this value of p is still not small enough for the approximation (2.10.10) to be very accurate. If one uses (2.10.3) to evaluate the capacity, one will

find that the loss in capacity is about thirty-three percent but, clearly, even a small fraction of turning traffic can have serious consequences for this particular geometry. Even for p equal to one percent, the (minimum) loss in capacity is about eighteen percent for C_0/L about 9.

The more general formulas (2.10.7) and (2.10.8) have similar qualitative behavior as the special case (2.10.9) and (2.10.10). Indeed one can write (2.10.7) and (2.10.8) in the latter form if one defines the p^* and s^* as suitable weighted average of the p_1, p_2 and s_1, s_2 respectively, specifically if

$$s^* \equiv s_1\mu + s_2(1 - \mu) \ .$$

$$p^* \equiv \frac{p_1 2_1 \mu^2 + p_2 s_2^2 (1 - \mu)^2}{[s_1\mu + s_2(1 - \mu)^2]} \ .$$

b. A F-C signal, multi-lane approach, no turn-bays

Suppose now that we had a two-lane approach in direction 1 but no turn bays. If there were also a two-lane approach in direction 3 with moderately heavy traffic, it would be quite unlikely that a turning vehicle in direction 1 would find a gap in both lanes of direction 3 before the end of the green interval. It is reasonable to assume, therefore, that a turning vehicle would block the left lane until the end of the green interval. There are, however, interval. It is reasonable to assume, therefore, that a turning vehicle would block the left lane until the end of the green interval. There are, however, other complications relating to the order in which vehicles are served.

Suppose, first, that when the red interval begins in direction 1 there are no vehicles remaining in the queue from the previous cycle, and, as the queue grows, vehicles distribute themselves between the two lanes so as to keep the queue lengths nearly equal at least until the first turning vehicle (if any) joins the queue in the left lane. If a turning driver, when he joins the queue, warns the vehicles behind him that he will turn left by flashing his turn signales, a subsequent driver in the left lane will know that he will be blocked

until the end of the green interval. If the latter driver is not also turning left, he would do better to join the queue in the right lane even if it is longer. The same is true also for any subsequent non-turning vehicles which might expect to be served in the right lane before the green interval expires. If a second left-turn vehicle arrives during that cycle, it will, however, join the queue in the left lane. This vehicle will likely manage to pass during the subsequent green interval, provided all the through vehicles, which arrived before it but after the first left-turn vehicle, moved to the right lane leaving two consecutive left-turn vehicles in the left lane. Unfortunately, a third left-turning vehicle would probably not pass during the next green interval and might be at the head of the line when the signal turns green the second time, thereby blocking the lane for the entire second green interval.

From the above (or an equivalent) starting state with no queue at the start of red, we can easily evaluate the expected number of vehicles which can be served in a green interval G_1. Suppose each lane has a saturation flow $s_1/2$ with $s_1G_1/2$ potential slots in time G_1, and p_1 is the fraction of turning vehicles in the combined arrival stream for both lanes. Analogous to the argument in (2.10.1), we assume that two vehicles can pass the intersection in the last two slots of each lane whether they turn or not. Two vehicles will pass in the first slots of the two lanes if they are both through-vehicles, with probability $(1 - p_1)^2 \cong 1 - 2p_1$ for $p_1 << 1$. Only one vehicle will pass (in the right lane) if one of the two vehicles wishes to turn, with probability $2p_1(1 - p_1) \cong 2p_1$. We will neglect the probability that two consecutive arriving vehicles both turn left. Vehicles will pass in each of the second slots of the two lanes if the first four arriving vehicles are through-vehicles with probability $(1 - p_1)^4$, etc.

The number of vehicles to leave in the left lane is, therefore

$$2 + (1 - p_1)^2 + \ldots + (1 - p_1)^{2[s_1G_1/2-2]}$$

$$= 1 + \frac{1 - (1 - p_1)^{s_1G_1-2}}{[1 - (1 - p_1)^2]}$$

$$\cong \frac{s_1G_1}{2} - \frac{p_1}{4}(s_1G_1 - 2)(s_1G_1 - 4) + \ldots .$$

Vehicles can presumably leave the right lane in every slot so

$$\text{number to leave} \cong s_1G_1 - \frac{p_1}{4}(s_1G_1 - 2)(s_1G_1 - 4) + \ldots \qquad (2.10.11)$$

As compared with (2.10.3) we see that this number is essentially the same as the combined numbers of two separate one-lane highways, each with saturation flow $s_1/2$, the same fraction p_1 of left-turn vehicles, and the same G_1. This is valid only for "sufficiently small" p_1 since it is based on an expansion in powers of p_1 to terms linear in p_1. The result is not surprising since, in effect, the two-lane approach behaves as if there were a fraction $2p_1$ of left-turn vehicles in the left lane and none in the right lane.

As for the single lane approach, there will be a cycle time C_0 which maximizes the capacity. Indeed it is nearly the same as for a single lane with the same p_j and the fractional loss in capacity due to switching and blocking is also nearly the same. If the p_j are not sufficiently small, one would ex-

the same p_j and the fractional loss in capacity due to switching and blocking is also nearly the same. If the p_j are not sufficiently small, one would ex-

pect the two-lane approach to have a smaller loss, but in the numerical example above with p = five percent, the loss is about thirty-three percent for a single lane but thirty-one percent for two lanes.

For less than about five percent left-turns, models of the above type are probably reasonable. The "optimal" cycle time is in a reasonable range and, for this cycle time, the majority of cycles will have no turning vehicles. Through-traffic would not avoid the left lane of a two-lane approach in fear of being

blocked, although the effect of blocking on the capacity may be appreciable.

If the fraction of left-turns is more than ten percent, but there is no place to store the left-turning vehicles, one may wish to consider other signal strategies. The above model is probably still reasonable for p = ten percent. The "optimal" cycle time is shorter than one would customarily use, but the capacity is quite insensitive to moderate changes in the cycle time. The loss in capacity increases only from about 31 percent to 37 percent as p increases from five percent to 10 percent. If, for p much larger than this, however, one does not reduce the cycle time to unrealistic values, there is a risk that in some cycles there may be three left-turning vehicles. This would cause the left lane to be blocked for the entire second green interval and cause additional loss not included in the above formulas. Also, through traffic may avoid the left lane and not take advantage of all the available slots.

Even if all through vehicles avoided the left lane, one could still serve $s_1 G_1/2$ in the right lane and presumably two (left-turn) vehicles in the left lane. If one had the same situation in direction 2, one would serve a total of $s_1 G_1 + 4$ per cycle in directions 1 plus 2, giving a capacity of

$$\frac{s_1 G_1 + 4}{C} = \frac{s_1}{2}\left[1 - \frac{(s_1 L - 8)}{s_1 C}\right]. \qquad (2.10.12)$$

The s_1 here refers to the two-lane flow so $s_1 L$ is likely to be around

The s_1 here refers to the two-lane flow so $s_1 L$ is likely to be around

10 or 12. This capacity is now an increasing function of C. As compared with just the single right lanes, the effect of adding four vehicles per cycle in the left lanes of directions 1 plus 2 is equivalent to subtracting four vehicles from the amount $s_1 L/2$ that is lost from the right lane due to switching, but we still expect $s_1 L/2$ to be larger than 4.

Formally, (2.10.12) would be a maximum for C arbitrarily large (infinite), but this would be equivalent to using the left lane as a parking lot for turning

vehicles which are never served. Presumably one would interpret the "capacity" as the maximum number of vehicles which could be served without causing any class of vehicles to be oversaturated. In the present case we would want the cycle time to be short enough that we could serve the left-turning vehicles in direction 1 at a rate of two per cycle while serving through vehicles at a rate of $s_1 G_1/2$, i.e.,

$$\frac{p_1}{1 - p_1} < \frac{4}{s_1 G_1} .$$

As applied to (2.10.12) this means approximately that

$$s_1 C \lesssim 8/p_1 \quad \text{or} \quad \frac{s_1 L - 8}{s_1 C} \sim p_1 \frac{(s_1 L - 8)}{8} .$$

Typically the factor in the brackets of (2.10.12) is between about 0.95 and one for $p < 20$ percent.

The models for turning traffic described so far are relatively simple because both the through vehicles and the turning vehicles in direction 1, for example, respond only to the signal but do not depend explicitly on the traffic behavior in any other directions. Primarily it illustrates the fact that even a small fraction of turning traffic (five percent or less) can seriously affect the capacity of an intersection, particularly one controlled by a F-C signal. For larger values of p (ten percent or more) the conse-

by a F-C signal. For larger values of p (ten percent or more) the conse-

quences could be very severe, possibly a 50 percent loss in capacity.

For small p, a suitably designed V-A signal will work much more efficiently. For the larger values of p it will generally be advantageous to use a multiphase signal as described in section 2.11 or to construct turn bays even if this means sacrificing a potential through-traffic lane on a multi-lane approach.

c. A V-A signal strategy

One of the deficiencies of a F-C signal, for small p_j, is that it has no way of detecting when a lane is blocked by a turning vehicle. The cycle time must be kept relatively short to limit the time during which a turning vehicle will block a lane. A V-A signal could be designed to detect a blocked lane and take appropriate action.

Ideally, one would like to detect a turning vehicle as soon as possible and remove it. If a detector had eyes, it might observe a vehicle flashing its turn signals; otherwise, one could not detect the vehicle at least until the vehicle stops some place in the intersection to wait for an opportunity to turn. Since different drivers will wait at different locations in the intersection, it seems impractical to place detectors in the intersection where they might also respond to other types of vehicle movements. If a detector placed near the stop line (of the left lane for a multi-lane approach) observes some minimum gap (say four seconds) after the signal has turned green and at least one vehicle has already crossed the detector, but a second detector further upstream indicates that the queue could not have vanished yet, then this should imply that the lane is blocked (and has already been blocked for several seconds). If the signal now switches to yellow and then to red, the turning vehicle will be removed. The signal will then proceed to switch back and forth in the usual way, when a queue vanishes or a left-turn ve-

turning vehicle will be removed. The signal will then proceed to switch back and forth in the usual way, when a queue vanishes or a left-turn ve-

hicle blocks a lane.

To see the consequences of this strategy, we first assume that there is negligible turning traffic in directions 3 and 4 but enough through traffic in these directions that a turning driver would have difficulty finding a gap. Suppose that the signal switches from direction 1 to 2 whenever either the

queue vanishes in direction 1 or a lane is blocked by a left-turning vehicle in direction 1. Similarly the signal switches from direction 2 back to 1 whenever the queue vanishes in direction 2 or a lane is blocked by a left-turning vehicle in direction 2. If there is a flow q_1 in direction 1, a fraction p_1 of turning vehicles, and the signal is undersaturated, then the signal will switch from direction 1 to 2 at a long-time average rate of p_1q_1 due to left-turning vehicles. Similarly it will switch from direction 2 to 1 at an average rate of p_2q_2 due to turning vehicles.

The number of switches from direction 1 to 2 must, of course, be equal to the number of switches from direction 2 to 1 (if the signal returns to where it started). If the q_1 and/or q_2 are barely at the limit of saturation, the queue in some direction will seldom vanish. The rate of switching from direction 1 to 2 or vice-versa will, therefore, be the larger of p_1q_1 or p_2q_2. If $p_1q_1 > p_2q_2$, then direction 1 will reach saturation while direction 2 is undersaturated. The signal will switch from direction 2 to 1 at a rate p_2q_2 due to left-turning vehicles and at a rate $p_1q_1 - p_2q_2$ because the queue vanishes. Otherwise, the signal will be kept busy serving vehicles either at a rate s_1 of s_2.

One could define four effective lost times per cycle depending on the type of signal switching, each of which should be (nearly) independent of the lengths of the green intervals, or therefore of the q_i or p_i. If one switches type of signal switching, each of which should be (nearly) independent of the lengths of the green intervals, or therefore of the q_j or p_j. If one switches from direction 1 to 2 because the queue vanishes, there will be a lost time, say L_1, the same as for a V-A signal with $p_j = 0$. There will also be a corresponding lost time L_2 for switching back. One may prefer, however, to define a lost time per cycle as done previously, $L_1 + L_2$, since only the sum is relevant. But we must now also consider a lost time $L_{1\ell}$ caused by a switch from direction 1 to 2 for a left-turning vehicle and a lost time $L_{2\ell}$

caused by a switch from direction 2 to 1 for a left-turning vehicle. Equivalently, one may wish to measure only the four combinations $L_1 + L_2$, $L_{1\ell} + L_2$, $L_1 + L_{2\ell}$, and $L_{1\ell} + L_{2\ell}$. These lost times will depend on the number of lanes, location of detectors, width of the intersection, etc., but there is not much point in trying to theorize about their values because, for any particular intersection, there is probably nothing one can do to change them. One can measure them from piecewise linear approximations to the cumulative departure curves as described in sections 2.1 and 2.2. We might expect the $L_{1\ell}$ to be larger than the L_1 because direction one will be blocked for a while before the detector recognizes it. On the other hand, a turning vehicle which is already in the intersection when the signal switches will clear during the yellow. We would expect, however, that all four lost times will be of comparable magnitude.

We can now evaluate the capacity of the intersection (for directions 1 and 2) in terms of these lost times. If the signal is at capacity and $p_1q_1 > p_2q_2$, the signal will spend a fraction of time q_1/s_1 serving vehicles in direction 1 at the rate s_1, a fraction q_2/s_2 serving vehicles in direction 2 at the rate s_2 and the rest of the time switching. The fraction of time lost in cycles with a switch from direction 2 to 1 caused by a left turn in direction 2 is $(L_{1\ell} + L_{2\ell})p_2q_2$. The fraction of time lost in cycles when the queue vanishes in direction 2 is $(L_{1\ell} + L_2)(p_1q_1 - p_2q_2)$. Thus, at capacity

$$\begin{aligned} 1 - q_1/s_1 - q_2/s_2 &= (L_{1\ell} + L_{2\ell})p_2q_2 + (L_{1\ell} + L_2)(p_1q_1 - p_2q_2) \\ &= p_1q_1(L_{1\ell} + L_2) + p_2q_2(L_{2\ell} - L_2) \,. \end{aligned} \tag{2.10.13}$$

Of course, the "capacity" depends on q_1/q_2 and s_1/s_2 but, for purposes of comparison with the F-C signal, suppose $q_1 = q_2 = q$, $s_1 = s_2 = s$, and $p_1 = p_2 = p$. The capacity would then be

$$q_1 + q_2 = 2q = \frac{2}{1 + (L_{1\ell} + L_{2\ell})ps/2} \quad . \tag{2.10.14}$$

This is to be compared with the F-C signal capacity (2.10.10). The L in (2.10.10) is the lost time per cycle (two switches) with a yellow time based on the approach speed. The $L_{1\ell} + L_{2\ell}$ is also a lost time per cycle potentially for a slower platoon speed but with some loss due to a delayed response to blocking. The two lost times should be similar. The main difference between the two formulas, however, is that loss in capacity due to turning traffic is proportional to the square root of the "small" dimensionless parameter psL in (2.10.10) but to the first power of the analogous parameter in (2.10.14). Besides, there is an extra factor of 1/2 in (2.10.14). If the loss $(psL)^{1/2}$ is 20 percent in (2.10.10), it would be about $(0.2)^2/2$ = two percent in (2.10.14). In the previous illustration with p = five percent, (2.10.10) gave $(\rho sL)^{1/2} \sim 1/2$ but the more accurate formulas gave a loss of only about 33 percent. The corresponding loss in (2.10.14) should be only about 12 percent for a single lane approach.

In the above comparisons, we have not considered the traffic in directions 3 and 4 except to postulate that it hinders the left turns in directions 1 and 2. For a F-C signal, the capacity is so sensitive to the p_1 and p_2 that if p_3 and p_4 were zero, but q_3 and q_4 were large enough to prevent turning traffic in directions 1 and 2, we would likely have q_3 and q_4 longer than q_1 and q_2. If there are turning vehicles from all four directions, we should interpret the directions 1 and 2 as the two cross directions which are most critical for the given ratios q_1/q_3 and q_2/q_4, and the given values of the p_i and s_i. These would not necessarily be the directions with the larger values of q_i/s_i as we would number them for $p_i = 0$. We could then argue that if the flows q_1 and q_2 could pass through the intersection, q_3 and q_4 could also, but the traffic in directions 3 and 4 would hinder

the turns in directions 1 and 2 (and vice-versa).

For the vehicle-actuated strategy described above, one may need to make some nontrivial modifications for p_3 , $p_4 > 0$. If the signal responds only to turning vehicles in directions 1 and 2, turning vehicles in directions 3 and 4 would block lanes in these directions for a time dictated by the green intervals for directions 1 and 2. In addition, traffic in directions 3 and 4 would be interrupted by the signal switches. In this case the capacity in directions 3 and 4 would be comparable with (perhaps even less than) that for a F-C signal operating with the mean cycle time for directions 1 and 2 and the fractions p_3 , p_4 of the turning traffic. Even for p_3 and p_4 equal to five percent or so, the capacity in directions 3 and 4 might be reduced by 30 to 50 percent. Even if the capacity is sufficient, the stochastic queueing in directions 3 and 4 might be unacceptably large.

If all the p_j are small, an alternative strategy would be to switch the signal whenever there is blocking in any direction. This would mean that the signal switches from directions 1, 3 to direction 2, 4 at a rate $p_1q_1 + p_3q_3$, and at a rate $p_2q_2 + p_4q_4$ from directions 2, 4 to 1, 3 due to turning vehicles. If the former is the larger rate, the capacity would be defined by a relation similar to (2.10.13) but with p_1q_1 , p_2q_2 replaced by $p_1q_1 + p_3q_3$, $p_2q_2 + p_4q_4$, respectively.

$p_2q_2 + p_4q_4$, respectively.

These formulas for the capacity of a V-A signal apply to either single or multiple lane approaches with the p_is_i and s_i interpreted as the values for combined traffic of all approach lanes in direction i. In order that turning traffic have only a "small" effect on the capacity, it is necessary that $p_is_iL_i$, for some suitable lost times L_i , be small compared with 1. For a single lane s_iL_i might be about five or six; but for a two-lane approach, it would be ten or twelve. Thus, a "small" p_i really means small compared with maybe 0.1, not 1.

Clearly the V-A strategy proposed here is questionable for p_i's much above five percent, particularly on multilane approaches. One certainly cannot profitably switch the signal every time any driver wishes to turn left if this means that the signal can serve only about five cars per lane per phase. If the p_i are too large, one cannot even guarantee that this V-A strategy is better than a F-C strategy. The F-C strategy would have the advantage for a multilane approach that it could serve two-left turning vehicles from the same direction per green interval, potentially four vehicles in the same signal switch for directions 1 plus 3, for example.

d. Left-turn bays

We have seen that left-turning vehicles can significantly reduce the capacity of either a F-C or V-A signal, particularly if there is five percent or more left turns. If, because of this, a signal becomes oversaturated or generates excessive queues, one can (a) ban left-turns, (b) add more lanes, (c) use a multiphase signal, or (d) construct left-turn bays.

The most common option (if the p_i are not too large) is to provide turn bays for storage. The main advantage of this over adding more lanes for through traffic is that a single length of road lane converted into turn bays will provide turn bays for both directions 1 and 3 (or 2 and 4) simultaneously. The turning traffic in a turn bay for direction 1 will not need to use the downstream continuation of the turning lane. Consequently, this downstream section can be used as a turn bay for direction 3.

It is difficult to formulate any well-defined criteria for the addition of turn bays because the typical number of lanes needed with no turn bays in directions 1 and 3, for example, is likely to be an even number (2 or 4) while the corresponding number of lanes with turn bays is likely to be an odd number (3 or 5). If there are physical constraints which virtually prohibit making a road more than two lanes wide (for two-way traffic), one has no choice, but

this would be a rare situation. Most roads have either a shoulder lane or a parking lane. One can usually take some of the shoulder lane or ban parking, restripe the lanes and squeeze in a turning lane of some reasonable length. This should certainly increase the capacity of the intersection, but these benefits must be balanced against possible issues of safety, construction cost, etc.

If, on the other hand, one had a four-lane road with two-lane approaches in directions 1 and 3 but no turn bays, we saw that, for sufficiently large p_1 and p_3, the two left lanes in directions 1 and 3 might be blocked so frequently that only left-turning vehicles would use them. In this case, each of these left lanes might serve only two (left-turning) vehicles per cycle. They would each behave almost like turn bays but on separate lanes. One might better use just one of these two lanes for turn bays for both directions 1 and 3 and give the extra lane to the through traffic in direction 1 or 3. (Unfortunately, it doesn't do much good to give just half a lane to each direction). Or perhaps one could take some of the shoulder lane or parking lane and squeeze in a fifth lane for a turn bay.

If, "at capacity" all traffic movements must be accommodated, the capacity of an intersection with turn bays will be constrained by the condition that the turning vehicles must be served. The capacity for the through traffic can be increased by increasing the cycle time (which would happen automatically for a turning vehicles must be served. The capacity for the through traffic can be increased by increasing the cycle time (which would happen automatically for a V-A signal driven by the through traffic), but increasing the cycle time will decrease the rate at which the signal can serve the turning traffic. If there are turn bays in all four directions and the signal can serve only two turning vehicles per direction per cycle, it would be necessary that the arrival rate of turning traffic per cycle be less than two, i.e.,

$$p_i q_i C \leq 2 \quad \text{for} \quad i = 1, 2, 3, 4 , \tag{2.10.15}$$

in which C is either the fixed cycle time for a F-C signal or the mean cycle time for a V-A signal (driven by the through traffic).

If L_t is the appropriate average lost time per cycle for the through traffic with the turn bays for the turning traffic, then a fraction L_t/C of time is used for switching the signal, which limits the capacity of the through lanes. The through flows are $(1 - p_i)q_i$ and we can number the directions so that the dominant directions are numbered 1 and 2, i.e., $(1 - p_1)q_1/s_1 > (1 - p_3)q_3/s_3$ and $(1 - p_2)q_2/s_2 > (1 - p_4)q_4/s_4$. The condition that all flows can be served is that

$$1 - (1 - p_1)q_1/s_1 - (1 - p_2)q_2/s_2 > \frac{L_t}{C} > \frac{p_i q_i L_t}{2} \quad \text{for all} \quad i , \qquad (2.10.16)$$

In the hypothetical case of a symmetric intersection with $q_i = q$, $s_i = s$ and either $p_i = p$ for all i or $p_1 = p_2 = p$ and $p_3 = p_4 = 0$, (2.10.16) simplifies to

$$\frac{2q}{s} < \frac{1}{1 - p + psL_t/4} . \qquad (2.10.17)$$

The term $-p$ in the denominator represents the contribution to the capacity from the two vehicles per direction which are in the turn bay. It is of little consequence compared with the other terms and would not even be there if $p_3 = p_4 = p$ and $p_1 = p_2 = 0$. The main feature of the formula is that it again contains the characteristic dimensionless product psL_t which is not necessarily small for reasonably small values of p.

This is to be compared with the analogous formulas (2.10.10) for a F-C signal and (2.10.14) for the specially designed V-A signal with no turn bays. The L_t in (2.10.17) is probably somewhat larger than the L in (2.10.10) because the intersection will be wider with a turn bay, but perhaps comparable with the $L_{1\ell} + L_{2\ell}$ in (2.10.14). The main difference is that (2.10.17) has a factor 1/4 where (2.10.14) has a factor 1/2. This is due essentially to

the fact that the turn bay serves two left-turn vehicles in directions 1 and 2 per signal switch, but the V-A signal serves only one. Actually (2.10.14) applies only for $p_3 = p_4 = 0$. For $p_3 = p_4 = p$, the turn bays could serve four left-turn vehicles per signal switch and (2.10.17) would still be valid, but the V-A signal would switch for only one vehicle and the factor of 1/2 would be replaced by one.

For typical acceptable cycle times of at least 40 seconds and $q_i = 1/5$ sec^{-1} per lane, the constraint (2.10.15) would mean that the p_i's should be less than 25 percent for a one-line approach and 12 percent for a two-lane approach. Thus the p's are limited to about the same range (somewhat larger) with turn bays as without, but an intersection with turn bays can accommodate considerably higher flows (for the same p's) than without turn bays.

If p_2 and p_4 were zero, one would not need a turn bay in directions 2 and 4. Indeed, we should not have one, because the turn bay would increase the width of the intersection and possibly increase the lost time in switching from direction 1 to 2. For sufficiently small p_2 and p_4 one might therefore build a turn bay for directions 1 and 3 but not for directions 2 and 4. In this case, it would clearly be preferable to have a V-A signal that switches from direction 2 to direction 1 whenever direction 2 or 4 is blocked by a turning vehicle than to have a F-C signal. In this case the signal would switch from direction 2 to 1 at a rate $p_2q_2 + p_4q_4$ due to turning vehicles but, presumably, the mean cycle F-C signal. In this case the signal would switch from direction 2 to 1 at a rate $p_2q_2 + p_4q_4$ due to turning vehicles but, presumably, the mean cycle time would be constrained also by the condition $p_1q_1C \leq 2$ or $p_3q_3C \leq 2$ as in (2.10.15). Thus, at capacity, the latter condition would dictate the mean rate of switching, the larger of $p_1q_1/2$ or $p_3q_3/2$, provided this is larger than $p_2q_2 + p_4q_4$.

If the signal is kept busy at all times except while switching, and $(1 - p_1)q_1/s_1 > (1 - p_3)q_3/s_3$, $q_2/s_2 > q_4/s_4$, the analogue of the conditions (2.10.13) or (2.10.16) would be of the form

$$1 - (1 - p_1)q_1/s_1 - q_2/s_2 > L'_{t\ell}(p_2q_2 + p_4q_4) + L'_t\left[\frac{1}{2}\max(p_1q_1 , p_3q_3) - p_2q_2 - p_4q_4\right] , \qquad (2.10.18)$$

in which $L'_{t\ell}$ is the mean lost time per cycle if the signal switches because of a left turn vehicle in direction 2 or 4 and L'_t is the mean lost time per cycle if the signal switches because it must return to direction 1 to accommodate the turning traffic in directions 1 or 3 (provided this term is positive).

If $L'_{t\ell} = L'_t$, the condition (2.10.18) would be independent of p_2 and p_4 . The cycle time, and therefore the capacity, would be determined by the turning vehicles in directions 1 and 3. If, as expected, the $L'_{t\ell} > L'_t$ because of the time lag in detecting a turning vehicle, the capacity would decrease somewhat with increasing p_2 and p_4 . As compared with the relation (2.10.16) with a turn bay in directions 2 and 4, (2.10.16) has a factor $(1 - p_2)$ multiplying the q_2/s_2 which results from the fact that the turn bay can accommodate the flow p_2q_2 of turning vehicles leaving only $(1 - p_2)q_2$ for the through lane. On the other hand, the L_t in (2.10.16) may be larger than the L'_t in (2.10.18). Chances are that there is little difference in the two strategies if $p_2q_2 + p_4q_4 < \frac{1}{2} p_1q_1$ and $\frac{1}{2} p_3q_3$. The loss in capacity due to turning traffic in directions 2 and 4 for a F-C signal could, however, be considerably larger under these conditions, with no turn bays in directions 2 and 4.

however, be considerably larger under these conditions, with no turn bays in directions 2 and 4.

e. Stochastic queueing

Most of the discussion above on turning traffic relates to the capacities. One could derive detailed formulas for queues lengths and delays in any of the above models, but the formulas would be so complex and contain so many parameters that they would only obscure the main issues.

There is a whole sequence of possible strategies with increasing cost to accommodate increasing traffic ; stop signs, F-C or V-A with or without

turn bays, or multiphase signals. Whether one uses a signal or not is likely to be based on safety, custom, politics, etc., rather than any measurable economic criteria. The choice among possible signal strategies depends on delays, in some qualitative sense, but the delays are so sensitive to the capacity that the first consideration should be whether or not the capacity for some signal strategy is sufficient to accommodate the (peak) flows for all traffic movements. If it is, one might then recognize that, if the demand for any movement is more than about 80 percent of the maximum service rate for that movement, then stochastic effects may cause appreciable delay (residual queues of three or four vehicles for a F-C signal or large fluctuations in cycle times for a V-A signal).

If the fractions of turning vehicles p_j are small (less than 10 percent), one is not particularly concerned with providing good service to the turning vehicles for their sake. We do not want to encourage more turning movements by making them too easy. We are mostly concerned with minimizing the damage they cause to others. Unfortunately, most strategies to minimize the damage give preferential service to the turning vehicles to prevent them from blocking the through traffic.

We will comment here only briefly on some of the queueing aspects of the strategies described above, more or less in order.

We will comment here only briefly on some of the queueing aspects of the strategies described above, more or less in order.

If there is only a single lane approach, no turn bay, and no room to pass stopped vehicles, vehicles will form a single lane queue. They will be served in order of their arrival whether they turn left or go straight. For a F-C signal, turning traffic will not only cause a substantial reduction in capacity, it will also generate a larger variance rate for the departures. The residual queue will still have a form similar to (2.3.7) but with a ρ determined by the reduced number served per green interval, and with an increased I_D .

A V-A signal which switches for each turning vehicle, or each vanishing queue will have considerably larger capacity (for sufficiently small p_j) . Unlike the F-C signal, it will automatically adjust to changing flows, partitioning the green intervals between the two phases so as to keep the queues from growing in any direction, if possible. The signal switchings for turning vehicles increase the lost time in switching and, therefore, reduce the capacity. They also cause both a residual queue and unequal cycle times, both of which cause increased stochastic delays. For flows close to capacity, the green interval in one of the directions will terminate with a turning vehicle nearly every cycle. The stochastic queueing in that direction will be even larger than for a hypothetical F-C signal with the same degree of saturation for that direction.

For multilane approaches or turn bays, the queue behavior is more complicated because the left-turn and through vehicles from the same directions are partially or completely separated. The vehicles are not necessarily served in order of their arrivals, and it is possible to form a large queue of one without the other, in particular for the turning vehicles. For a two-lane approach with no turn bays and sufficiently small p_j , the through traffic would use both lanes. The turning vehicles would reduce the capacity, but turning vehicles and through vehicles would be served almost in order of their arrival if each driver is clever enough to try to minimize his trip time. The queue behavior of and through vehicles would be served almost in order of their arrival if each driver is clever enough to try to minimize his trip time. The queue behavior of a F-C or V-A signal would be analogous to that of a hypothetical single lane approach with the same s_i and p_i .

An intersection with turn bays clearly has a larger capacity than one without turn bays, but the same number of through lanes. The capacity of the system, however, is restricted by the service rate of one of the turn bays, i.e., by (2.20.15). At capacity, one of the turnbays and two of the through traffic directions will be at capacity because, in an attempt to maximize the flow of

through traffic which can be accommodated in direction 1, for example, one might first increase the green interval for direction 1 by varying the split with direction 2 until the through traffic in direction 2 is at capacity, then one would increase the cycle time to reduce the loss in switching until one of the turning movements is at capacity.

In section d there was no discussion of whether the signal was F-C or V-A. In the latter case, the signal would presumably be driven by the through traffic in directions 1 and 2. The split between directions 1 and 2 would automatically adjust to changes in the flows in these two directions; but as the flows increase, so will the mean cycle time. If the mean cycle time becomes too large, however, one of the turning movements will become oversaturated. In theory, the capacity would be the same for both the F-C and V-A signal if all directions have turn bays and the F-C green intervals are properly split. In essence, the signal is serving four independent traffic streams simultaneously in each green interval. In the two directions for through traffic, the maximum number of vehicles that can be served per cycle is proportional to the green interval, but in the turn lanes the maximum is assumed to be two (independent of the green interval).

We saw in section 2.3 that a F-C signal serving stationary flows q_i with no turning vehicles would cause stochastic queues in all directions. If the q_i were sufficiently close to saturation $(1 - q_1/s_1 - q_2/s_2 << 1)$ the residual queues would be very large but one could balance the deterministic and stochastic were sufficiently close to saturation $(1 - q_1/s_1 - q_2/s_2 << 1)$ the residual queues would be very large but one could balance the deterministic and stochastic queueing by choosing a sufficiently large cycle time and some appropropriate splits so as to distribute the stochastic queueing equitably between the 1, 3 and 2, 4 directions. In particular, we saw that the optimal cycle time would be appoximately twice the minimum admissible cycle time which would keep the signal undersaturated.

With turning traffic and turn bays in all directions (and unconstrained storage space), there will be eight separate queues and the situation will be

quite different. If the q_i (and $p_i q_i$) are close to saturation for some choice of cycle time and splits (i.e., at least two through traffic directions and one turning direction are close to saturation) increasing the cycle time might reduce the delays for the through traffic in directions 1 and 2, but the capacity for the turning traffic is proportional to $1/C$. Even a small increase in C could oversaturate one of the turn bays or at least cause a very large increase in the expected queue length.

One could propose to determine cycle times and splits so as to minimize the total delay to all drivers. Since the fraction of turning vehicles is assumed to be small, the turning vehicles would have relatively little "weight" in the formula for total delay, but the delay for turning vehicles is very sensitive to the only parameters one can vary, namely, the two green intervals (or the cycle time and split). The net result would favor keeping the queues for the turning vehicles relatively short because this could be done without increasing delays to the through traffic very much.

Actually one would do this anyway but for quite different reasons. Although we may not wish to favor the turning vehicles, the turn bays have only a finite length. If a queue exceeds the storage capacity of its turn bay, the queue will back up onto the through lane. For a one-lane approach, this could block the through lane completely, but for a two-lane approach it might still limit the number of through vehicles which could reach the intersection during the green interval. For any given length of the turn bay, one is almost forced to choose a cycle time so that the queue of turning vehicles seldom exceeds the storage of the turn bay.

If a signal should become oversaturated causing a large queue of both through and turning traffic in some direction, this does not necessarily imply that the turning traffic will block the through traffic lanes or that even the turn bays will be full. Turning vehicles cannot enter a turn bay, even if there is room,

until they can reach the entrance of the turn bay. If the turn bay is not full, but a queue of through traffic extends upstream beyond the entrance to the turn bay, drivers who join the queue will pass the turn bay in order of their arrivals if there is only one approach lane. If there is more than one lane, drivers may jockey between the lanes, but they will presumably do so in an attempt to equate the waiting time in all lanes. They would also pass the turn bay approximately in order of their arrival.

If the signal is oversaturated, the turn bay is not full, and $s_i G_i$ through vehicles are served in direction i per cycle, then an average of

$$p_i s_i G_i / (1 - p_i)$$

turning vehicles will enter the turn bay per cycle. If the ith direction is undersaturated and the turn bay is not full, an average of $p_i q_i C$ turning vehicles will enter the turn bay per cycle. In either case for a F-C signal, the number of vehicles entering the turn bay per cycle should be approximately Poisson distributed and, with no restrictions on the storage, the residual queue should behave like any typical queue with Poisson arrivals and a degree of saturation

$$\rho_{\ell i} = \frac{p_i s_i G_i}{2(1 - p_i)} \quad \text{or} \quad \frac{p_i q_i C}{2} \tag{2.20.19}$$

for oversaturated or undersaturated conditions for the through traffic in direction i, respectively, and two departures per cycle from the ith turn bay. The for oversaturated or undersaturated conditions for the through traffic in direction i, respectively, and two departures per cycle from the ith turn bay. The probability, i.e., the fraction of signal cycles, that the residual queue length in the ith turn bay exceeds some number n will be approximately

$$\exp\left(-n2(1 - \rho_{\ell i})\right) ,$$

an exponential (geometric) distribution with mean length of approximately $1/2(1-\rho_{\ell i})$.

Most engineering recipes for the design of turn bays specify that the length of the turn bay should be chosen so that there is only a small probability that

the number of new arrivals per cycle exceed the storage capacity of the turn bay. In the present case with a maximum service rate per cycle of only about 2 vehicles and thus an arrival rate of less than 2 per cycle, this recipe would require a storage for only 3 or 4 vehicles (for a F-C signal). Actually the turn bay should be designed so that there is only a small probability that the number of new arrivals plus the residual queue exceed the storage capacity of the turn bay. The uncertainty of the latter is typically more important that the uncertainty of the former.

If we let $n_{\ell i}$ denote the storage capacity of the turn bay, a more reasonable recipe for choosing $n_{\ell i}$ is that there be a small probability that the residual queue exceed approximately $n = n_{\ell i} - 2$ (allowing a storage for two new vehicles plus the residual queue). We are not so much concerned here with estimating accurately the probability that the turn bay is full, as in establishing some criteria which will guarantee that this probability is "small." A reasonable criteria would be that

$$(n_{\ell i} - 2)2(1 - \rho_{\ell i}) \geq 2$$

giving a probability of about e^{-2} for a full turn bay.

Alternatively, for a turn bay of given size $n_{\ell i}$, one can interpret this as a restriction on the value of $\rho_{\ell i}$ which can be accommodated without danger of overflowing the turn bay, namely

as a restriction on the value of $\rho_{\ell i}$ which can be accommodated without danger of overflowing the turn bay, namely

$$\rho_{\ell i} \leq 1 - \frac{1}{n_{\ell i} - 2} \qquad (2.10.20)$$

Thus, if a turn bay can hold six vehicles, the $\rho_{\ell i}$ should not exceed about 3/4.

The main point here is that a turn bay, to be effective, must be capable of absorbing some of the fluctuations in the arrivals. For reasonable size turn bays, the degree of saturation should be kept appreciably less than one, for all turn bays. Actually, for small values of p_i, the restriction (2.10.19), (2.10.20) is not very severe. For a one-lane approach and p_i = 10 percent or

a two-lane approach with p_i = five percent, the G_i would be typically limited to be less than 30 seconds (for a F-C signal).

A V-A signal with no turning traffic has the obvious advantage over a F-C signal that it will automatically adjust the mean cycle times and splits to the prevailing values of the q_i. It will also respond more efficiently to stochastic fluctuations, provided that q_1/s_1 is not so close to q_3/s_3 or q_2/s_2 so close to q_4/s_4 that the queue in direction 1 or 2 vanishes before that in direction 3 or 4 respectively. A two-phase V-A signal can respond efficiently to stochastic fluctuations in two directions (1 and 2) but not four (or eight). As the cycle time drifts in response to fluctuations in the traffic in directions 1 and 2, however, the time needed to serve the traffic in directions 3 and 4 drifts in the same way. If $q_1/s_1 > q_3/s_3$ and $q_2/s_2 > q_4/s_4$, the queue in directions 3 and 4 will usually vanish before those in directions 1 and 2, regardless of the cycle time.

For sufficiently small $p_i > 0$ with turn bays which are (almost) never full, one certainly would not change the signal strategy to cater to the turning vehicles. The problem is that the natural response of a V-A signal to an excess of vehicles in directions 1 and 2 is for the cycle time to drift to higher values and stay there for several cycles until the excess has been served or there is a fluctuation in the opposite direction. This will, of course, cause an increase in the expected number of arrivals per cycle in the turn bays. Unlike the
a fluctuation in the opposite direction. This will, of course, cause an increase in the expected number of arrivals per cycle in the turn bays. Unlike the
through traffic in directions 3 and 4, however, an increase in the cycle time does not increase the number of vehicles served per cycle from the turn bays. One obviously should make some modifications in strategy if there is a danger that some turn bay will overflow and block an approach lane.

One could instrument the turn bays to measure flows, queue lengths (in the turn bay) or any other relevant information about past behavior. The only action one would take, however, in response to an excess of vehicles in

the turn bay is to limit future cycle times so as gradually to reduce future inputs per cycle to the turn bay. One has no mechanism with a two-phase signal to serve these vehicles immediately, and one probably would not do so even if one could, at the expense of interrupting other traffic movements.

To know only the length of the queue in the turn bay is not sufficient information itself to take some meaningful action. Any restriction of future cycle times should be based on an estimate of the future inputs per cycle $p_i q_i C$ or $p_i s_i G_i/(1 - p_i)$. Since the p_i is not likely to vary rapidly with time, one could make separate field measurements of them, or, if one had any type of detectors in the turn bays, one could infer values of the p_i from cumulative counts "on-line." Without elaborate equipment and sophiticated logic, it would seem that the only simple strategy to prevent overloading the turn bays would be to impose a maximum green interval G_{Mi} on the V-A signal chosen, so that

$$p_i s_i G_{Mi}/2(1 - p_i) \leq 1 - 1/(n_{\ell i} - 2) \text{ for all } i .$$

Under such a strategy, a green interval which runs to the maximum will leave a residual queue of through traffic (an unavoidable situation). The detectors will not sense how long the queue may be, but limiting the green interval in one direction will, in effect, also limit the green interval needed by the cross will not sense how long the queue may be, but limiting the green interval in one direction will, in effect, also limit the green interval needed by the cross traffic. The signal will return to the original direction as soon as it has served the cross traffic; and, if the signal is really undersaturated, the residual queue will be served eventually.

2.11. Multiphase Signals

At a four-way intersection there are potentially four through and four left-turn movements. At more complex intersections there may be even more movements; and, if there are pedestrians, one may need to worry about right-turning vehicles as well. We will continue, however, to limit our discussion

to the typical four-way intersections, disregarding effects of right-turning vehicles.

Of the eight possible movements, the signal can serve only two at a time without conflict and only a few combinations of these. The through traffic in direction 1 can be combined only with the through traffic in direction 3 or with the left-turns in direction 1. The left-turns in direction 1 can also be combined with the left-turns in direction 3; similarly for directions 2, 3, and 4. There are eight possible nonconflicting pairs in all; but, actually, one would never use more than six phases at a single intersection, three for the 1 plus 3 directions and three for the 2 plus 4 directions. There are, however, several possible three phases sequences for the 1 plus 3 directions, each of which can be combined with any of several possible sequences for directions 2 plus 4. It suffices, however, to consider the sequences for the 1 plus 3 directions separately from those for the 2 plus 4 directions.

In the discussion of the previous section it was, in effect, assumed that one would serve only the through traffic in direction 1 with the through traffic in direction 3 (and the same for directions 2 and 4). The left-turn movements could be accommodated only at the end of the green or during the yellow interval. In the absence of left-turn bays, a single left-turn vehicle would block a lane and the signal would typically serve only one left-turn vehicle per phase. It could conceivably serve as many as four if the left-turn vehicle in direction 1 and the signal would typically serve only one left-turn vehicle per phase. It could conceivably serve as many as four if the left-turn vehicle in direction 1 were followed (in the left lane of a multilane approach) by another left-turn vehicle; and, in the same phase, a left-turn vehicle blocked a lane of direction 3 and it were also followed by a left-turn vehicle. Even with left-turn bays, it was assumed that the yellow interval could accommodate at most two left-turn vehicles in direction 1 and/or two in direction 3 (although it may be difficult to accommodate two left-turn vehicles in the two directions simultaneously). With only limited storage in the left-turn bays (say for about six vehicles),

however, one should typically choose a cycle time so as to serve only an average of about one and one-half vehicles per cycle from any turn bay.

The final conclusion was that for $p_j \geq 10$ percent, the capacity of the intersection would be severely limited by the turning traffic. If one chose a cycle time long enough to serve the through traffic, there would be too many left-turn vehicles per cycle to be served in the yellow interval. Conversely, if one chose a cycle time short enough to serve the left-turn vehicles, one may lose so much time in switching that one cannot accommodate the through traffic.

The purpose in using some multiphase signal strategy is to exploit the fact that, for any pair of traffic movements being served, the amount of time needed to serve one of the movements will be larger than the other. Any excess time not needed by one movement can possibly be used by another traffic stream in a different pairing. The most common example would be that the through traffic in direction 1 (by our convention for numbering) requires more green time than direction 3. Any time not needed by direction 3 might, therefore, be assigned to the left-turns in direction 1 paired with the through traffic in direction 1. Similarly for directions 2 and 4.

There are quite a number of possible situations to consider. As in the last section, we will analyze first cases with no left-turn bays, then cases with separate left-turn lanes but no storage restrictions, and finally the last section, we will analyze first cases with no left-turn bays, then cases with separate left-turn lanes but no storage restrictions, and finally the general case of turn-bays with finite (possibly zero) storage.

a. No left-turn bays, four-phase signal

If for p_1 sufficiently large, one were to serve the through traffic in directions 1 and 3 simultaneously, a left-turn vehicle would likely block a lane so quickly that only a few vehicles could pass before one must suffer a lost time to switch the signal. There may be a cycle time C_0 which maxi-

mizes the capacity, but for $p_1 \geq 20$ percent, it is likely to be too short to be practical.

An alternative strategy would be to serve the 1 and 3 directions in separate phases, serving the through, right turns, and left turns in direction 1 simultaneously, then those in direction 3. For this strategy it is reasonable to assume that the saturation flow in direction 1 is (nearly) independent of p_1. The number of vehicles that can be served in a green interval G_1 is $s_1 G_1$ and the fraction of time needed to serve a flow q_1 in direction 1 is q_1/s_1. Similarly, $s_3 G_3$ vehicles can be served in an interval G_3 for direction 3 and a flow q_3 needs a fraction of time q_3/s_3. Together they need a fraction of time $q_1/s_1 + q_3/s_3$. If, however, there are any pedestrian crossings, this time might be considerably larger. A left (or right) turning vehicle must yield to any pedestrians and, in doing so, may disrupt the through traffic.

If (for sufficiently large p_2) one chooses also to serve direction 3 and 4 in separate phases, the fraction of time needed by all traffic will be

$$\frac{q_1}{s_1} + \frac{q_3}{s_3} + \frac{q_2}{s_2} + \frac{q_4}{s_4} \leq 1 - \frac{L_4}{C}. \qquad (2.11.1)$$

The L_4 is the effective lost time per cycle for four signal switches and L_4/C is the fraction of time spent in switching the signal. The L_4 may depend on the sequence of phases since one has the option of serving the four L_4/C is the fraction of time spent in switching the signal. The L_4 may depend on the sequence of phases since one has the option of serving the four directions in any order. If drivers making a left-turn do not try to "cut the corner" too close, it would seem that the sequence of directions 1, 4, 3, 2 in figure 2.2 would involve the least switching loss because drivers in direction 4, for example, could start to move as soon as the last vehicle in direction 1 has had time to pass the midpoint of the intersection.

The most important feature of (2.11.1), as compared with the corresponding relation (2.2.7) for a two-phase signal with $p_i = 0$, is that (2.11.1) has $q_1/s_1 + q_3/s_3$ and $q_2/s_2 + q_4/s_4$ where we previously had max $(q_1/s_1,$

q_3/s_3) and max (q_2/s_2 , q_4/s_4), which by our numbering convention are equal to q_1/s_1 and q_2/s_2 , respectively. In the worst case with $q_3/s_3 = q_1/s_1$ and $q_4/s_4 = q_2/s_2$, the left hand side of (2.11.1) would be $2(q_1/s_1 + q_2/s_2)$, which means that the capacity for $C \to \infty$ is only half that of the two-phase signal with $p_i = 0$. But if the traffic is unbalanced with $q_1/s_1 = 2q_3/s_3$ and $q_2/s_2 = 2q_4/s_4$, for example, the left hand side of (2.11.1) is only (3/2) $(q_1/s_1 + q_2/s_2)$ and the capacity is reduced by only a factor of 2/3.

The relation (2.11.1), however, is independent of p_i and is valid for single or multiple-lane approaches. For a F-C signal, it could be compared with the results in section 2.10a and b where we obtained a loss in capacity of 1/3 due to turning traffic even for p = five percent (and a larger loss for larger p). The V-A strategy of section 2.10c gave a larger capacity (2.10.13) but would still give a loss in capacity of this magnitude for p_i about 10 percent. Obviously, for sufficiently large p_i (typically for p_i larger than about 10 percent) and no turn bays, the capacity of this four-phase strategy will exceed that of any two-phase strategy.

The unpleasant feature of (2.11.1) is that, to achieve a capacity close to this maximum, it is necessary that C be large compared with L_4 , but L_4 could potentially be twice as large as the L for a two-phase F-C signal possibly as much as 20 seconds. Actually, drivers are likely to behave more

could potentially be twice as large as the L for a two-phase F-C signal possibly as much as 20 seconds. Actually, drivers are likely to behave more

aggressively at a four-phase F-C signal knowing that they may have a long wait if they miss their turn, so the L_4 may not be quite this large. A V-A signal, on the other hand, will be controlled by only one traffic direction at a time (one need not worry about fast vehicles arriving in direction 3 at the time the end of the platoon reaches the intersection in direction 1) and could be designed to put further pressure on stragglers at the end of the platoon.

For a F-C signal, one must choose a cycle time and split it four ways so as to balance stochastic queueing effects against the deterministic queueing caused by the alternating service. There will be a cycle time and splits which minimizes the total delay for all four directions, a generalization of (2.3.13) but with four queues instead of two. Theoretically, the ratio of the "optimal" cycle time to the minimum cycle time satisfying (2.11.1) will be similar for a four-phase signal as for a two-phase signal, i.e., about a factor of two. This may, however, be somewhat academic because one would probably limit the cycle time to one and one-half to two minutes even if this were less than optimal.

Whatever one chooses for a cycle time, one should first assign the minimum time $q_i C/s_i$ to each phase, then evaluate the excess time

$$\left(\frac{q_1}{s_1} + \frac{q_2}{s_2} + \frac{q_3}{s_4} + \frac{q_4}{s_4}\right)C - L_4$$

and distribute this excess among the four phases in the ratio of the $(q_i/s_i)^{1/2}$.

For a V-A signal one no longer has the problem of waiting for two queues to vanish, as discussed in section 2.7. The theory would be an obvious generalization of that discussed in section 2.6 for a two-phase signal. The mean cycle time for an undersaturated intersection would be the value of C which gives an equality in (2.11.1), but the L_4 for the V-A signal

The mean cycle time for an undersaturated intersection would be the value of C which gives an equality in (2.11.1), but the L_4 for the V-A signal

could be made appreciably less than for a F-C signal with a yellow interval based on the platoon speed rather than the approach speed. The cycle time will drift from cycle to cycle with a standard deviation comparable with the mean, and it may drift to unacceptably large values. One may wish to impose some maximum green intervals to limit this drift (at the expense of causing residual queues).

Since the optimal cycle time for a F-C signal may be unacceptably large, there is a strong incentive for having a V-A signal with four phases (or constructing some turn bays).

b. No left-turn bays, three-phase signal

One could combine any strategy for serving directions 1 and 3 with any other strategy for serving directions 2 and 4. If there were no turn bays in directions 1 and 3 but heavy turning traffic in at least one of these directions, one may choose to serve these directions with separate phases as described in part a. This does not necessarily imply that one must do the same for directions 2 and 4. If, for example, $p_2 = p_4 = 0$, directions 2 and 4 could be served simultaneously, consuming a fraction of time $\max (q_2/s_2, q_4/s_4)$. The analogue of (2.11.1) would then be

$$\frac{q_1}{s_1} + \frac{q_3}{s_3} + \frac{q_2}{s_2} < 1 - \frac{L_3}{C} , \qquad (2.11.2)$$

with L_3 the total lost time per cycle for the three-phase sequence. This obviously would give larger capacity, shorter cycle times, etc., than the four-phase strategy of part a.

One may encounter difficulties with this strategy, however, for p_2, $p_4 > 0$ even if p_2 and p_4 are "small." Even if one had long turn bays in directions 2 and 4, one could accommodate only two left-turn vehicles per cycle in directions 2 and 4. Thus, it would be necessary that

$$p_2 q_2 C < 2 \quad \text{and} \quad p_4 q_4 C < 2 , \text{ i.e., } \frac{1}{C} > \frac{p_2 q_2}{2}, \frac{p_4 q_4}{2} .$$

If one had only limited storage in the turn bays, one may need to restrict this to an average of 3/2 vehicles per cycle instead of two; and if there were no turn bays, this number would be less than one. Thus (2.11.1) would take the form

$$\frac{q_1}{s_1} + \frac{q_3}{s_3} + \frac{q_2}{s_2}\left(1 + \frac{p_2 s_2 L_3}{u}\right) < 1$$

or

$$\frac{q_1}{s_1} + \frac{q_3}{s_3} + \frac{q_2}{s_2} + \frac{q_4}{s_4}\left(\frac{p_4 s_4 L_3}{u}\right) < 1$$

with $u = 2, 3/2$, or less than one being the average number of turning vehicles which can be served per cycle in directions 2 and 4.

The problem here is that L_3 is larger than for a two-phase signal, perhaps fifteen seconds and $s_3 L_3$ or $s_4 L_3$ is likely to be about seven for a single-lane approach, 14 for a two-lane approach. The range of p_3 and/or p_4 in which one can serve the vehicles in directions 2 and 4 is even more restricted than for a two-phase signal.

Equivalently, one could say that by going to a two-phase sequence for directions 1 and 3, one has attempted to gain capacity by using a longer cycle time. But this reduces the capacity for the left-turn traffic in directions 2 and 4. Consequently, one may be forced also to use a two-phase sequence for the latter directions.

c. Capacity for six-phase signal, turn bays with unrestricted storage

If an intersection has turn bays with unlimited storage (turning lanes), the through traffic and the turning traffic do not interfere with each other as they approach the intersection. There will be eight separate queues and

the through traffic and the turning traffic do not interfere with each other as they approach the intersection. There will be eight separate queues and the traffic signal can serve any nonconflicting pairs in any sequence. The objective, generally, will be to serve two traffic movements simultaneously at the maximum rates as much of the time as possible.

To analyze some of these combinations, it will be convenient to use similar notation for all arrival streams. Let $q_{\ell i}$ be the arrival rates of left-turn vehicles in direction i (previously denoted by $p_i q_i$) and let

q_{ti} be the arrival rate of through traffic in direction 1 (previously denoted by $(1 - p_i)q_i$). We will also assume here that the number of left-turn vehicles which can be served in any interval $G_{\ell i}$ is a linearly increasing function of $G_{\ell i}$ with saturation flow $s_{\ell i}$, independent of which traffic movement it may be paired, but with some effective starting time or lost time which may depend on the strategy of sequencing. Similarly, as before, we will assume that there is a saturation flow s_{ti} for the ith through traffic, independent of the pairing, but also with a lost time which may depend on the sequencing. It is implied by the above assumptions that these saturation flows can be maintained as long as one pleases, which means that there is no limit on storage space, or, particularly, that the storage for one movement does not block the flow for any other (in contrast with the cases discussed in a and b with no turn bays).

If the arrival rates are so close to (but below) saturation that one must use a very long cycle time to accommodate the flows, and the $q_{\ell i}$ are not zero, one will not be able to serve the left-turn movements during the yellow intervals. They will require their own signal phases. One can define "effective green intervals" $G_{\ell i}$ or G_{ti} so that the output flow for a fully utilized green interval is (by definition) $s_{\ell i}G_{\ell i}$ or $s_{ti}G_{ti}$. Even if some traffic movements should receive more than one green interval per cycle or be paired with more than one other traffic movement, we can define the $G_{\ell i}$ or G_{ti} as the total effective green interval per cycle for that movement. It may be important also to identify how various switching strategies affect the total effective lost time per cycle, but, for any given sequencing strategy, we can formally define the total lost time per cycle as the difference between the cycle time C and the sum of all fully utilized component effective green intervals contributing to C.

Over a long period of time during which the arrival rates $q_{\ell i}$, $q_{\ell i}$ are stationary, a necessary condition for these traffic movements to be undersaturated is that the average number of arrivals per cycle be less than the maximum

$$q_{\ell i}C < s_{\ell i}G_{\ell i} \quad \text{and} \quad q_{ti}C < s_{ti}G_{ti} \quad \text{for all} \quad i \ .$$

Thus the flows $q_{\ell i}$, q_{ti} require fractions $q_{\ell i}/s_{\ell i}$ and q_{ti}/s_{ti} of the signal time. The signal, however, can serve two movements at a time in appropriate pairs.

In directions 1 and 3, the movements t1 and $\ell 3$ cannot be served simultaneously. Together they require a fraction of signal time $q_{t1}/s_{t1} + q_{\ell 3}/s_{\ell 3}$. Similarly, the movements t3 and $\ell 1$ require separate phases and a fraction of time $q_{t3}/s_{t3} + q_{\ell 1}/s_{\ell 1}$. The time allocated to the 1 and 3 directions, however, can be partitioned so that either t1 or $\ell 3$ is paired with either t3 or $\ell 1$. The total fraction of time needed for directions 1 and 3 is, therefore, the larger of the two sums, i.e.,

$$\max\left\{ \frac{q_{t1}}{s_{t1}} + \frac{q_{\ell 3}}{s_{\ell 3}} \ , \ \frac{q_{t3}}{s_{t3}} + \frac{q_{\ell 1}}{s_{\ell 1}} \right\} . \qquad (2.11.3)$$

Regardless of any previous convention of numbering directions, it will be convenient now to renumber the 1 and 3 directions so that the first of the two expressions in (2.11.3) is the larger. Similarly, the total fraction of time needed by directions 2 and 4 is

$$\max\left\{ \frac{q_{t2}}{s_{t2}} + \frac{q_{\ell 4}}{s_{\ell 4}} \ , \ \frac{q_{t4}}{s_{\ell 4}} + \frac{q_{\ell 2}}{s_{\ell 2}} \right\} \qquad (2.11.4)$$

$$\max\left\{ \frac{q_{t2}}{s_{t2}} + \frac{q_{\ell 4}}{s_{\ell 4}} \ , \ \frac{q_{t4}}{s_{\ell 4}} + \frac{q_{\ell 2}}{s_{\ell 2}} \right\} \qquad (2.11.4)$$

and we can renumber the 2 and 4 directions so that the first term is the larger.

The traffic in directions 1 and 3 cannot be paired with any traffic in directions 2 and 4, so the total fraction of time needed for all traffic movements is the sum of (2.11.3) and (2.11.4). If the signal is undersaturated, the total fractions of time needed must be less than one. If the signal time is fully utilized during each signal phase, one can define the total lost time per cycle due to switching L_6 so that

$$\max\left\{\frac{q_{t1}}{s_{t1}}+\frac{q_{\ell 3}}{s_{\ell 3}},\frac{q_{t3}}{s_{t3}}+\frac{q_{\ell 1}}{s_{\ell 1}}\right\}+\max\left\{\frac{q_{t2}}{s_{t2}}+\frac{q_{\ell 4}}{s_{\ell 4}},\frac{q_{t4}}{s_{t4}}+\frac{q_{\ell 2}}{s_{\ell 2}}\right\}$$

$$=\frac{q_{t1}}{s_{t1}}+\frac{q_{\ell 3}}{s_{\ell 3}}+\frac{q_{t2}}{s_{t2}}+\frac{q_{\ell 4}}{s_{\ell 4}}<1-\frac{L_6}{C}. \qquad (2.11.5)$$

In principle, one could choose a cycle time sufficiently large that L_6/C is arbitrarily small, so this equation, in effect, defines the "capacity" of the intersection. As a practical matter, however, the lost time with possibly six signal phases per cycle may be so large that one cannot make L_6/C "small" without making C unacceptably large. One may, therefore, be willing to sacrifice some capacity by using a smaller cycle time, possibly with fewer signal phases per cycle.

Equations (2.11.3) to (2.11.5) already illustrate the benefit of reassigning any unused time from one movement to some other movement. If, for example, the flows are highly imbalanced with q_{t1}/s_{t1} considerably larger than q_{t3}/s_{t3}, it is likely that there is more turning traffic in direction 1 than 3 and $q_{\ell 1}/s_{\ell 1} > q_{\ell 3}/s_{\ell 3}$. But the q_{t1}/s_{t1} and $q_{\ell 1}/s_{\ell 1}$ appear in different terms of (2.11.3). The flow $q_{\ell 1}$ will not affect the capacity for the through traffic in direction 1 unless the second term of (2.11.3) exceeds the first term. On the other hand, the first term of (2.11.3) is quite sensitive to $q_{\ell 3}$. Chances are that one can afford to assign only one lane to the turning term. On the other hand, the first term of (2.11.3) is quite sensitive to $q_{\ell 3}$. Chances are that one can afford to assign only one lane to the turning traffic in direction 3 even though there may be two or more lanes for the through traffic in direction 1, i.e., $s_{\ell 3} < s_{t1}$. This means that the turning traffic $q_{\ell 3}$ is using a disproportionally larger fraction of the signal time than q_{t1} but, in any case, it is taking time that would otherwise be available to the through traffic.

d. Signal sequences

It would seem from (2.11.5) that the order of the various signal phases is irrelevant. Indeed, it should be for sufficiently large C, but the order does have an important effect on the L_6, which, in turn, influences the admissible values of C and the delays. Also, some sequences are safer than others.

Figure 2.12 shows the possible sequences of signal phases for directions 1 and 3 (which cannot be mixed with those for directions 2 and 4). There are four possible starting phases a, b, c, d, indicated by the solid line arrows in phase 1. Depending on which of the two movements served in phase 1 requires more time, the second phase may be either 2 or 2' either of which will be followed by the same phase 3. For example, in sequence a starting with movements t1, t3, the second phase will be 2 if $q_{t1}/s_{t1} > q_{t3}/s_{t3}$, i.e., the movement t1 requires more time than t3, but otherwise the second phase is 2'.

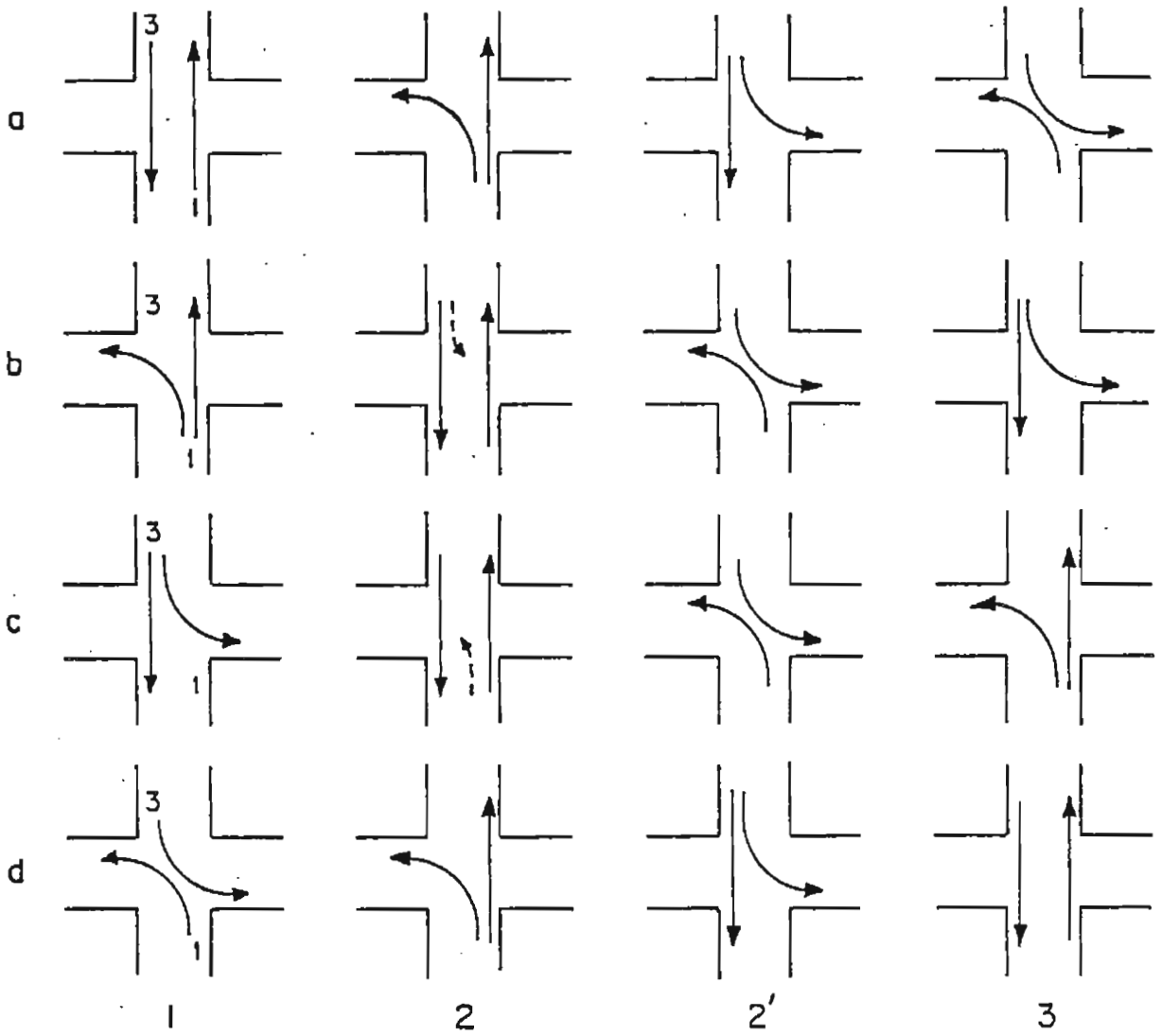
The sequences a and d are the same except that they are in the opposite orders, similarly for b and c. Sequence b starts with a "leading left" for direction 1 and ends with a "lagging left" for direction 3 if the intermediate phase is 2. Sequence c is also similar to b but with directions 1 and 3 interchanged.

Although each of these pairings is admissible by itself, the switch from

tions 1 and 3 interchanged.

Although each of these pairings is admissible by itself, the switch from

phase 2 or 2' to 3 in sequence a and the switch from phase 1 to 2' in sequence b or c is not advisable unless the left-turn movements can be confined to well-defined channels. In phase 2 of sequence a, for example, the left-turn vehicles, with no interference from the $\ell 3$ vehicles, may turn wide and confiscate some of the channel that the $\ell 3$ vehicles will want to use in phase 3. Even though the $\ell 3$ vehicles in phase 3 may have a left-



1 2 2' 3

Fig. 2.12 – Sequences of signal phases for directions 1 and 3.

turn arrow, they may not be able safely to enter the intersection. In sequence d, on the other hand, the two left-turn movements start at the same time and each should be aware that it must leave room for the other.

In sequence b or c it would be unlikely that one would want to go from phase 1 to 2' in both. For this to be true, the left-turn movements would require more time than the through movements, $q_{\ell 1}/s_{\ell 1} > q_{t1}/s_{t1}$ and $q_{\ell 3}/s_{\ell 3} > q_{t3}/q_{t3}$. In this case one should definitely use sequence d, but if the left turn movements are this heavy on multilane approaches, one would probably use double left-turn lanes.

If one insists that all left-turn movements have protected turn phases with a clear right-of-way over through traffic, sequence d would probably have less lost time in switching than b or c. In b, c, or d one could have a short yellow interval (if any) between phase 1 and 2 because the turning vehicles should be traveling slowly, also between phases 2 and 3 in d. In b or c, however, if the turn movements had a red signal in phase 2 but a turn arrow in phase 3, one could not start phase 3 until after the through traffic in phase 2 had been given a yellow interval which would virtually guarantee a safe start for the left turns in phase 3. In b, c, or d one would also need such a yellow interval following phase 3.

Suppose, however, one allows unprotected green intervals for turning traffic, in particular for the movement $\ell 3$ in sequence b or $\ell 1$ in sequence c. These

Suppose, however, one allows unprotected green intervals for turning traffic, in particular for the movement $\ell 3$ in sequence b or $\ell 1$ in sequence c. These vehicles could move out into the intersection as indicated by the broken line arrows in phase 2. They would be required to yield to the t1 or t3 traffic, respectively, but they would be ready to start phase 3 as soon as the through vehicles of phase 2 come to a stop. Also, the first left-turn vehicle would be starting from a position within the intersection. If per chance, the through traffic queue should vanish during phase 2 and there are at most two left-turn vehicles to be served in phase 3, then one can eliminate phase 3 and expect the

turning vehicles to pass during the yellow interval following phase 2.

The preferred choice between sequence b or c with an unprotected green for the turns in phase 2 is determined by which of the terms in (2.11.3) is larger (not necessarily by the $q_{1\ell}$ and $q_{3\ell}$ themselves). If, for example, the first term is the larger, then one would want to minimize the time lost in switching between movements t1 and ℓ3. One would use the sequence b, and see if one can also eliminate phase 3.

In the likely situation in which the fractions of turning vehicles are small, but the flows are highly imbalanced with q_{t1}/s_{t1} considerably larger than q_{t3}/s_{t3}, it is also likely that $q_{\ell 1}/s_{\ell 1}$ is larger than $q_{\ell 3}/s_{\ell 3}$. In this case, the sequence b with a leading green for $q_{\ell 1}$ is preferable to both d and to the two-phase signal discussed in the last section. For the two-phase signal, the cycle time is constrained by the condition that $p_i s_i C = q_{\ell i} C < 2$ for all i but one can still eliminate phase 3 of sequence b if only $q_{\ell 3} C < 2$. Thus with sequence b one can use a larger cycle time with essentially the same lost time per cycle for switching in directions 1 and 3, but a smaller fraction of time lost in switching and, therefore, larger capacity.

e) F-C signal, unrestricted storage

For a F-C multiphase signal, one must not only choose a cycle time, but one must split it into 4, 5, or 6 phases. The larger the number of phases,

For a F-C multiphase signal, one must not only choose a cycle time, but one must split it into 4, 5, or 6 phases. The larger is the lost time per cycle, so one would prefer to operate the signal with only four phases. For known values of the q_{ti}, s_{ti}, $q_{\ell i}$, $s_{\ell i}$ one should, however, first check to see if (2.11.5) can be satisfied for any sufficiently large C. If not, the signal will be oversaturated under any strategy and one would tend to prefer any socially acceptable strategy which would favor those movements with the largest saturation flows. This would typically imply giving preference to the through traffic over the left-turn traffic; but if the queue

of left-turn vehicles will block the through lanes, one may need to serve the turning vehicles at some minimum rates. We will not pursue this issue further because, as we have seen before, the objective is not clear if the signal is oversaturated.

If the signal is undersaturated, one should next check to see if the traffic can be accommodated with a four-phase signal strategy in which phase 3 is eliminated in sequence b for both the 1 plus 3 and 2 plus 4 directions. From (2.11.3) and (2.11.4) one first identifies which are the two constraining movements for directions 1 plus 3 and 2 plus 4 . Regardless of any previous numbering convention, we can relabel the directions so that the constraining movements are t1, ℓ3, t2 and ℓ4 .

In order that a four-phase signal which can accommodate two left-turn vehicles per cycle in each of the ℓ3 and ℓ4 directions be undersaturated, it is necessary that $q_{\ell 3}C < 2$ and $q_{\ell 4}C < 2$. If, however, these movements are served from turn bays which can hold only about six vehicles and the left turn vehicles can block the through traffic if they overflow the turn bays, it would be desirable to restrict the C even more so that

$$q_{\ell 3}C \leq 3/2 \quad \text{and} \quad q_{\ell 4}C \leq 3/2 \ . \tag{2.11.6}$$

If we know the total lost time per cycle L (for the four-phase signal), and, therefore, with (2.11.6), a lower bound on L/C

If we know the total lost time per cycle L (for the four-phase signal), and, therefore, with (2.11.6), a lower bound on L/C

$$L/C \geq (2/3)q_{\ell 3}L \ , \ (2/3)q_{\ell 4}L \ ,$$

one can then check if the through traffic can be accommodated, i.e., if

$$\frac{q_{t1}}{s_{t1}} + \frac{q_{t2}}{s_{t2}} < 1 - \frac{L}{C} < 1 - \max\left\{2/3\ q_{\ell 3}L \ , \ 2/3\ q_{\ell 4}L\right\} \ . \tag{2.11.7}$$

If not, then the four-phase signal will be oversaturated. One may need to use a five or six-phase signal.

We have used the notation L for the lost time here, the same notation as for the two-phase F-C signal in section 2.2. Although the turning vehicles in directions 3 and 4 might delay the start of vehicles in the next phase, the lost time here is essentially the lost time for just the movements t1 and t2 . It is assumed that the movements ℓ1, t3, and ℓ2, t4 can be accommodated during the phases for t1 and t2, respectively.

There is a slight complication here. It is possible (but not very likely) that, for the values of C satisfying (2.11.6), the time needed to serve the t1 traffic, Cq_{t1}/s_{t1}, may exceed that needed to serve the movements ℓ1 plus t3, $C(q_{\ell 1}/s_{\ell 1} + q_{t3}/s_{t3})$, by an amount less than the differences in the lost times for the two sequences; similarly for the t2 vs ℓ2 plus t4 movements. In such cases the role of the 1 and 3 and/or the 2 and 4 directions are, in effect, reversed.

If (2.11.7) is satisfied, one might now seek a C and splits so as to minimize the total delay to all vehicles. We would expect the delay to be most sensitive to the q_{t1} and q_{t2}, however, so we might choose the cycle time and the split between the 1 plus 3 and 2 plus 4 directions (the G_{t1} and G_{t2}) to minimize the delay to these movements. This, however, is equivalent to the problem discussed in section 2.3 with no turning traffic so we might next evaluate the "optimal cycle time" (2.3.13). If this cycle time, equivalent to the problem discussed in section 2.3 with no turning traffic so we might next evaluate the "optimal cycle time" (2.3.13). If this cycle time, which is likely to be about twice as large as the minimum C satisfying (2.11.7), also satisfies (2.11.6), then it would be a logical choice. Otherwise, we would use the maximum C satisfying (2.11.6).

Having chosen a C larger than the minimum value satisfying (2.11.7), we must now split this among the various movements, but we can do this in two

steps. First, the excess time not required by the t1 and t2 movements could be partitioned between these two in the ratio (2.3.15) so as to minimize the sum of the stochastic queueing delays for these two movements. This will determine the green intervals G_{t1} and G_{t2} which, in turn, will also determine the total time available to the ℓ1 plus t3 and ℓ2 plus t4 movements.

By our numbering convention, the minimum times needed per cycle for the latter movements are less than those for the t1 and t2 movements, respectively, so we now must decide how to split these times between the ℓ1 and t3, and between the ℓ2 and t4 . It is likely now that $q_{\ell 1}C$ and/or $q_{\ell 2}C$ are larger than 2 or at least larger than 3/2, otherwise we probably would not have been considering a multiphase signal; a two-phase signal would suffice. With possibly several left-turn vehicles served per cycle, it may be necessary to have longer turn bays or a separate left-turn approach lane. If there is no constraint due to storage of the queue, however, one might choose to split G_{t1} between ℓ1 and t3 so as to minimize the sum of delays to the latter movements. This problem is essentially the same type as the problem of splitting the cycle time C between t1 and t2 . We have some excess time to distribute, namely

$$G_{t1} - C\left(\frac{q_{\ell 1}}{s_{\ell 1}} + \frac{q_{t3}}{s_{t3}}\right) ,$$

$$G_{t1} - C\left(\frac{q_{\ell 1}}{s_{\ell 1}} + \frac{q_{t3}}{s_{t3}}\right) ,$$

(or actually, this less some effective lost time for switching between ℓ1 and t3). This excess time should presumably be distributed between the ℓ1 and t3 in the ratio

$$(q_{\ell 1}/s_{\ell 1})^{1/2}/(q_{t3}/s_{t3})^{1/2} .$$

Similarly, any excess time to be split between the ℓ2 and t4 directions should be split in the ratio

$$(q_{\ell 2}/s_{\ell 2})^{1/2}/(q_{t4}/s_{t4})^{1/2} .$$

If the four-phase signal is oversaturated, i.e., (2.11.7) cannot be satisfied, it is possible that a five-phase signal will suffice. Suppose, for example, that $q_{\ell 3} > q_{\ell 4}$ (if not, interchange the numbering 1,3 with 2,4). If we do not provide a separate phase for the $\ell 4$ traffic, then the cycle time should be chosen so that $q_{\ell 4}C \leq 3/2$, $L_5/C \geq (2/3)q_{\ell 4}L_5$, but now with L_5 the total lost time per cycle for the critical movements of the five-phase signal. We now check to see if the remaining traffic can be accommodated, i.e., from (2.11.5).

$$\frac{q_{t1}}{s_{t1}} + \frac{q_{\ell 3}}{s_{\ell 3}} + \frac{q_{t2}}{s_{t2}} < 1 - \frac{L_5}{C} < 1 - \frac{2}{3}q_{\ell 4}L_5 . \qquad (2.11.8)$$

We could now determine cycle times and splits so as to minimize the total delay, or generalize the approximate procedure described for the four-phase signal. The main difference is that if we split the cycle time in two stages, the first stage involves splitting C three ways to accommodate the traffic movements t1, $\ell 3$, and t2 . In the second stage, we redistribute the times for the t1 plus $\ell 3$ and for the t2 into two parts each for the t3 plus $\ell 1$, and t4 plus $\ell 2$ movements.

In the first stage one might also determine the C so as to minimize the total delay for movements t1, $\ell 3$, and t2 , subject to the condition $q_{\ell 4}C \leq 3/2$. There would be a formula, a generalization of (2.3.13) to three traffic movements, for the (nearly) optimal C , but the optimal will typically be at least twice the minimum C satisfying (2.11.13). It would be reasonable to use twice the minimum cycle time provided that it satisfies $q_{\ell 4}C \leq 3/2$ and is not unacceptably large.

Having chosen a C one can first assign each of the movements t1, $\ell 3$, and t2 its minimum time Cq_{t1}/s_{t1} , $Cq_{\ell 3}/s_{\ell 3}$, and Cq_{t2}/s_{t2} , respectively.

Then distribute any excess time among the three movements in proportion to $(q_{t1}/s_{t1})^{1/2}$, $(q_{\ell 3}/s_{\ell 3})^{1/2}$ and $(q_{t2}/s_{t2})^{1/2}$, respectively. The times for t1 plus $\ell 3$ and t2 can then be partitioned between the t3 plus $\ell 1$ and t4 plus $\ell 2$ as for the four-phase signal.

The existence of pedestrians may further restrict the suitability of a four or five phase sequence. Since pedestrians should not cross through a protected left turn movement, a minimum crossing time would be imposed on the phases for simultaneous through traffic in directions 1, 3, and 2, 4. For an asymmetric intersection (a minor road intersecting a major road) the crossing time for pedestrians is also likely to be larger for the crossing direction requiring the lesser time for the vehicles (pedestrians crossing a major road during the through traffic phase of the minor road). The problem here for the four or five phase sequence is that the pedestrian constraints may force the cycle time to such a large value as to violate one of the constraints (2.11.6) for elimination of one of the turn phases.

The six-phase signal will definitely accommodate longer flows than the five-phase signal. The analogue of (2.11.8) for the six-phase signal is (2.11.5) with no restriction on C other than it be socially acceptable and that the turning lanes be long enough that the queues for turning vehicles do not block the through lanes. If we view (2.11.8) and (2.11.5) as a constraint on the $q_{\ell 4}$, the two formulas are quite similar;(2.11.8) would imply

not block the through lanes. If we view (2.11.8) and (2.11.5) as a constraint on the $q_{\ell 4}$, the two formulas are quite similar;(2.11.8) would imply

$$\frac{2}{3} q_{\ell 4} L_5 < 1 - \frac{q_{t1}}{s_{t1}} - \frac{q_{\ell 2}}{s_{\ell 2}} - \frac{q_{t2}}{s_{t2}},$$

whereas (2.11.5) would imply that

$$\frac{q_{\ell 4}}{s_{\ell 4}} < 1 - \frac{q_{t1}}{s_{t1}} - \frac{q_{\ell 2}}{s_{\ell 2}} - \frac{q_{t2}}{s_{t2}}.$$

The former is the more severe restriction if $2/3L_5 > 1/s_{\ell 4}$, i.e., $s_{\ell 4} L_5 > 3/2$. But even for a single left-turn lane $s_{\ell 4} L_5$ is likely to be at least 5.

The L_6 in (2.11.5) is for a six-phase signal, probably in the range of 15 to 20 seconds. One would like to choose a C about twice the minimum which satisfies (2.11.5) but presumably the q/s values are already so large that the four or five-phase signals were unsatisfactory. The optimal cycle time is almost certain to be quite large, bordering on (or perhaps exceeding) the limit of acceptability. If one chooses to use a six-phase signal during peak traffic conditions, one should switch to a strategy with fewer phases at other times.

The partition of the cycle time can be made in a manner analogous to that described for the four or five-phase signal except that in the first stage one must split C four ways. Each traffic movement should be first assigned its minimum time Cq_{t1}/s_{t1} , $Cq_{\ell 3}/s_{\ell 3}$, etc., and then any excess time can be partitioned in the ratios of the $(q/s)^{1/2}$ values. The time allocated to 1 plus 3 and 2 plus 4 directions can then be partitioned again between the t3 plus $\ell 1$ and t4 plus $\ell 2$ movements. If there are pedestrians and the resulting split does not provide adequate time for pedestrian crossings, one should increase the cycle time until it does.

f) V-A signals

A V-A signal would not have a larger capacity than a six-phase F-C signal with unrestricted storage for left-turn vehicles, but it certainly could be designed so as to give considerably less delay. For an intersection with re-

with unrestricted storage for left-turn vehicles, but it certainly could be designed so as to give considerably less delay. For an intersection with restricted storage, however, a F-C signal would have restrictions on the cycle time, possibly so severe that one would not even use a six-phase strategy. A V-A signal might, in this case, have a higher capacity than any F-C strategy, but the equipment and the strategy needed to operate a V-A signal efficiently may be more complicated than what is commonly available.

For an intersection with no storage constraints, the main advantage of the V-A signal is that the detectors can observe if the queue for any traffic movement vanishes and can possibly terminate a phase earlier than for a F-C

strategy. The purpose would be to eliminate some of the time wasted by allowing a signal phase to continue after the output flow has dropped from some prevailing s to the corresponding q. Complications arise, however, from the fact that the signal will usually be serving two traffic movements simultaneously. The strategy must be such as to guarantee (if possible) that none of the eight traffic movements becomes oversaturated.

If there is restricted storage for left-turn vehicles, a V-A signal with the potential for operating on a six-phase cycle would have an advantage over either a F-C signal or a two-phase V-A signal in that it could regulate the time allocated to the left-turn phases and even keep a count of the vehicles in the turn-bays so as to prevent blocking while operating on a longer average cycle time than would be possible for a F-C or a two-phase V-A signal. The longer cycle time would reduce the fraction of time lost in switching and give a higher capacity.

One must first choose the sequence of signal phases. For a six-phase F-C signal, one would probably prefer the sequence d of figure 2.12, for both 1-3 and 2-4 directions because one can probably use shorter yellow intervals than in sequence b or c. During the off-peak, however, one should change to four or two-phase strategies at predetermined times. For a V-A, one might prefer the sequence b or c accordingly as the first or second term of (2.11.3) is larger. The advantage would be that if vehicle counts or

one might prefer the sequence b or c accordingly as the first or second term of (2.11.3) is larger. The advantage would be that if vehicle counts or a presence detector in the turnbay $\ell 3$ indicates that there are two or less vehicles waiting at the end of phase 2, then one can skip phase 3 and expect these vehicle to turn during the yellow interval. The signal would, therefore, automatically adjust to a four-phase strategy during the off-peak. The signal could also eliminate phase 1 if there are zero or one vehicles in the $\ell 1$ queue at the start of this phase. One would probably not eliminate this phase, however, if two vehicles were waiting, since this traffic movement is

not likely to affect the total time needed to serve the traffic in the 1, 3 directions.

Actually, the preference of b or c over d is not completely obvious because one can make similar modifications in sequence d for light turning traffic. If at the start of phase 1 of sequence d there are two or less vehicles waiting in the $\ell 3$ queue, for example, one could eliminate phase 1, go to phase 2, and then give the $\ell 3$ movements an unprotected green in phase 3. Of course, these vehicles will not be served until the end of phase 3, and, in the meanwhile, more $\ell 3$ vehicles may have joined the queue.

Most traffic controllers would not count the number of vehicles in the turn bays and it is probably not necessary anyway. The main advantage of sequence b or c is that one need not provide a turn signal for the turn movements in phase 3. During phase 2 one or two turning vehicles can move into the intersection and wait for a gap in the through traffic or the end of this phase. They will be ready to move at the earlier opportunity with little loss in switching time.

Typically, the turn bay would have a loop detector close to the intersection to detect the presence of a vehicle waiting in the turn bay itself, but, if a vehicle has moved into the intersection, it will not be on the detector. The controller would automatically skip phase 3 if there were no vehicles ac-

if a vehicle has moved into the intersection, it will not be on the detector. The controller would automatically skip phase 3 if there were no vehicles actually in the turn bay. If there is a vehicle on the detector at the end of phase 2, the signal would proceed to phase 3 and continue this phase until the detectors indicate that the queue has dissipated.

The type of strategy needed to infer when the queue vanishes in the turn bay is similar to that described in section 2.5 for the through lanes. Vehicles in the turn bay, however, will typically be moving at slower speed and drivers anticipating that the turn phase may be rather short, will tend to move with relatively close spacing in fear of being cut off. They do not expect the

Since the queues in the turn bays are likely to be short, a loop detector is much preferred over an impulse detector. It would be advantageous to have the inner edge of the loop displaced maybe 10 or 15 feet from the stop line so as barely to detect a vehicle stopped at the stop line. This extra space would serve a dual purpose. If at the end of phase 2 there were only two vehicles waiting, they would probably have moved off the detector (past the stop line) and the signal could skip phase 3. But if there were vehicles on the detector and the signal did switch to phase 3, one would wish to terminate the phase with a yellow signal at a time when some vehicle is so close to the intersection that it cannot stop.

To detect the "end of the queue," one must also have a mechanism for finding a minimum gap or the lack of vehicles over some section of road, or some combination of the two. For a single turning lane, it should suffice, in effect, to detect about a 3-second gap or the lack of vehicles over about a 75 foot section. The same type of detector system could be used for the turn bay served in either phase 1 or 3. The difference is that before the start of phase 1, the vehicles will not have moved into the intersection. One would probably choose not to skip phase 1 if any vehicles are present. If one has chosen sequence b or c in figure 2.12 according to whether the first or second term of (2.11.3) is the longer, the sluggish start of the turning traffic in phase 1 is not likely to affect the time needed to the first or second term of (2.11.3) is the longer, the sluggish start of the turning traffic in phase 1 is not likely to affect the time needed to complete the signal sequence.

If the turn bays have a restricted length which may cause the turn bays to become full, it would be desirable to have a loop detector near the far end of the turn bay to detect if the turn bay is full. If a moderate size loop (10 or 15 feet) becomes occupied and stays occupied for a time longer than it would normally take a moving vehicle to cross the detector, the controller could infer that a vehicle is stopped on the detector, i.e., the turn bay is

full. If this happens, presumably in phase 2, one would like to terminate this phase so that excess turning vehicles will not block the through traffic.

A possible V-A control strategy based on sequence b might proceed as follows. Phase 1 would terminate when the $\ell 1$ queue vanishes, regardless of any other movements (except that one can skip phase 1 if there are no vehicles in the turn bay). Phase 2 will terminate if either the $\ell 1$ or $\ell 3$ turn bay becomes full or the t1 queue vanishes. If there are pedestrians, however, this phase may require a specified minimum time. If at the end of phase 2, the t3 queue has already vanished and there are at most two vehicles in the $\ell 3$ queue, phase 3 will be skipped and the signal will turn to phase 1 for the 2-4 directions. Otherwise, phase 3 will continue at least until the $\ell 3$ vanishes, possibly until both the t3 and $\ell 3$ queues vanish or until some preset maximum G^*_{M3} .

For any V-A strategy, serving two independent traffic movements simultaneously, one encounters the type of complication discussed in section 2.7, at the end of phase 3. Although, by our convention of numbering so that the first term of (2.11.3) is less than the second, the t3 queue should usually (more than half the time) vanish before the $\ell 3$ queue, sometimes it will be the reverse. Particularly if the terms of (2.11.3) are nearly equal, one cannot afford always to wait for the t3 queue to vanish because the output flow will drop to $q_{\ell 3} + s_{t3}$, but if one does not wait at all, there is a risk that stochastic effects may cause the t3 queue to become too large. There is some "optimal" maximum time G^*_{M3} that the signal should wait for the t3 queue to vanish if the $\ell 3$ has already vanished. A corresponding strategy applies also for the directions 2 and 4.

The above strategy may not be the "optimal" strategy, but it should provide nearly maximum capacity, adjust automatically to changing flows, and give considerably less delay than any F-C strategy (typically, probably

less than half as much). Most of the time the V-A signal should behave as if it were serving only the traffic streams t1, ℓ3, t2, and ℓ4 in order, switching whenever a queue vanishes, but skipping the "lagging left" ℓ3 and ℓ4 phase whenever they are not needed. The mean cycle time should be close to the minimum needed to serve just these four movements, none of which can be served simultaneously.

Aside from the problems created when the t3 or t4 queues fail to vanish before the ℓ3 and ℓ4 queues respectively, the only modification in strategy occurs in phase 2 if the turn bays become full. Unlike any F-C strategy, however, there should not be any residual queue left in any turn bay at the end of its signal phase.

The size of the turn bays will restrict the capacity of the intersection with a V-A signal but not nearly as severely as for a F-C signal. The "capacity" is defined by a condition that, at capacity, all input flows can be served, but just barely for at least one of the input streams. Suppose, for illustration, that direction 1 is the critical direction; at capacity the $q_{\ell 1}$ and/or q_{t1} can barely be served.

As the flows approach capacity, signal phase 2 may terminate before the t1 queue vanishes, because a turn bay is full, causing a residual queue of through traffic in direction 1. This overflow may accumulate over many cycles, causing queue vanishes, because a turn bay is full, causing a residual queue of through traffic in direction 1. This overflow may accumulate over many cycles, causing the queue to back up beyond the entrance to the turn bay, preventing new ℓ1 vehicles from reaching the turn bay. In phase 1 of sequence b , any vehicles in the turn bay at the start of this phase plus perhaps one or two more which are swept into the turn bay by the moving t1 stream will be served during this phase but none will be served in phase 2 or until the next cycle.

In phase 2, ℓ1 vehicles will enter the turn bay only as fast as they can be swept in by the passing stream of through vehicles, i.e., at an average

rate $s_{t1}q_{\ell 1}/q_{t1}$. If the queue of $t1$ vehicles is sufficiently large, phase 2 will continue until the turn bay is full. Then all movement in this direction ceases until phase 1 of the next cycle, when the pattern repeats.

Clearly, if direction 1 is the critical direction, the flow at capacity is restricted only by the condition that the number of $\ell 1$ vehicles served per cycle is equal to the storage capacity of the turn bay, plus maybe one or two vehicles which can catch up with the queue while it is discharging. More generally, if $n_{\ell i}$ is the storage capacity of the turn bay for direction i and C is the average cycle time, the capacity of the intersection will be determined by the condition

$$q_{\ell i} \leq n_{\ell i}/C \quad \text{for} \quad i = 1, 2, 3, 4 \quad . \tag{2.11.9}$$

If it were known, for sure, that the first terms of (2.11.3) and (2.11.4) were the larger terms, one could theoretically operate the V-A signal so that phase 3 terminates when the $\ell 3$ queue (or $\ell 4$ queue for directions 2 and 4) vanishes, regardless of whether or not the $t3$ (or $t4$) has also vanished. The $t3$ or $t4$ queues might become very large if the two terms of (2.11.3) and (2.11.4) are nearly equal, but they would stay finite. The average cycle time would then be the minimum value satisfying (2.11.5) and the capacity would be defined by (2.11.5) and (2.11.9), i.e.,

$$q_{t1} \; . \; q_{\ell 3} \; . \; q_{t2} \; . \; q_{\ell 4} \; . \; . \quad L_6 \; . \; , \qquad q_{\ell i} L_6$$

defined by (2.11.5) and (2.11.9), i.e.,

$$\frac{q_{t1}}{s_{t1}} + \frac{q_{\ell 3}}{s_{\ell 3}} + \frac{q_{t2}}{s_{t2}} + \frac{q_{\ell 4}}{s\ell_4} < 1 - \frac{L_6}{C} < 1 - \max_i \frac{q_{\ell i} L_6}{n_{\ell i}} \; . \tag{2.11.10}$$

We do not have an analogous formula for the capacity of a F-C signal with restricted size of turn bays because we have not actually quantified the consequences of blocking when a turn bay overflows. The implication has been that the consequences are so severe that one would try to avoid it by using a much shorter cycle time than implied by (2.11.9).

Even if the intersection should become oversaturated, the V-A strategy

described above would yield a finite mean cycle time and well-defined splits, but not necessarily the most desirable ones. If direction 1 is oversaturated, phase 1 for directions 1 and 3 will last only long enough to empty the finite turn bay $\ell 1$ and phase 2 will last only enough to fill either the turn bay $\ell 1$ or $\ell 3$. Phase 3 will last long enough to empty the $\ell 3$ turn bay but will have a preset maximum after that. With an upper bound on the time allotted to the 1 and 3 directions, there will also be a bound on the number of arrivals per cycle in the 2 and 4 directions. If the V-A signal is capable of serving the vehicles in the 2 and 4 directions without exceeding the bounds on the turn bay storages $n_{\ell 2}$ and $n_{\ell 4}$, it will do so. If not, the time allotted to the 2 and 4 directions will be limited also by the size of the turn bays and the maximum green time extension in phase 3.

In principle, one could achieve any desirable average cycle time and splits for an oversaturated intersection by appropriate choice of the size of the turn bays, but this would seem to be a rather clumsy way of doing it. Since the fluctuations in the number of turning vehicles arriving in a fixed time (one minute, for example) is likely to be quite large, a signal driven by the turning vehicles is likely to have a rather large coefficient of variation for the cycle time. Normally one would try to make the $n_{\ell i}$ as large as possible, given various constraints on cost, geometry, etc., and try to control the splits or cycle times by some more direct control. Chances are that one would

given various constraints on cost, geometry, etc., and try to control the splits or cycle times by some more direct control. Chances are that one would provide turn bays so large that a cycle time constrained only by the turn bays would be unacceptably large.

If an intersection becomes oversaturated, one would like to give some preference to the traffic movements with the largest saturation flow so as to keep the sum of queues as small as possible. In particular, one would typically like to give low priority to the turning movements, but this is not possible. If one imposes a maximum time on phase 1 or 3 which could terminate those phases before the queue vanishes in the turn bay, the residual

queue in the turn bay might cause the turn bay to become full in phase 2 and indirectly control the duration of phase 2. Conversely, if one imposes a maximum time on phase 2, this will limit the number of vehicles which can enter the turn bays which, in turn, will limit the time needed to serve the turning vehicles in phase 3 and phase 1 of the next cycle. One certainly does not need to impose maximum times on all phases. If the dominant movement is the through traffic, it would be more effective to impose the maximum time only on the through movements in phase 2, and use the turn bays and phases 1 and 3 to absorb the fluctuations in turning traffic.

If one imposes maximum times on phase 2 for both the 1-3 and 2-4 directions, the final question is how one should choose the relative magnitudes of these. This is, of course, somewhat arbitrary since one is not willing to impose all the delays on one group of travelers. Some possible issues might be the following. If one has decided that direction 1 will be oversaturated, one might try to keep direction 3 undersaturated, perhaps also direction 2 and 4. If $q_{\ell 3}$ or $q_{\ell 4}$ is sufficiently small, one might try to limit phase 2 for directions 1-3, or 2-4 so as to give a high probability that one can skip phase 3. If one has turning lanes, in effect, infinite storage and no interference between the through and turning traffic, the strategies may be quite different because the input to the turn lane is deter-

storage and no interference between the through and turn vehicles are swept

strategies may be quite different because the input to the turn lane is determined by the $q_{\ell j}$ itself, not by the rate at which the turn vehicles are swept in by the through traffic. In this case one can control the turning traffic independent of the through traffic.

g. Short turn bays

In the discussion of parts e and f, it was assumed, or implied, that if one built turn bays or turning lanes, one would build them with sufficient storage capacity that they would not seriously affect the capacity of a five

or six phase signal. Actually for a F-C signal most of the analysis dealt with the possibility of skipping phase 3 for the 1-3 and/or 2-4 directions, but it was still assumed that there was adequate storage in those turn bays for which there was a separate left-turn phase. For the V-A signal, there was some discussion of the effect of a finite storage $n_{\ell i}$ but the capacity (2.11.10) is clearly not consistent with the formulas of parts a and b for $n_{\ell i} = 0$. Obviously, the theory is not quite complete yet. If the $n_{\ell i}$ are "small," there must be some transition between the strategies discussed in parts e and f for "large" $n_{\ell i}$ and those of parts a and b for $n_{\ell i} = 0$.

We first note that the existence of short turn bays (for one or two vehicles), or even some room for through traffic to sqeeze past a turning vehicle in the intersection, will considerably increase the capacity of the two-phase signal discussed in section 2.10a and b for small p_i. For a F-C signal with short turn bays in all directions, one could possibly use a cycle time constrained by the condition that an average of about one left-turn vehicle appears in any direction per vehicle per cycle, i.e., $p_i q_i C = q_{\ell i} C < 1$ for all i, so that the capacity constraint would be

$$\frac{q_{t1}}{s_{t1}} + \frac{q_{t2}}{s_{t2}} < 1 - \frac{L}{C} \leq 1 - \max_i q_{\ell i} L \ . \qquad (2.11.11)$$

$$\frac{q_{t1}}{s_{t1}} + \frac{q_{t2}}{s_{t2}} < 1 - \frac{L}{C} \leq 1 - \max_i q_{\ell i} L \ . \qquad (2.11.11)$$

For a two-phase V-A signal it was suggested in section 2.10c that with no turn bay one might switch the signal whenever the intersection was blocked by a turning vehicle. In section 2.10e, however, with reasonable size turn bays, it was suggested that one might use a maximum green to control the queue in the turn bays. Actually, in the latter case, one could achieve a larger capacity with a strategy in which one switched the signal whenever a turn bay became full (or the queue vanishes in the through direction). For flows close

to capacity, however, this strategy would lead to highly fluctuating cycle times and large stochastic queueing.

Short turn bays cannot absorb fluctuations in the number of turning vehicles per cycle. If each turn bay can store at least two vehicles, a reasonable strategy for a two-phase signal would be to switch the signal whenever the queue vanishes for the through traffic or a turn bay becomes full. If a turn bay is full, presumably two vehicles can leave that turn bay during the switch.

If, for flows close to capacity, only one of the turn bays becomes full in any signal cycle, and it is always the same turn bay, the capacity of the signal will be constrained only by the condition that the average cycle time must be such that $q_{\ell i}C \leq 2$ for all i. Thus the flows would be constrained by

$$\frac{q_{t1}}{s_{t1}} + \frac{q_{t2}}{s_{t2}} \leq 1 - \frac{L}{C} \leq 1 - \max_i \frac{q_{\ell i} L}{2} , \qquad (2.11.12)$$

and one could operate with an average (but highly fluctuating) cycle time perhaps twice that for a F-C signal. If, however, the turn bay which becomes full during any signal cycle is not always the same one, the signal will switch more frequently and give a lower capacity. The actual capacity constraint in this case is a rather complicated function of several of the $q_{\ell i}$.

more frequently and give a lower capacity. The actual capacity constraint in this case is a rather complicated function of several of the $q_{\ell i}$.

Neither of the above two-phase strategies can accommodate very high fractions of turning vehicles. Our main concern in this section is to establish a bridge between the multiphase strategies discussed in sections 2.11a, b, and those discussed in sections c, d, e, and f. The former could be interpreted as special cases of the latter in which phase 2 of sequence b in figure 2.12 (or both phase 1 and 3) are eliminated. The main difference between the two, however, is that, with no turn bays and no phase 2, the main function of

phases 1 and 3 is to serve the through traffic (if the time needed for the through traffic exceeds that for the turning traffic, i.e., $q_{t1}/s_{ti} > q_{\ell i}/s_{\ell i}$). With large turn bays, however, we expect that much of the through traffic will be served in phase 2. The main function of phases 1 and/or 3 would then be to serve the excess turning traffic that cannot be accommodated in phase 2 or during the signal switch.

For short turn bays suppose now that the duration of phase 2 is limited by the storage of the turn bays and cannot be adjusted so as to provide needed capacity for the through traffic. The duration of phases 1 and 3 must now be adjusted to accommodate both the through traffic and the turning traffic, but particularly the former.

In serving directions 1 and 3, suppose that the (effective) time G_1^* for phase 2, which may depend on the storage of the turn bays, has been specified. During the time G_1^*, the signal can serve $s_{t1}G_1^*$ and $s_{t3}G_1^*$ through vehicles in directions 1 and 3, respectively. If, during a cycle time C, $q_{t1}C$ and $q_{t3}C$ through vehicles arrive, then one will need to serve $\max\{0, q_{t1}C - s_{t1}G_1^*\}$ and $\max\{0, q_{t3}C - s_{t3}G_1^*\}$ through vehicles in phase 1 and 3. The fraction of the cycle time needed to serve the through traffic in directions 1 and 3 during all three phases is therefore

$$\max\left\{0, \frac{q_{t1}}{} - \frac{G_1^*}{}\right\} + \max\left\{0, \frac{q_{t3}}{} - \frac{G_1^*}{}\right\} + \frac{G_1^*}{} . \qquad (2.11.13)$$

during all three phases is therefore

$$\max\left\{0, \frac{q_{t1}}{s_{t1}} - \frac{G_1^*}{C}\right\} + \max\left\{0, \frac{q_{t3}}{s_{t3}} - \frac{G_1^*}{C}\right\} + \frac{G_1^*}{C} . \qquad (2.11.13)$$

Presumably at least one of the first two terms of (2.11.13) must be positive; otherwise, the through traffic in both directions 1 and 3 could be accomodated in the time G_1^*. The time G_1^* could then be reduced so as no longer to be constrained by the storage of the turn bays. If $q_{t1}/s_{t1} > q_{t3}/s_{t3}$, then certainly the first term of (2.11.13) must be positive.

If $q_{t3}/s_{t3} < G_1^*/C$, phase 2 can accommodate the through traffic in direction 3. Again we could reduce the time G_1^* by increasing the time given to phase 1 (if necessary) to accommodate the extra $t1$ traffic eliminated from phase 2. Actually, the first two terms of (2.11.3) can not only be made positive, they should even exceed the times (if any) needed to serve the left-turn vehicles in phases 1 and 3. Otherwise, one could reduce the time G_1^* by shifting some time from phase 2 to phases 1 and/or 3, so that the size of the turn bay is no longer a constraint. Thus, the minimum fraction of time needed to serve the traffic in directions 1 and 3 (including left turns) would be

$$\frac{q_{t1}}{s_{t1}} + \frac{q_{t3}}{s_{t3}} - \frac{G_1^*}{C} . \qquad (2.11.14)$$

In the special case with $G_1^* = 0$ (no turn bays and p_1 and/or p_3 too large for a two-phase signal), (2.11.14) reduces to that described in sections 2.11a, b with separate phases for the 1 and 3 directions. At the other extreme, one will use the above strategy only if (2.11.14) is larger than the corresponding fraction of time needed to serve the 1 and 3 directions by any strategy for large turn bays as described in sections 2.11c-f.

One can combine the above strategy for serving the 1 and 3 directions with any appropriate strategy for serving the 2 and 4 directions, but if one employs the same type of strategy in the 2 and 4 directions, the capacity would be re-

any appropriate strategy for serving the 2 and 4 directions, but if one employs the same type of strategy in the 2 and 4 directions, the capacity would be re-

stricted by the condition

$$\frac{q_{t1}}{s_{t1}} + \frac{q_{t3}}{s_{t3}} + \frac{q_{t2}}{s_{t2}} + \frac{q_{t4}}{s_{t4}} - \frac{G_1^*}{C} - \frac{G_2^*}{C} \leq 1 - \frac{L_6^*}{C} , \qquad (2.11.15)$$

in which G_2^* is the constrained time on phase 2, of sequence b of figure 2.12, for the 2-4 directions and L_6^* is the lost time per cycle for this six-phase sequence.

This should now be compared with corresponding expression (2.11.1) for the four-phase sequence (with $G_1^* = G_2^* = 0$) but with L_6^* replaced by the L_4 for the four-phase signal. If the vehicles in the turn bays do not interfere with the through traffic in directions 1 and 3 during phase 2, the L_6^* and the L_4 should be essentially equal. Since the durations of phases 1 and 3 are dictated by the through traffic, the G_1^* or G_2^* are, in effect, the time of "overlap" between phases serving the 1 and 3 directions (or the 2 and 4 directions). Clearly any overlapping of these phases when one can serve the through traffic in two directions simultaneously will give a higher capacity than with no overlap.

For a F-C signal one would like to choose G_1^* as large as possible, yet small enough so that it is unlikely that the turn bays will fill during any cycle and block a through traffic lane. If the turn bays empty during phases 1 and 3 (they should usually do so), the number of t1 and t3 vehicles served in time G_1^* is $s_{t1}G_1^*$ or $s_{t3}G_1^*$, respectively. The <u>average</u> number of turning vehicles swept into the turn bay during this time is $(q_{\ell 1}/q_{t1})s_{t1}G_1^*$ or $(q_{\ell 3}/q_{t3})s_{t3}G_1^*$. If the storage capacity of the turn bay is $n_{\ell 1}$ or $n_{\ell 3}$ (not large), one should probably choose the G_1^* so that an average of only about half this number of vehicles enter the turn bay each cycle. Thus a reasonable choice of G_1^* might be

about half this number of vehicles enter the turn bay each cycle. Thus a reasonable choice of G_1^* might be

$$G_1^* \equiv \min\left\{\frac{q_{t1}n_{\ell 1}}{2s_{t1}q_{\ell 1}} \ , \ \frac{q_{t3}n_{\ell 3}}{2s_{t3}q_{\ell 3}}\right\} . \qquad (2.11.16)$$

It is implied, of course, that this value of G_1^* is a genuine constraint. The G_1^*/C is less than one would need to operate phase 2 with no storage constraint as in (2.11.3), i.e.,

$$\frac{q_{t1}}{s_{t1}} + \frac{q_{t3}}{s_{t3}} - \frac{G_1^*}{C} > \frac{q_{t1}}{s_{t1}} + \frac{q_{\ell 3}}{s_{\ell 3}}$$

$$\frac{G_1^*}{C} < \frac{q_{t3}}{s_{t3}} - \frac{q_{\ell 3}}{s_{\ell 3}} .$$

Also this G_1^* is less than one would use in a two-phase strategy (if it is possible to use a two-phase strategy).

Having chosen the G_1^*, one must still choose a cycle time and a split between phases 1 and 3 for the 1-3 directions and between the phases for the 1-3 and 2-4 directions so as to balance the deterministic and stochastic queueing. Needless to say, if it is necessary to use a strategy such as this, (with a rather large total lost time in switching), the cycle time will be quite large.

We might note here that one has a choice of using either the sequence of phases b or c in figure 2.12. If one uses an unprotected green in phase 2 as indicated in figure 2.12, the intersection itself can store one or two left-turn vehicles for one of the two directions. It would be advantageous to choose between sequences b and c so that the extra storage in the intersection is assigned to the left-turn movements that constrain the G_1^* in (2.11.16).

Unfortunately, it is quite awkward to design some V-A strategy for this situation in which short turn bays constrain the time for phase 2. One would like to run phase 2 until a turn bay becomes full, but phase 1 of either sequence b or c must be such that the through traffic in directions 1 or 3 is served during the combined time of phases 1 and 2. One cannot know the quence b or c must be such that the through traffic in directions 1 or 3 is served during the combined time of phases 1 and 2. One cannot know the minimum time needed in phase 1 without knowing how long phase 2 will last.

If one interchanged phases 1 and 2, one could first serve the two through traffic streams simultaneously until one of the turn bays filled, then run separate phases for the 1 and 3 directions until both the left turn and through traffic queues are served in each phase. The awkward feature of this strategy is that one of the through traffic directions must be interrupted in the second phase and started again in the third phase. Also, the duration of the first

phase would be highly irregular. This does not appear to be a very satisfactory strategy.

We will not pursue the details of this further. It is obvious that to have inadequate turn bays can be very disruptive.

2.12. Time-dependent output flow

In all the theory discussed so far, it has been assumed that the cumulative output from a traffic signal could be approximated by a piecewise linear curve as in figure 1.6; and that the number of vehicles which could leave in an "effective" time G while the queue is discharging could be approximated by sG, for some appropriate saturation flow s and any time G sufficiently large for at least three or four vehicles to pass.

There is evidence to suggest, however, that the (average) output flow s may decrease with time after a while, possibly because slow moving vehicles do not keep up with the platoon. The further back these slow vehicles are in the queue when the signal turns green, the longer is the time it takes for them to reach the intersection and the larger is the gap between them and faster vehicles ahead of them. To verify this, however, one must carefully distinguish a decrease in flow while the queue is discharging from a decrease in average flow because the queue may empty during some signal cycles.

There are other situations in which the output flow will definitely decrease because the highway upstream of the intersection cannot maintain a flow s of through traffic. Suppose, for example, that the approach lanes upstream of the intersection can carry a flow s but there are separate turn bays for left and/or right turning vehicles. During the red signal, the queue for through traffic may back up beyond the entrance to the turn bays. The turning vehicles do not overflow the turn bays, but, after the signal turns green for the through traffic (and possibly also for the turning traffic), the highway upstream of the entrance to the turn bays can feed the through lanes plus the turn bays at a combined flow of at most s. Thus, after the signal has served

the through vehicles which were stored in the queue out to the entrance to the turn bays (at a rate s), the output flow may drop to s times the fraction of through vehicles in the approach stream. The signal cannot serve vehicles faster than they can pass the entrance to the turn bays (at an earlier time).

In either case, the question is: under what conditions should one terminate the green even before the queue vanishes, because the output flow has decreased to the point where one would do better to serve the cross traffic and come back in the next cycle with a fresh start at a higher output flow? The objective, presumably, would be to maximize the "capacity" of the intersection.

Suppose, for example, that one had a two-phase fixed-cycle signal serving traffic possibly in four directions but the cycle time and capacity are dictated by the through traffic in directions 1 and 2. For $i = 1$ or 2, let

$\bar{n}_i(t_i)$ = average number of (through) vehicles that leave the intersection in direction i in a time t_i after the start of the "effective" green, if there is a queue at time t_i.

We assume that the output rate at the time t_i

$$s_i(t_i) = d\bar{n}_i(t_i)/dt_i$$

is a decreasing (or nonincreasing) function of t_i.

Actually, it is more convenient here to deal with the inverse relation

is a decreasing (or nonincreasing) function of t_i.

Actually, it is more convenient here to deal with the inverse relation

$\bar{t}_i(n_i)$ = time required to serve an average of the n_i vehicles in direction i

(with $d\bar{t}_i(n_i)/dn_i = 1/s_i(\bar{t}_i(n_i))$ an increasing function of n_i), because the cycle time needed to serve an average of n_1 vehicles in direction 1 plus n_2 vehicles in direction 2 is

$$C = \bar{t}_1(n_1) + \bar{t}_2(n_2) + L$$

The average time per vehicle served in directions 1 and 2, the reciprocal of the combined service rate in directions 1 and 2 is

$$\text{average time per vehicle} = \frac{\bar{t}_1(n_1) + \bar{t}_2(n_2) + L}{n_1 + n_2} . \qquad (2.12.1)$$

We would presumably define the "capacity" as the maximum rate at which the intersection could serve vehicles (in directions 1 and 2) for a specified ratio of service rates in the two directions; namely, for $n_1/n_2 = q_1/q_2$ with q_1, q_2 the arrival rates in directions 1 and 2. Thus, the capacity is the reciprocal of the minimum of (2.12.1) with respect to n_1, for the given ratio n_1/n_2, i.e., the minimum of

$$\text{average time per vehicle} = \frac{\bar{t}_1(n_1) + \bar{t}_2(n_1 q_2/q_1) + L}{n_1(1 + q_2/q_1)} . \qquad (2.12.2)$$

One can formally minimize (2.12.2) by setting the derivative with respect to n_1 equal to zero. This leads to a (necessary) condition

$$\frac{q_1}{s_1(\bar{t}(n_1))} + \frac{q_2}{s_2(\bar{t}_2(n_1 q_2/q_1))} = \frac{q_1 C}{n_1} = 1 , \qquad (2.12.3)$$

with $q_1 = n_1/C$, $q_2 = n_2/C$ equal to the maximum arrival rates (capacity) for directions 1 and 2, respectively. If s_1 and s_2 were constant, this would be the familiar form for the capacity of an intersection

$$\frac{q_1}{s_1} + \frac{q_2}{s_2} = 1 - \frac{L}{C} \qquad (2.12.4)$$

would be the familiar form for the capacity of an intersection

$$\frac{q_1}{s_1} + \frac{q_2}{s_2} = 1 - \frac{L}{C} \qquad (2.12.4)$$

with a maximum capacity for $C = \infty$. In this more general form, however, the s_1 and s_2 are the service rates at the end of times t_1 and t_2 or after one has served n_1 and n_2 vehicles in directions 1 and 2, but with the t_1 and t_2 constrained so that $n_1/n_2 = q_1/q_2$. If the s_1 and s_2 are decreasing functions of t_1 and t_2, we would expect that (2.12.3) would be satisfied for some finite values of t_1, t_2, and C.

One can see the implications of (2.12.2) more easily from some graphical constructions. From a graph of $\bar{n}_i(t_i)$, one can obtain the graph of $\bar{t}_i(n_i)$ simply by switching the n and t axes. A graph of $\bar{t}_2(n_1 q_2/q_1)$ can be obtained from the graph $\bar{t}_2(n_2)$ by just changing the scale of the n . The numerator of (2.12.2) can be represented by adding the $\bar{t}_1$ and $\bar{t}_2$ graphs and displacing the coordinate by L .

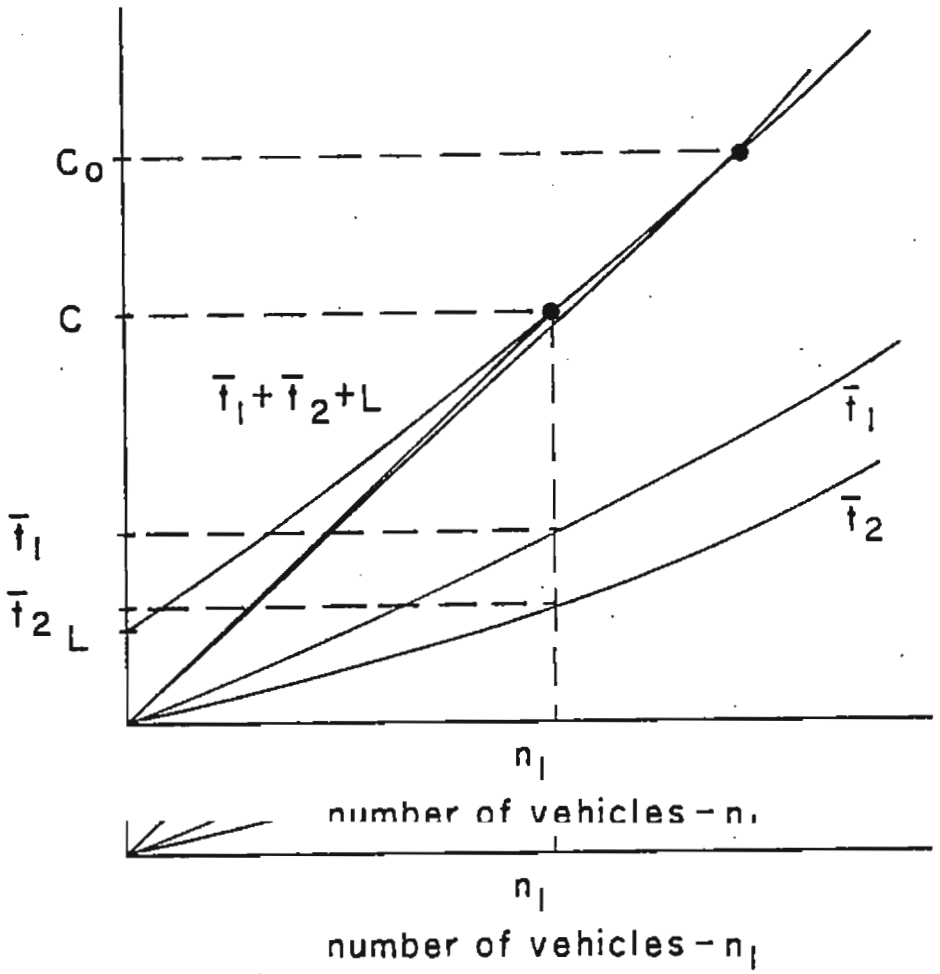


Fig. 2.13 - Graphical construction for the maximum capacity.

Figure 2.13 shows a hypothetical graph of $\bar{t}_1 + \bar{t}_2 + L$ vs. n_1 . For $n_1 = 0$, it has a value L and a slope $1/s_1(0) + q_2/q_1 s_2(0)$, but the slope increases with n_1 . Note now, that for any value of n_1 , the value of $(\bar{t}_1 + \bar{t}_2 + L)/n_1$, is the slope of the line from the origin to the point $(n_1 , \bar{t}_1 + \bar{t}_2 + L)$ on the graph. The value of $\bar{t}_1 + \bar{t}_2 + L$ is also the

cycle time C. If one also draws the graphs $\bar{t}_1(n_1)$ and $\bar{t}_2(n_1 q_2/q_1)$, one can read from these graphs the component splits $\bar{t}_1$ and $\bar{t}_2$ associated with any choice of n_1. The slope of the line is, of course, also the reciprocal of the flow q_1 which can be accommodated, and this flow multiplied by $1 + q_2/q_1$ gives the combined flow in both directions. Thus, for any choice of n_1, one can easily measure all quantities of interest from this figure.

From this graph one can readily see that as n_1 increases so does C, $\bar{t}_1$, and $\bar{t}_2$, but the slope of the line from the origin reaches a minimum when this line becomes the tangent line. The condition (2.12.3) is equivalent to the condition that the slope of the tangent line $1/q_1$ is equal to the slope of the curve $\bar{t}_1 + \bar{t}_2 + L$.

If, in figure 2.13, the cycle time C_0 which gives the maximum capacity is considerably larger than L (which one would expect to be true), there is another graphical construction which shows more clearly the interplay between the lost time L and the loss in time due to the increasing $s_i(t_i)$.

Suppose we define $s_1 = s_i(0)$ and write

$$\bar{t}_i(n_i) = \frac{n_i}{s_i} + \left[\bar{t}_1(n_i) - \frac{n_i}{s_i}\right] \qquad (2.12.5)$$

in which the second term represents the cumulative excess time needed to serve

$$\bar{t}_i(n_i) = \frac{n_i}{s_i} + \left[\bar{t}_1(n_i) - \frac{n_i}{s_i}\right] \qquad (2.12.5)$$

in which the second term represents the cumulative excess time needed to serve the n_i vehicles because $s_i(t_i) \geq s_i$. Of course, we expect that the second term will be much smaller than the first term. The n_1 in the denominator of (2.12.2) cancels the n_1 in the first term of (2.12.5) so that (2.12.2) can also be written as

$$\text{average time per vehicle} = \frac{1 + q_2 s_1/q_1 s_2}{s_1(1 + q_2/q_1)}$$

$$+ \frac{\left[\bar{t}_1(n_1) - \frac{n_1}{s_1}\right] + \left[\bar{t}_1(n_1 q_2/q_1) - \frac{n_1 q_2}{s_1 q_1}\right] + L}{n_1(1 + q_2/q_1)} . \qquad (2.12.6)$$

The first term of (2.12.6) is independent of n_1 and L. It represents the average time per vehicle with no losses, or the reciprocal of the "theoretical capacity" for $s_i(t) = s_i$. The second term of (2.12.6) has the same form as (2.12.2) itself except with the $\bar{t}_i(n_i)$ replaced by the $[\bar{t}_i(n_i) - n_i/s_i]$. The determination of the n_1 which gives the maximum capacity or the minimum average "time loss" per vehicle can now be done by the same construction as in figure 2.13 but with $\bar{t}_i(n_i)$ replaced by the $[\bar{t}_i(n_i) - n_i/s_i]$.

The advantage of this second scheme is that it, in effect, "magnifies" the deviation between the straight line and the curve in figure 2.13 so as to give better accuracy. The disadvantage is that it does not show directly the values of the $\bar{t}_1$, $\bar{t}_2$, C and the capacity; it shows the average time losses per vehicle.

As an illustration of the second scheme, suppose that (because of finite length turn bays in direction 1) the flow of through vehicles $s_1(t_1)$ has a

As an illustration of the second scheme, suppose that (because of finite length turn bays in direction 1) the flow of through vehicles $s_1(t_1)$ has a
value s_1 until n_1^* vehicles pass but it then drops to a value $s_1^* < s_1$ thereafter. The flow in direction 2, however, remains constant, $s_2(t_2) = s_2$. Thus

$$\bar{t}_1(n_1) - n_1/s_1 = \begin{cases} 0 & \text{for } n_1 < n_1^* \\ (n_1 - n_1^*)\left(\frac{1}{s_1^*} - \frac{1}{s_1}\right) & \text{for } n_1 > n_1^* \end{cases}$$

and

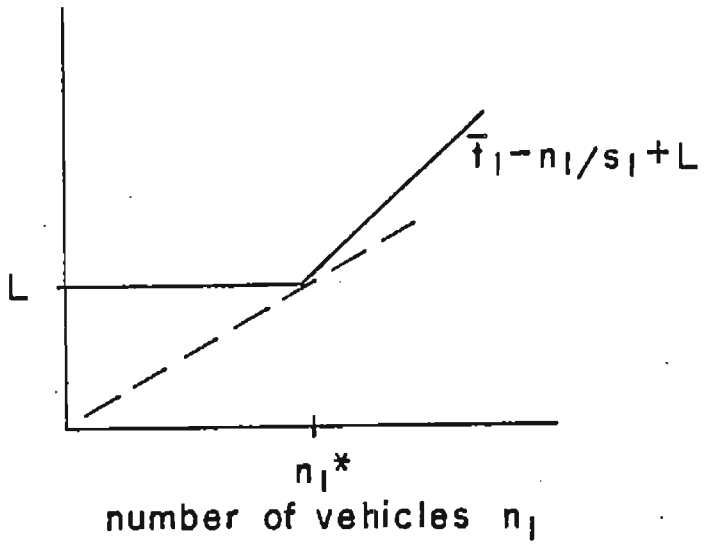


Fig. 2.14 - Minimizing the average time loss per vehicle.

The numerator of the second term of (2.12.6) would be as shown in figure 2.14 and the average loss in time per vehicle is proportional to the slope of a line from a point on this curve to the origin. This slope will be a minimum either for $n_1 = n_1^*$ or for $n_1 = \infty$ accordingly as the slope of the curve for $n_1 > n_1^*$ is larger or smaller than the slope L/n_1^* of the broken line of figure 2.14. Thus, we should switch the signal after n_1^* vehicles have passed if

$$\frac{1}{s_1^*} - \frac{1}{s_1} > \frac{L}{n_1^*} \quad \text{or} \quad \frac{s_1}{s_1^*} - 1 > \frac{s_1 L}{n_1^*} = \frac{L}{t_1^*} , \qquad (2.12.7)$$

$$\frac{1}{s_1^*} - \frac{1}{s_1} > \frac{L}{n_1^*} \quad \text{or} \quad \frac{s_1}{s_1^*} - 1 > \frac{s_1 L}{n_1^*} = \frac{L}{t_1^*} , \qquad (2.12.7)$$

i.e., if the fractional change in $s_1(t_1)$, $s_1/s_1^* - 1$ is larger than the ratio of the lost time L to the green interval t_1^* for direction 1.

This illustration shows that one would typically need a fairly substantial drop in flow to justify switching the signal after n_1^* vehicles (for $L \sim 10$ seconds, t_1^* perhaps 30 seconds, one would need $s_1^*/s < 3/4$). One would be more likely to switch the signal, however, if there were also a drop in flow for direction 2.

The most important aspect of this is not so much the question of whether or not one should switch the signal, but that the reduced flow causes a reduction in capacity no matter what one does. Traffic engineers customarily worry about whether turn bays are large enough to store the turning vehicles, they are less apt to consider the possibility that the turning traffic may limit the through traffic if the queue of through traffic extends beyond the entrance to the turn bays.

It would be rather tedious to generalize the theory in section 2.3 where we tried to choose a cycle time so as to minimize the sum of deterministic and stochastic queueing effects. Suffice it to say, that choosing a cycle time larger than the C_0 which maximizes the capacity would increase both the deterministic and the stochastic queueing. Perhaps a reasonable recipe would be to choose a cycle time as described in section 2.3 provided that it is less than C_0, otherwise choose the cycle times and splits as in figure 2.13 to maximize the capacity. (Actually the "optimal" cycle times and splits would be shorter than this).

If, for a V-A signal, the decrease in $s_i(t_i)$ with t_i is due to some restriction upstream of the signal, one could modify the usual V-A signal strategy by imposing maximum green intervals equal to those which maximize the capacity (with the appropriate lost time L for the V-A signal). To strategy by imposing maximum green intervals equal to those which maximize the capacity (with the appropriate lost time L for the V-A signal). To allow green intervals larger than these would negate the benefits one hoped to achieve by allowing the cycle time to drift to larger values. It is conceivable that the mechanism for detecting the "end of the queue" by observing individual headways in the traffic stream could be set so that the signal would automatically switch when the flow dropped below the level specified in figure 2.13, but the observation of individual headways is not a very good measure of the flow.

There is no satisfactory theory to describe how one should observe and

Even if one did have such a theory, it would be difficult to implement. Most existing equipment cannot distinguish a large headway caused by a slow vehicle from a large headway following the "end of the platoon." In the former case one expects that the flow will return (temporarily) to a value of approximately s_i after the slow vehicle passes, whereas in the latter case one expects the long headway to be followed by a flow of approximately q_i. To distinguish between these, one would obviously need more detectors upstream of the detector which is recording the gaps. The strategy one should use in the latter case has already been discussed in sections 2.6 and 2.7.

If one had many detectors at various locations upstream of the signal, one might be able to estimate, at any time t, the values of several successive headways after time t (with decreasing accuracy). In effect, one could estimate, say for direction 1, the actual values of the $t_1(n_1)$ or the $t_1 - n_1/s_1$ rather than their average values.

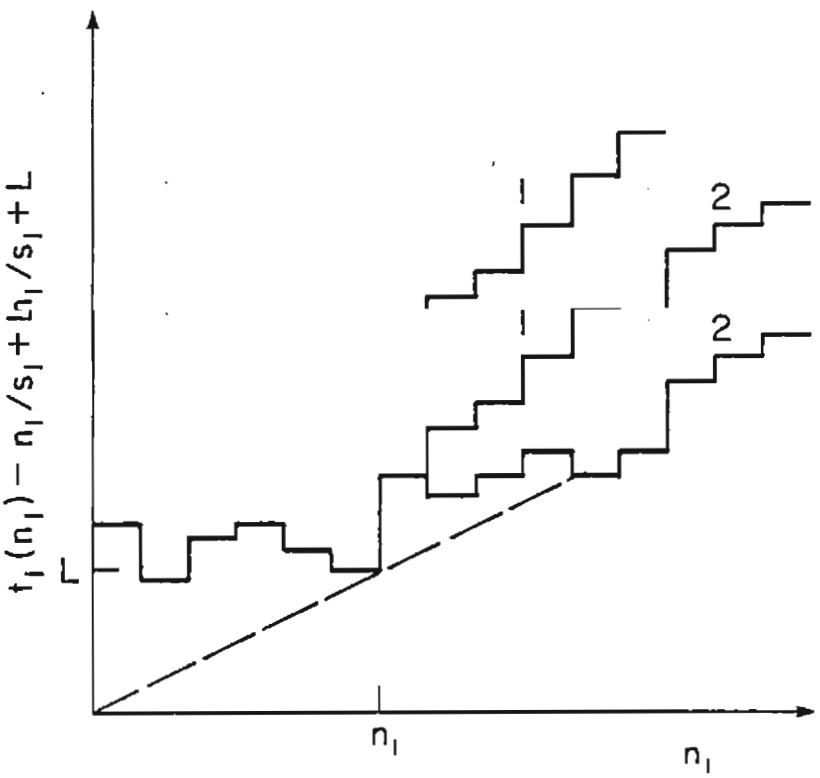


Fig. 2.15 - Observed cumulative time.

One can see what is at issue here by imagining that one can draw a graph of $t_1 - n_1/s_1 + L$ vs. n_1 as in figure 2.15 for integer values of n_1. For the "normal" saturation flow one expects the actual values of $t_1(n_1) - n_1/s_1 + L$ to fluctuate around the value L. Suppose, however, for some particular value of n_1, we observe a headway appreciably larger than $1/s_1$. If this indicates the end of the queue, the subsequent curve of $t_1(n_1) - n_1/s_1 + L$ would grow at an average rate of $(1/q_1 - 1/s_1)$ per vehicle, as illustrated by curve 1 of figure 2.15. A properly set V-A signal would presumably switch just at the start of the long headway, absorbing the headway in the lost time, and achieving an average lost time per vehicle corresponding to the slope of the broken line of figure 2.15. If the subsequent curve of $t_1(n_1) - n_1/s_1 + L$ stays above this broken line, then one should indeed switch the signal when the n_1th vehicle passes.

If the (future) curve will look like curve 2 of figure 2.15, i.e., the flow returns to a value of approximately s_1 and it stays there long enough to intersect the broken line, then one should typically continue the green so as to achieve a smaller average time loss per vehicle. Unfortunately, one cannot generally obtain estimates of the future headways accurately enough to predict whether or not the curve will cross the broken line. Certainly there is insufficient data available now to justify trying to develop any theory

predict whether or not the curve will cross the broken line. Certainly there is insufficient data available now to justify trying to develop any theory

which might be useful here.

REFERENCES

1. McClintock, Miller, Street Traffic Control, McGraw Hill, New York, 1925.

2. Matson, T. M. "The principles of traffic signal timing." Transactions of the National Safety Council III, Chicago, 1929.

3. Gerlough, D. L., and Huber, M. J., Traffic Flow Theory. Transportation Research Board Special Report 165, 1975.

4. Edie, L. C., "Flow Theories," in Traffic Science, D. C. Gazis, editor, Wiley, 1974.

5. Preparata, F. P., "Analysis of traffic flow on a signalized one-way artery," Trans. Sci., 6, 32-51 (1972).

6. Wardrop, J. G., "Some theoretical aspects of road traffic research," Proc., Instn. Civil Engineers 1, 325-62 (1952).

7. Webster, F. V., Traffic Signal Settings, Road Research Technical Paper 39 H. M. Stationery Office, London, 1961.

8. Newell, G. F., Applications of Queueing Theory, 2nd ed., Chapman and Hall, London, 1982.

9. Newell, G. F., "Approximate method for queues with application to the fixed-cycle traffic light," SIAM Rev. 7, 223-240 (1965).

10. Webster, F. V., and Cobbe, B. M., Traffic Signals, Road Research Technical Paper #56, H. M. Stationery Office, London, 1966.

11. U. S. Department of Transportation, Federal Highway Administration, Traffic Control Devices Handbook, 1983.

12. U. S. Department of Transportation, Federal Highway Administration. Traffic Control Systems Handbook, 1985.

13. Tanner, J. C., "Problem in the interference of two queues," Biometrica 40, 58-69 (1953).

Control Systems Handbook, 1985.

13. Tanner, J. C., "Problem in the interference of two queues," Biometrica 40, 58-69 (1953).

14. Darroch, J. N.; Newell, G. F.; and Morris, R. W. J., "Queues for a vehicle-actuated traffic light," Operations Res. 12, 882-895 (1964).

15. Newell, G. F., "Properties of vehicle-actuated signals I One-way streets," Trans. Sci., 3, 30-52 (1969).

16. Newell, G. F., and Osuna, E. E., "Properties of vehicle-actuated signals II Two-way streets," Trans. Sci. 99-125 (1969).

3. COORDINATION ON A ONE-WAY ARTERIAL

3.1. Introduction

Given the number and complexity of issues relating to the design of a single isolated signal as described in Chapter 2, one might think that any extension of this theory to a sequence of signals along an arterial or to a network would be extremely complex. The number of potentially important parameters could be very large. It might include not only the p_i , q_i , s_i, L, etc., at each intersection but trip times between intersections and vehicle interactions. From this, one must determine cycle times and splits (possibly with turn phases) for each intersection, and "off-sets," i.e., the time displacement of one signal relative to others. Actually most of the issues relating to signal coordination are quite different from those of isolated signals, and one should not think of the theory of coordinated signals as a "generalization" of that for single intersections.

Although one might use information from vehicle detectors to make local accommodations, an "optimal" strategy at one signal for serving a pulsed traffic stream from a neighboring signal is to switch the signal more or less in phase with the pulses. It would seem that the overall pattern of signal switches should be (nearly) periodic in time. This does not mean that all signals should necessarily operate on the same "cycle time" as defined for the isolated signals. It is often advantageous for some signals to run through two or three sequences

necessarily operate on the same "cycle time" as defined for the isolated signals.

It is often advantageous for some signals to run through two or three sequences of phases in each cycle of some other signal, i.e., operate on a half or third the cycle time of the latter. Certainly one would not choose the cycle time of one signal independent of the others or allow a V-A signal to drift at will in response to local conditions.

The first step in the analysis of any network of signals is to determine whether or not any intersection will, necessarily, be oversaturated regardless of what signal strategy one uses, for a specified routing of vehicles. If some signals are oversaturated, one should then see if it is possible to accommodate

the demand by rerouting the traffic (particularly the turning vehicles at critical intersections). If nothing succeeds, then the main issue is where one can store the excess vehicles which cannot be accommodated, so as to cause the least damage to others (blocking, "grid lock," etc.), how one can use the signals at noncritical intersections to keep the vehicles there, and how one should meter the flows into the critical intersections. The ideal strategy would be to meter the traffic from all parking lots and other sources (where there is certainly storage space), but this is not usually possible.

Whether or not any individual intersection is oversaturated is essentially independent of any strategy of off-sets or cycle times of neighboring signals. If a signal is undersaturated, one can measure the (cumulative) number of vehicles to arrive (or leave) during several consecutive cycles (perhaps 10 to 15 minutes) and divide this by the period of observation so as to obtain an (average) flow q_i or $q_{\ell i}$. One may (if necessary) also average this over corresponding observations on "similar" days. The assumption here is that such observations of the q_i will vary only slowly with the time of day (on a scale of 10 or 15 minutes, perhaps). They represent approximately the (nearly uniform) rate at which travelers would arrive at the intersection if there were no other signals.

Most of the discussion in Chapter 2 regarding capacity (but not that regarding the stops and/or delays) applies here as well, but with the q_i or $q_{\ell i}$ interpreted as above. These capacity constraints simply require that the (average) number of arrivals of any type per cycle shall be less than the corresponding maximum number which can be served. For example, at the intersection of one-way streets with flows $q_i^{(m)}$ at the mth intersection in direction i, the constraint on the cycle time is

$$C^{(m)} \geq \frac{L^{(m)}}{1 - q_1^{(m)}/s_1^{(m)} - q_2^{(m)}/s_2^{(m)}} \tag{3.1.1}$$

in which $L^{(m)}$ is the appropriate lost time per cycle at the mth intersection.

The same formula would apply also for two-way streets with no turning traffic if directions 1 and 2 are labeled so that $q_1^{(m)}/s_1^{(m)} > q_3^{(m)}/s_3^{(m)}$ and $q_2^{(m)}/s_2^{(m)} > q_4^{(m)}/s_4^{(m)}$. It also applies for intersections with left-turn bays (but with the $q_i^{(m)}$ replaced by $q_{ti}^{(m)}$) provided the turning traffic does not overflow the turn bays and block the through traffic. In this case, however, there will also be upper bounds on $C^{(m)}$, namely

$$q_{\ell i}^{(m)} C^{(m)} < 2 \tag{3.1.2}$$

or whatever average number of left turns vehicles can be accommodated per cycle in direction i. For multiphase signals, there are more complex constraints, as described in section 2.11.

For any admissible $C^{(m)}$ it is necessary also that the cycle time be split so that no through direction is oversaturated. Specifically, the green interval for direction i, $G_i^{(m)}$ must satisfy

$$s_i^{(m)} G_i^{(m)} > q_i^{(m)} C^{(m)} \tag{3.1.3}$$

for all i and m.

Although the "optimal" settings of signals in a network could be a function of all the $q_i^{(m)}$, $s_i^{(m)}$ and $L^{(m)}$, most of the dependence on these parameters

Although the "optimal" settings of signals in a network could be a function of all the $q_i^{(m)}$, $s_i^{(m)}$ and $L^{(m)}$, most of the dependence on these parameters

enters through equations of the above type which guarantee that each intersection is undersaturated. These conditions do not depend on the off-sets or any issues regarding delays, stops, etc.

Much of the following analysis will deal with the choice of off-sets (and the choice among admissible cycle times) so as to minimize delays, stops, etc. The models discussed in chapters 1 and 2 will be used to relate the departure times at the mth intersection to the "expected" arrival times at the mth

intersection and to evaluate delays and stops at the mth intersection. To complete the picture, however, one must also have some models to relate the expected arrival times of vehicles at the mth intersection to their departure times at neighboring intersections.

Obviously the choice of off-sets must be related to some "average" trip time between intersections but, fortunately, these choices are not very sensitive to the details of the vehicle trajectories, for two reasons. First, the total "delay" is actually the total trip time of all vehicles measured relative to some hypothetical motion with "zero delay." How one defines the latter is actually irrelevant. In particular, it is not necessary that one have a precise definition of "expected arrival time" as long as one can identify which vehicles are delayed, for any choice of offsets. Secondly, as pointed out in Chapter 1, the times at which stopped vehicles leave an intersection relative to the start of green are (nearly) independent of when they arrived.

To illustrate some issues as simply as possible, we first imagine some hypothetical vehicle motion in which each vehicle is either stopped or is traveling at a some specified speed v. If two intersections are a distance d apart, the nonstop trip time of the vehicle between these intersections is, therefore, $\tau = d/v$.

Obviously one should recognize that the trip time is an increasing function of d, but the geometry of the network is presumably given and the distances cannot be changed. One could calculate the trip time from an assumed velocity v and distance d or one could simply measure the time directly. The key postulate here is not that the τ increases linearly with d, but that the trip time is the same for all vehicles, independent of their positions in a platoon, or if they are stopped at some intersection, or even who the drivers are.

After we have analyzed some of the implications of such a model, we will

then investigate what sort of modifications one must make to correct for the fact that the trip times are not the same for all drivers.

We will first consider, in some details, the coordination of signals along a one-way arterial for various models of vehicle motion. We will then consider the coordination on two-way arterials without turning traffic and with turning traffic.

3.2 Pretimed signals, no platoon spreading

a) Simple progresson, no turning traffic

If all vehicles travel at some design speed v when they are not stopped, the standard scheme for coordination of signals on a one-way arterial is as illustrated in figure 3.1. All signals operate on the same cycle time C and

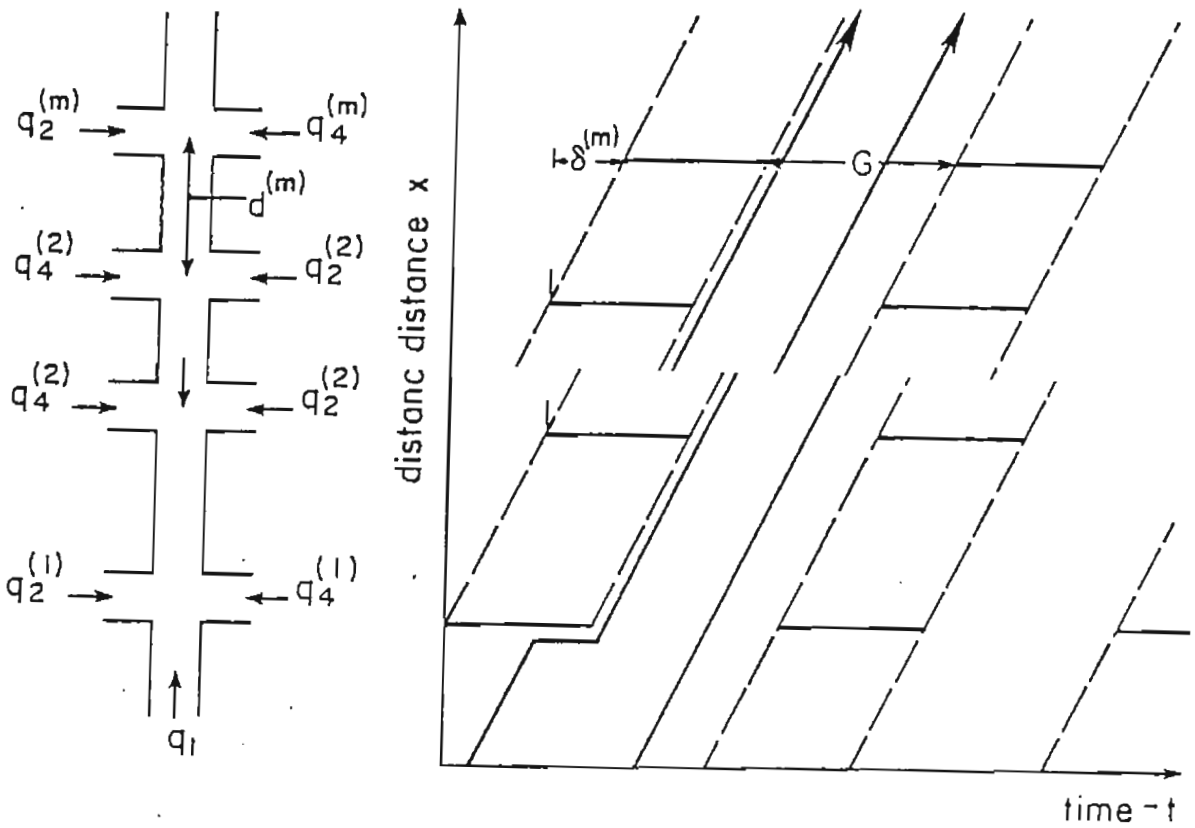


Fig. 3.1 - An idealized signal progression.

have the same green interval G for the arterial direction. If here is no turning traffic, the flow $q_1^{(m)} = q_1$ along the arterial will be the same at all intersections. We will also assume that $s_1^{(m)} = s_1$ is the same at all intersections. The cross-traffic may be either one-way or two-way, but if it is the latter, we will label directions 2 and 4 so that $q_2^{(m)}/s_2^{(m)} > q_4^{(m)}/s_4^{(m)}$ (not necessarily the same directions at every intersection).

The cycle time and the green interval should (if possible) be chosen so that all intersections are undersaturated, i.e.,

$$G = G^{(m)} > (q_1^{(m)}/s_1^{(m)})C \text{ and } (C - G - L^{(m)}) > (q_2^{(m)}/s_2^{(m)})C \qquad (3.2.1)$$

for all m, which together also imply that

$$C > \frac{L^{(m)}}{1 - q_1^{(m)}/s_1^{(m)} - q_2^{(m)}/s_2^{(m)}} \text{ for all } m,$$

or

$$C > \max_m \left\{ \frac{L^{(m)}}{1 - q_1^{(m)}/s_1^{(m)} - q_2^{(m)}/s_2^{(m)}} \right\}. \qquad (3.2.2)$$

The off-sets are chosen so that the start of green at the mth intersection occurs at a time $\delta^{(m)}$ after that for the $(m-1)$th intersection with $\delta^{(m)} = \tau^{(m)} = d^{(m)}/v$ equal to the trip time at velocity v between the two intersections at spacing $d^{(m)}$.

$\delta^{(m)} = \tau^{(m)} = d^{(m)}/v$ equal to the trip time at velocity v between the two intersections at spacing $d^{(m)}$.

As one can see from the time-space diagram of figure 3.1, with this signal coordination any vehicle traveling on the arterial at speed v which passes the first signal at the start (end) of the green, will pass every other signal at the start (end) of the green.

There are a number of interesting features of this scheme which one should recognize immediately from figure 3.1, some of which may be consequences of an overly idealized picture.

A vehicle approaching intersection 1 on the arterial may suffer some delay at this intersection, but it should not be stopped or delayed at any subsequent intersections. Any idealized trajectory with (exactly) constant velocity v may depend on the setting of the first signal but is unaffected by the existence of any other signals. One can insert other signals as close as one pleases or take them out.

With no turning traffic, vehicles on the arterial will pass any point x only during a time of duration G, but the arterial will be clear for any cross traffic during an (effective) time $C - G - L^{(m)}$, regardless of whether or not there is a signal at x. For sufficiently light traffic on both the arterial and the cross street, the delay to the cross street would likely be less without a signal because a vehicle on the cross street may find gaps in the traffic during the arterial flow. For heavy traffic on the arterial and/or the cross street, however, a traffic signal would have an advantage over a stop or a yield sign on the cross street because the saturation flow $s_2^{(m)}$ during a cross street green is considerably larger than at a stop sign (during the time when there is no traffic on the arterial).

b) Coordinate transformations

Before we proceed to consider other possible strategies and issues, it is convenient to develop some other graphical representations of trajectories

Before we proceed to consider other possible strategies and issues, it is convenient to develop some other graphical representations of trajectories and signals.

The key assumption made above about the vehicle motion is that all vehicles have the same trip time $\tau^{(m)}$ between intersections $m - 1$ and m. Whether or not the actual trajectories between intersections have a constant velocity between intersections is irrelevant. For any choice of signal settings, the assumption of a constant velocity is used only to provide a simple graphical scheme for relating the arrival times at one intersection to the departure times

of a previous intersection. Otherwise, the vertical scale in the t, x graph of figure 3.1 is quite arbitrary.

Generally, it would be more convenient to measure the position x along the arterial by the uninterrupted trip time from some reference location (intersection 1, for example). If all vehicles did travel with a constant velocity, we would measure "distance" as x/v. A graph such as figure 3.1 would then be rescaled so as to show x/v vs. t. If x/v and t were drawn on the same scale, a vehicle would travel a unit "distance" x/v in unit time; an idealized trajectory would have slope 1. But even if v were not independent of location and could be described as some function $v(x)$, a graph of the uninterrupted trip time $\int_0^x dx'/v(x')$ as a description of location vs. the actual time t would yield vehicle trajectories with slope 1, independent of the $v(x)$. The same would be true if we simply chose the "distance" between intersections as the $\tau^{(m)}$ and said nothing about the detailed motion between intersections.

To illustrate further the fact that only the differences between (expected) vehicle arrival times and departure times from individual intersections (which in the present scheme are all zero) are important, we might also note that observers at different locations could have clocks with different time origins. Suppose, for example, that an observer at $x = 0$ starts his clock with the passage of some vehicle. An observer at x would start his clock at the time when the vehicle would pass him if it were not interrupted, i.e., at time x/v passage of some vehicle. An observer at x would start his clock at the time when the vehicle would pass him if it were not interrupted, i.e., at time x/v if v is constant, or, more generally, at time $\Sigma_{\ell=1}^{m}\tau^{(\ell)}$ at the mth intersection.

Each observer now records all events, signal switches or passing vehicles, relative to his clock which measures the time $t(x) = t(0) - x/v$. If one now draws a vehicle trajectory with the coordinates x or x/v and $t(x)$, an uninterrupted trajectory would appear as a vertical line as in figure 3.2.

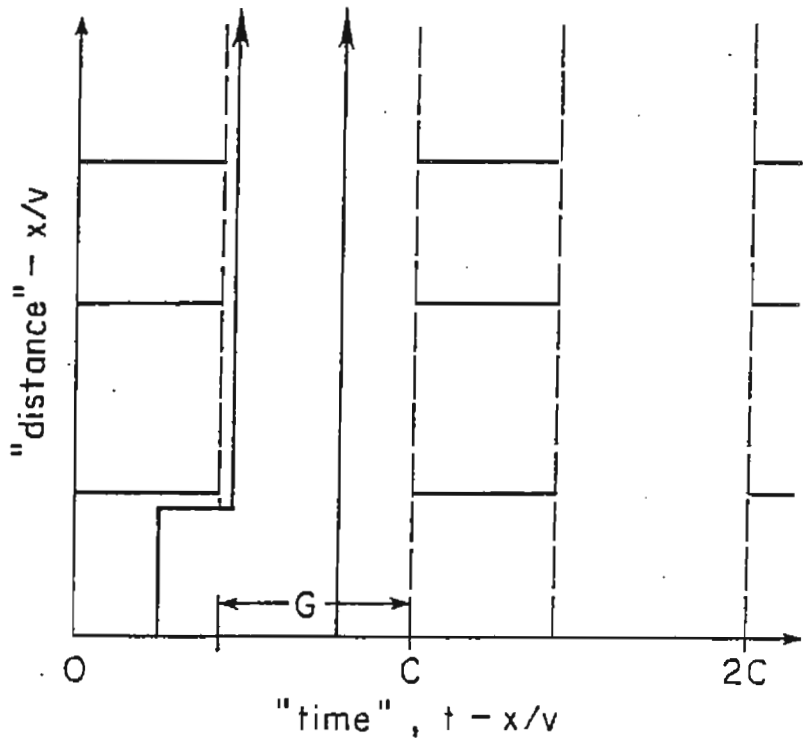


Fig. 3.2 - A transformed signal progression.

For any specified cycle time C, it may also be advantageous to measure times in units of C , i.e., draw graphs of x/vC vs. t/C or t/C - x/vC .

c) Progression with unequal splits, no turning traffic

Suppose that for the above arterial system we choose some common cycle time C , and arterial green intervals $G^{(m)}$ satisfying (3.2.1) and (3.2.2). For some reason, however, we choose the $G^{(m)}$ unequal. Perhaps there are some (unequal) pedestrian constraints at each intersection, or perhaps there is some minor cross street (particularly m = 1) which requires only a short green interval and, to minimize delays at this intersection, we propose to give some extra green time to the arterial.

One of the common procedures for choosing off-sets (particularly for two-way traffic) is to maximize the (time) width of the "through band." The through band consists of the family of all (perhaps hypothetical) trajectories such that,

if all vehicles are injected into the first intersection at times within the band, they will traverse the system with no delay. In Figures 3.1 and 3.2 the width of the through band is the common green interval G. For systems with unequal $G^{(m)}$, the maximum bandwidth is

$$b = \min_m G^{(m)} \tag{3.2.3}$$

and the through band is constructed by drawing through this minimum green interval the trajectories which would exist if there were no other signals. Since every intersection has a green interval at least b, it is possible to select the offsets at every intersection so that the through band passes through every green interval. The off-sets are obviously not unique.

The strategies which minimize the total delay are not unique. They include all those which maximize the bandwidth and others, but, among these, the strategy which also minimizes the number of stops does <u>not</u> maximize the bandwidth.

To analyze arterial signal coordination, it is advantageous first to identify the most critical intersections (in this case those for which $G^{(m)} = b$) and then see what would happen if there were no other signals (or associated cross traffic) in the system. Certainly the addition of more signals (and cross traffic) cannot decrease the total delay or the number of stops on the arterial, so the optimal strategy should be that which minimizes the penalties for the addition of more signals.

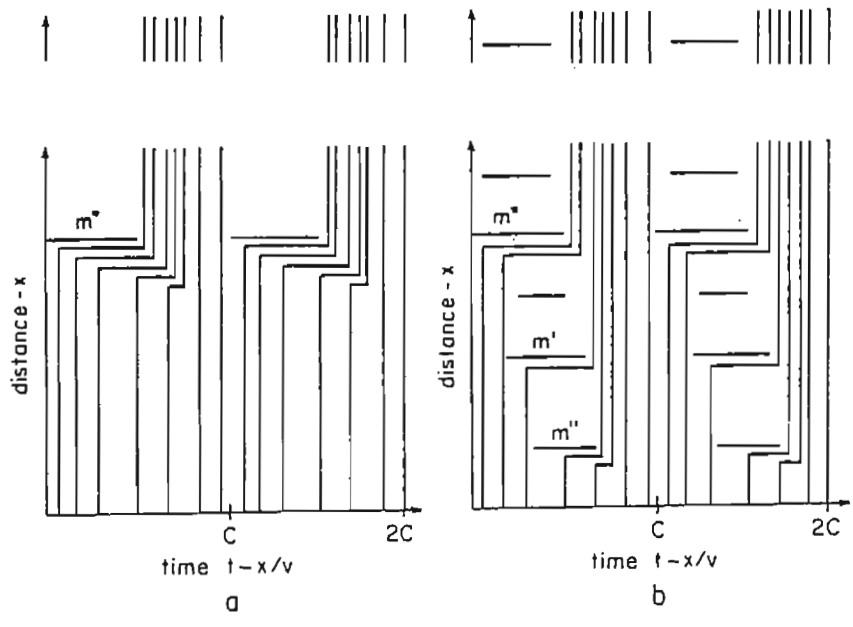


Fig. 3.3 - Signal coordination with unequal green times.

Figure 3.3a shows a schematic picture of some possible trajectories in an x vs. t - x/v graph with all signals having $G^{(m)} > b$ eliminated. These vehicles must have entered the system at some times not necessarily in the through band and, in the absence of some signals (particularly $m = 1$), these vehicles would not be stopped until they reach the first critical intersection m^*. Those which are stopped presumably leave at a flow $s_1^{(m^*)}$ from the start of green until the queue vanishes. For simplicity, let's assume that the $s_1^{(m)}$ are all equal, $s_1^{(m)} = s_1$, so that vehicles leaving m^* will not be delayed again because of some smaller $s_1^{(m)}$ downstream.

From the arguments in section 1.5, we now claim that the introduction of other signals will have no effect on the total delay (travel time) on the arterial provided that each vehicle in figure 3.3 arrives at m^* prior to its departure in figure 3.3a and the new signals do not interrupt the throughband downstream of m^*. If a vehicle is certain to be delayed, it makes no difference when it is delayed. Furthermore (with $s_1^{(m)} = s_1$ for all m), it is always possible to choose the off-sets for any signals with $G^{(m)} > b$ so that each vehicle is stopped at most once. Thus, it is possible to minimize both the total delay and the number of stops on the arterial simultaneously.

This "optimal" strategy is illustrated in figure 3.3b. (It is possible to derive some formulas for the off-sets $\delta^{(m)}$, particularly for a constant input flow, but it is typically easier to construct graphically). If the to derive some formulas for the off-sets $\delta^{(m)}$, particularly for a constant input flow, but it is typically easier to construct graphically). If the input flow is time-dependent due to some signals operating on a cycle time C which are outside our system, we will first choose the start of green in figures 3.3a so as to be "optimal" according to any criteria, and insert any traffic signals downstream of m^* so as not to interrupt the through band. For given values of the $G^{(m)}$, we will next identify the signal m' with the smallest $G^{(m)}$, $m < m^*$ and suppose, for now, that we eliminate all other signals.

Since $G^{(m')} > b$, some vehicles which pass m' without delay will be stopped at m* (they would have been stopped even without the signal at m'). The objective now is to choose the off-sets so that a vehicle which is stopped at m' is not stopped again at m* .

We could start a "trial solution" by switching the signals to green "simultaneously" in figure 3.3b (actually with an off-set equal to the trip time from m' to m*). Vehicles leaving m' will then arrive at m* just as the signal turns green but would join a queue of those vehicles which were stopped at m* for the first time. Now proceed to delay the start of green at m' (which will, generally, increase the number of vehicles which slip past m' but are stopped at m*) until the vehicles leaving m' arrive at m* , at flow s_1 , just in time to keep the signal at m* busy when the queue vanishes.

We can now proceed iteratively. Insert any signals between m' and m* so as not to interrupt any traffic between m' and m* and identify the signal m" with the smallest $G^{(m)}$, $m < m'$. We can disregard, for now, any other signals and choose the offset of the signal at m" so that the vehicles stopped at m" arrive at m' just as the queue vanishes. Continue now upstream until one reaches the entrance.

We have now seen that, for given $G^{(m)}$ and C , it is possible to choose the off-sets of all signals in such a way that the total delay and number of stops on the arterial is the same as if there were no signals except at the the off-sets of all signals in such a way that the total delay and number of stops on the arterial is the same as if there were no signals except at the intersection m* with $G^{(m*)} = \min_m G^{(m)}$. If there were no other intersections, the delays and stops would all occur at m* , but if the $G^{(m)} = G$ were all equal (or m* = 1), they would occur at the entrance. Otherwise, the queues would be distributed over the signals with $m \leq m^*$.

For heavy traffic and large values of C , there may be some advantage to distributing the queues over several intersections so as to prevent queues from blocking intersections. Otherwise, there seems to be no reason why one

should give any excess green time $G^{(m)} - b$ to the arterial. Unless an intersection is also part of a signal progression in the cross direction (which has no need for the excess time either), assigning the excess green time to the cross direction would typically reduce delays and stops for the cross direction at no expense to the arterial traffic. Thus (in this situation), one should choose $G^{(m)} = b$ for all m.

d. Turning traffic, local optimal

We have seen that, with no turning traffic on a one-way arterial, the delays and stops on the arterial could be reduced to the point that they depend only on how vehicles enter the system (at the "first" intersection). For sufficiently long arterials, however, a significant fraction of the flow on the arterial is likely to enter (and leave) via the cross streets even though there may be only a small fraction of turning traffic at any single intersection.

Traffic leaving the (one-way) arterial during a green interval should not seriously affect the saturation flows $s_1^{(m)}$ unless there is interference from pedestrians. With turning traffic, however, the flow $q_1^{(m)}$ approaching the mth intersection (interpreted as in section 3.1) will vary with m. If the signals are to be undersaturated, the $G^{(m)}$ must satisfy (3.1.3). We will assume (for now), however, that the $s_1^{(m)}$, $G^{(m)}$, and $C^{(m)}$ are the same at every intersection and

assume (for now), however, that the $s_1^{(m)}$, $G^{(m)}$, and $C^{(m)}$ are the same at every intersection and

$$s_1 G > q_1^{(m)} C \quad \text{for all} \quad m .$$

If some vehicles on the arterial can turn on red without affecting the output rate s_1 during the subsequent green, they can be eliminated from the flow $q_1^{(m)}$ above, so that the $q_1^{(m)}$ includes only those vehicles which must be served during the green interval.

Although the flow leaving the $(m-1)$th intersection during the green may have the value s_1, some of this flow may exit the arterial at the $(m-1)$th intersection. Other vehicles may also enter the arterial from the cross streets during the arterial red (or even during the green). Suppose that the (average) cumulative curve of expected arrivals (from all sources) approaching the mth intersection on the arterial has a periodic behavior as illustrated in figure 3.4 with a flow $q_1^{(m)}(t)$ such that $q_1^{(m)}(t) \leq s_1$ for all t.

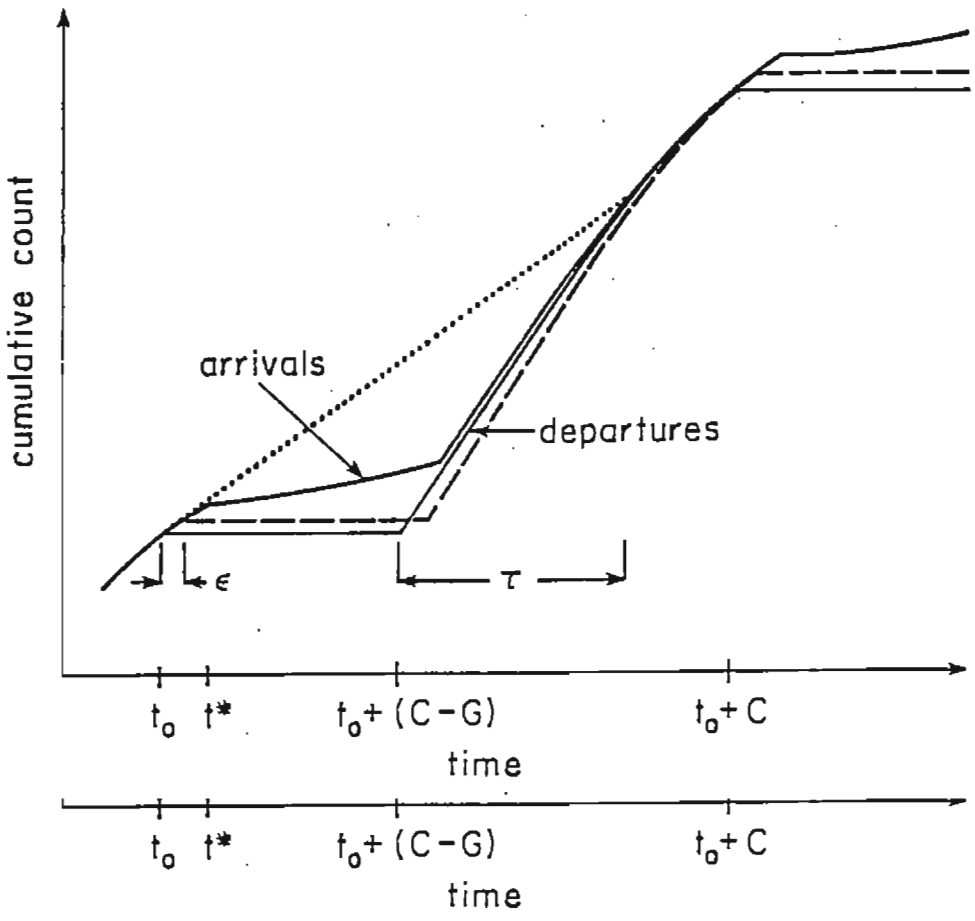


Fig. 3.4 - Cumulative arrivals and departures from a signal.

For any choice of off-sets, the departure curve from the mth intersection (including traffic which will leave the arterial during the green interval at the mth intersection) is assumed to have a flow s_1 during the green interval whenever there is a queue and otherwise to give zero delay.

Suppose that we propose some off-set such that the signal turns red at time t_0 but we then delay the start of the red by an arbitrarily small time

mth intersection, so that any change in the off-sets will merely displace the "time origin" on the mth cross street and not affect the (average) delays to the cross street. One should, however, worry about possible consequences of changes in the off-sets at the mth intersection on the arterial traffic delays downstream.

Condition (3.2.5) does not itself guarantee a minimum delay (only a stationary value), but for typical arrival curves there will usually be only one solution. If there is more than one, it should not be difficult to identify which one gives the minimum. The formula does not give an "explicit" solution for the time t_0, but one can easily determine it graphically by drawing some tangent lines to the arrival curve at various proposed values of t_0 and observing which tangent line intersects the arrival and departure curves at time $t_0 + (C - G) + \tau$. The formal solution, however, is not as important as the qualitative issues and typical conclusions.

For small fractions of turning traffic at each intersection, the typical shape of the arrival curve is as shown in figure 3.4. There is a (near) discontinuity in slope at a time t^* generated by the termination of green at the $(m - 1)$th intersection. From time t^* to $t^* + (C - G)$ the arrival curve will increase slightly due to vehicles turning into the arterial from the $(m - 1)$th cross street. The detailed shape of the curve here will depend on the signal strategy and traffic pattern for turning traffic from the cross street, but there may be a queue on the arterial at the start of the arterial green. After time $t^* + (C - G)$ the arrival flow would have a value s_1 (for our idealized model with equal trip times for all vehicles and piecewise linear cumulative departure curves) until the flow drops due to the termination of the queue at the $(m - 1)$th intersection, provided no vehicles exit the arterial at the $(m - 1)$th intersection. For a small fraction of turning vehicles, however, the flow during this time period will be slightly less than s_1. For

later times we expect the flow $q_1^{(m)}(t)$ to be decreasing with t, having values dependent on strategies of control upstream of the $(m-1)$th intersection.

Because of the discontinuity in $q_1^{(m)}(t)$ at $t = t^*$ from $q_1^{(m)}(t^*)$ nearly to 0, the optimal choice of t_0 is quite likely to be at $t_0 = t^*$, i.e., the off-sets will be equal to the transit time, despite the possible existence of a queue at the start of green. To test this, suppose, as illustrated in figure 3.5a, we draw the departure curve for $t_0 = t^*$ and locate the time $t^* + (C - G) + \tau$ at which the queue vanishes. Then draw a (dotted) line between the intersection of the arrival and departure curves at time t^* and $t^* + (C - G) + \tau$. If the slope of this line is less than the slope $q_1^{(m)}(t^*)$ of the arrival curve just prior to the time t^*, it is <u>not</u> advantageous to advance the off-sets.

It is interesting to note here that if the arrivals were uniform at the first intersection of figure 3.1, the tangent line to the arrival curve and the actual arrival curve would coincide, as illustrated in figure 3.5b. The condition (3.2.5) is satisfied for every choice of t_0. This is consistent with the obvious fact that the delays at this intersection are independent of when the signal cycle starts. It follows, also, that, if the mth intersection is undersaturated, the time τ in figure 3.5a must be less than G, and the slope of the tangent line in figure 3.5a must be less than the (average) flow $q_1^{(m)}$. Thus, if the arrival flow $q_1^{(m)}(t)$ is larger than $q_1^{(m)}$ for all t during the green interval, but particularly near the end of the green, then the optimal off-set is $t_0 = t^*$ regardless of any other features of the arrival curve or the size of the queue at the start of green.

The strategy in figure 3.5a is contrary to the "conventional wisdom" that if a queue forms at the mth intersection during the red interval due to turning

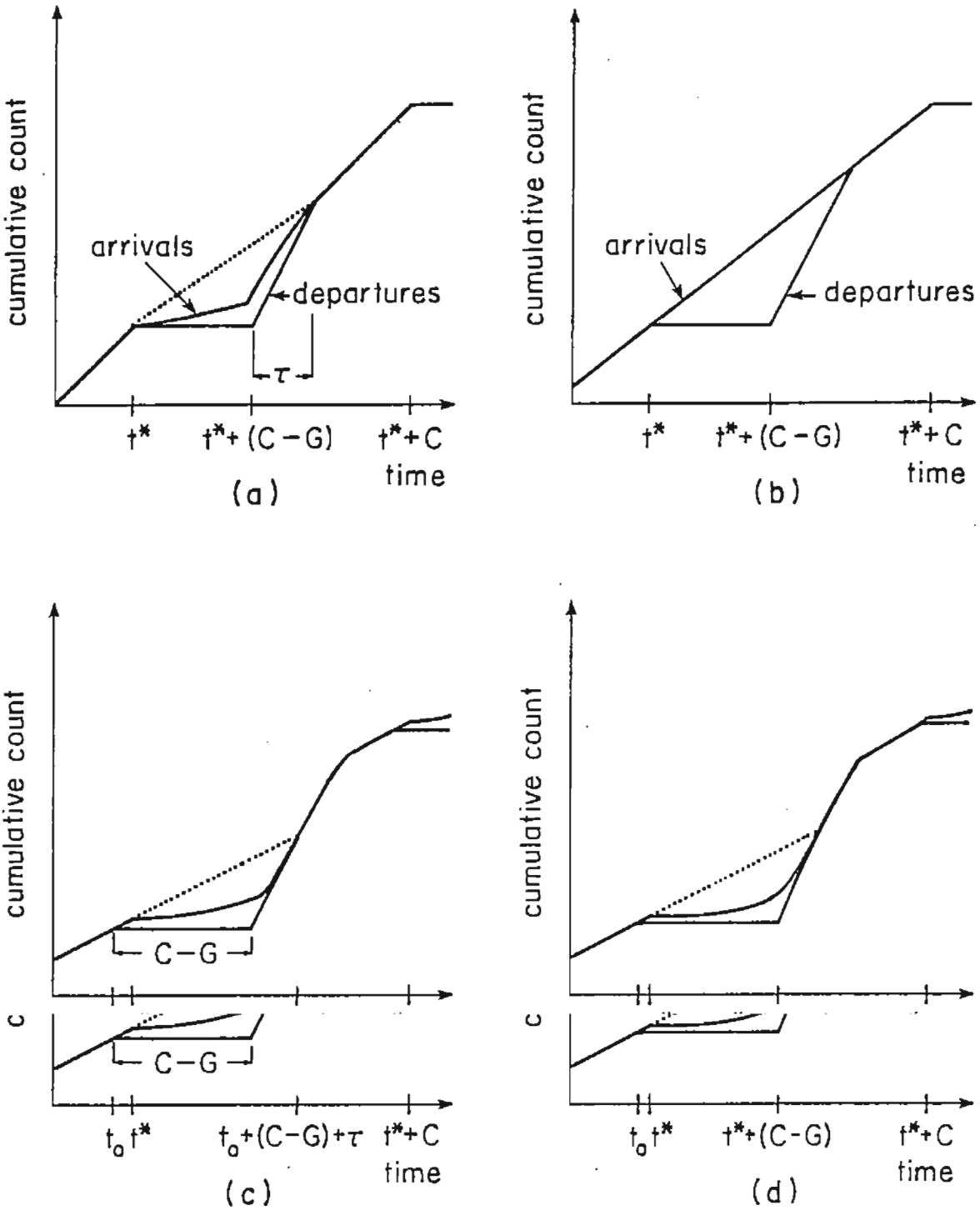


Fig. 3.5 - Examples of arrival patterns and offsets.

vehicles at the $(m-1)$th intersection, that one should advance the off-sets so that the queue vanishes just as the platoon arrives. If the flow $q_1^{(m)}(t^*)$ is sufficiently large, one would rather not advance the signal, cutting off the flow, and forcing some vehicles to miss the green and wait a whole red interval (plus contribute to the queue at the start of the next green). Instead, one would let the queue of turning vehicles compress the platoon back to a flow s_1 and (if necessary) even extend the time τ needed to discharge the queue.

At the other extreme, suppose, as illustrated in figure 3.5c, that the arrival rate $q_1^{(m)}(t^*)$ is sufficiently low and nearly constant for t "close" to t^*, that some vehicles enter the arterial at intersection $(m-1)$, but that very few leave. If one draws a tangent line to the arrival curve at or near time t^*, it will intersect the arrival curve again at some time $t_0 + (C-G) + \tau$. Now draw a line of slope s_1 from this point and locate the time t_0 such that the horizontal segment of the departure curve has width $C-G$.

This illustrates the conventional strategy. If the arrival curve has a slope very close to s_1, the arrival and departure curves will nearly coincide once the arrival curve starts to rise rapidly signalling the arrival of the platoon from intersection $m-1$. Thus, this strategy implies that the queue nearly vanishes just as the platoon arrives.

platoon from intersection $m-1$. Thus, this strategy implies that the queue nearly vanishes just as the platoon arrives.

If, as illustrated in figure 3.5d, some vehicles leave the arterial at intersection $(m-1)$, one may make some compromise between these extreme strategies. One may advance the off-sets slightly, but not enough so that the queue (nearly) vanishes, so as to compress the platoon and fill in the vacancies left by vehicles which departed at intersection $m-1$.

e. Turning traffic, global strategies

In the above strategy, we described how to minimize the delay at a single intersection. The objective, however, is to minimize the total travel time

(delays) for all trips or the sum of delays at all intersections. The discussion of part c should serve as a warning that a reduction of delay at one intersection will not necessarily reduce the total delay at all intersections, it may just transfer the delay from one intersection to another. It is possible that it could even result in a net increase in total delay.

To verify that a strategy as in part d is reasonable, one should consider a sequence of many signals as in figure 3.6. One could draw the x-coordinate

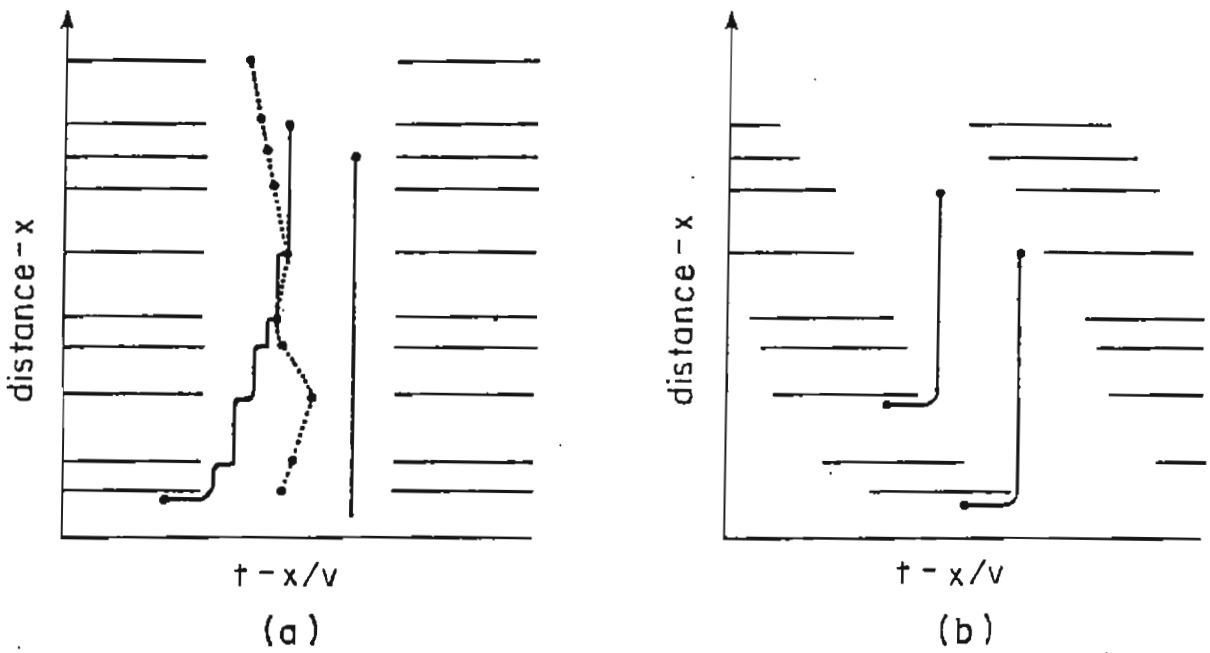


Fig. 3.6 - Some strategies with turning vehicles.

Fig. 3.6 - Some strategies with turning vehicles.

on any scale, but if there is a large number of intersections, they would appear on the graph as if they were "closely" spaced. The red and green intervals would appear as nearly continuous strips.

Vehicles enter the "green band" on the left due to traffic turning onto the arterial from cross streets and they leave from the interior of the band due to vehicles turning off (at discrete intersections). If one were to choose the off-sets equal to the transit time, the green band would be a vertical

strip in the x vs. $t - x/v$ graph as in figure 3.6a.

At each intersection the entering vehicles cause the flow to have a value s_1 until the queue passes the intersection. We could also draw a "path" of the times at which the end of the queue passes each intersection. This is defined only at the intersections but one could connect these points by a smooth curve as indicated by the broken line of figure 3.6a. The important point here is that if every signal is undersaturated, the queue will vanish during every green interval (we are neglecting for now, "stochastic effects"). The "most critical intersection" will be the one which requires the longest time to clear the queue.

Suppose now that some vehicle entered the green band at a time so to pass the critical intersection later than the time needed to clear the queue at the critical intersection. This vehicle would not be delayed at any signal, despite the existence of vehicles turning in and out. Actually this vehicle must have entered the band at the first intersection; it could not be pushed into this position if it entered from a cross street unless the driver intentionally stalled to avoid congestion.

With this strategy, any vehicle which enters from a cross street cannot be delayed more than the green interval. Actually its "delay" is its final position in the green band when it leaves the arterial, which cannot exceed be delayed more than the green interval. Actually its "delay" is its final position in the green band when it leaves the arterial, which cannot exceed the maximum time needed to clear the queue among all intersections between its entrance and exit. Thus, the most one can gain by advancing the off-sets of some signals is that some vehicles may save some fraction of a green interval. This must be balanced against the potentially much larger loss for other (fewer) vehicles which may be delayed a whole red interval possibly repeated several times if they travel far enough.

There is a trade-off, however. For a sufficiently long arterial, the strategy in figure 3.6a is not the optimal strategy. Any vehicle which is

not delayed by any signal should eventually exit the arterial. The flow at the end of the green will, therefore, decrease with x (if it started at the first intersection with some nonzero value) until the expected loss from advancing the off-set of some signal becomes arbitrarily small. Indeed the strategy in part d would advance the off-set when the penalty at some intersection becomes less than the benefit.

The above example illustrates that the evolution of the arrival curves at successive intersections (under any strategy) depends on the distribution of trip lengths (not just on the number of vehicles entering or leaving at each intersection). Figure 3.6b illustrates a case in which the optimal strategy is to advance the off-sets so that the queue clears at each intersection just as the platoon arrives. In this example the trip length of each vehicle is such that it will leave the arterial before the cumulative off-sets of all signals between its entrance and exit exceeds the green interval. No vehicle is delayed. Furthermore, the flow at the end of the green interval is always zero so the strategy in part d will also give these off-sets.

There are some situations in which the strategy described in part d does not give the minimum total delay. In the typical circumstances for this, the gains achieved from some modification in strategy at the mth intersection directly affect the total travel time because the vehicles realize this benefit. The losses to other vehicles, however, are to vehicles which continue through several more intersections. These losses might be ones that would have occurred even without the modification.

Suppose, for example, that by following the strategy in part d, one advanced the off-set at the mth intersection to reduce the delay to vehicles which had entered the arterial at the $(m-1)$th intersection, at the expense of cutting off some vehicle at the end of the green. The latter vehicle, however, was due to exit at intersection m (during the green interval)

so its trip time was actually increased.

As a modification of this strategy, suppose one does not advance the off-set, to the immediate benefit of the vehicle that is leaving. This causes delay to the vehicles at the start of green. It is possible, however, that in the first strategy one did not choose to advance the signal at intersection $(m + 1)$ despite the existence of some new vehicles entering at intersection m, so the vehicles continuing through the mth intersection are delayed at the $(m + 1)$th intersection. In the modified strategy, however, one may now choose to advance the signal at $(m + 1)$, because the flow near the end of the green is low (the vehicles which might have been there exited at the mth intersection). The vehicles which were delayed at the mth intersection may now receive no delay at the $(m + 1)$th intersection.

The number of possible modifications in strategy is astronomical even if one considers only two choices at each of many intersections, either advance the signal to clear the queue or not at all. One should not, however, take this formal problem of minimizing total delay too seriously. It implies that one is willing to trade a delay equal to a red interval for one vehicle to gain (possibly) a much smaller benefit to each of several other vehicles. This is not what people really want but, in any case, one must make some trade-offs possibly giving heavier weight to long delays than to short delays. Whatever one does is somewhat arbitrary, but one certainly should be reluctant to advance off-sets of signals if it leads to some vehicle being cut off at the end of the green.

f) Cross street delay

It is possible that some of the cross streets may be part of a signal progression in the cross direction and give negligible delays to though traffic on the cross street at the intersection with the arterial, but the delays on

other cross streets may be similar to those for uniform arrivals. The total delay to the cross traffic at all intersections is likely to be considerably larger than the total delay on the arterial.

The delay to the through traffic on the cross street depends mostly on the choice of the cycle time C and the arterial green intervals but not on the sequence of phases for turning traffic. The choice of the cycle time and green intervals involves a balance between stochastic and deterministic queues, which will be discussed later. The delays for the traffic turning onto the arterial, however, do depend on the sequence of signal phases. Since, for sufficiently long arterials, most traffic on the arterial is likely to have entered from a cross street, the delays to these vehicles entering the arterial progression may also represent a significant part of the total delay on the arterial. The delays for turning traffic will, of course, depend on the flows and geometry of the cross street.

If the mth cross street is a one-way street, it would likely have a turn bay or turn lane. Vehicles which turn on red (either left or right) would join the end of the arterial platoon or enter gaps in the platoon. This would cause an increase in the arterial flow leaving the mth intersection near the end of the arterial green, perhaps even a significant flow during the yellow interval. There will also be some flow of turning vehicles during the cross street green.

the arterial green, perhaps even a significant flow during the yellow interval. There will also be some flow of turning vehicles during the cross street green.

According to the strategy in part d, the off-sets at intersection m + 1 would be influenced by the turning movements (in or out) at the mth intersection. The net result of this is that one is not likely to advance the off-sets at intersection m + 1 ; one may even retard the off-sets to accommodate some vehicles which turned on the yellow at intersection m because they could not squeeze into the platoon. It is difficult, however, to imagine how one could improve on this strategy. Any signal control of the turning movements at intersection m would only delay the turning vehicles at the mth intersection,

and delaying them at this point certainly would not result in a net gain.

If the cross street is a two-way street, the vehicles turning right onto the arterial would behave as above. If there were no separate left-turn signal, the vehicles turning left onto the arterial would likely turn at the end of the cross street green (after waiting for opposing cross street traffic to clear) or during the following yellow. They would leave the mth intersection just ahead of the arterial platoon and experience, at most, only a short delay at the (m + 1)th intersection.

If there is a separate left turn phase (and a turn bay) and vehicles arrive on the mth cross street at a uniform rate, the delay to left turning traffic at the mth intersection would be (nearly) independent of when the left turn phase occurs in the phase sequence of the signal. The delays on the arterial at the (m + 1)th intersection, however, do depend on the sequence. If the left turn vehicles pass at the start of the cross street green, they will follow an arterial platoon, but, if they pass at the end of the green, they will lead the platoon. In the former case, one must either force the turning vehicles to wait most of a red time at the (m + 1)th intersection or retard the off-set and delay the whole arterial platoon. Obviously, the latter strategy (lagging left) gives less delay.

If the cross street is part of some signal progression in the cross direc-

left) gives less delay.

If the cross street is part of some signal progression in the cross direction, there is even more reason to have a lagging rather than a leading turn phase. Most turning vehicles would enter the turn bay as the cross street platoon is approaching the intersection and during the cross street green. With a lagging turn phase, these vehicles would be served after only a short delay, but with a leading turn phase they would be delayed until the start of of the next green (and then be delayed again when they reach the (m + 1)th intersection).

g. Unequal cycle times

So far, it has been assumed that one would use the same cycle time at all intersections and that the cycle time would be chosen (if possible) so that all intersections are undersaturated, i.e., the cycle time is at least as large as the maximum of the minimum cycle times for all intersections. The minimum cycle times, however, are very sensitive to the flows on the cross street and will typically vary considerably among intersections. If, as is often the case, the combined delay on all cross streets is considerably larger than on the arterial, one should try to reduce the cross street delays by using different cycle times at different intersections rather than spending too much effort trying to fine-tune the off-sets.

As a first illustration, suppose that there is no turning traffic and some intersection m^* requires a cycle time at least twice as large as any other intersection. Thus, if one chooses to operate with a cycle time C and an arterial green interval G at intersection m^*, the fraction of time needed by the m^*th cross street is less than $1 - G/C - L^{(m^*)}/C$, but the other intersections could operate on a cycle time $C/2$, a green interval of $G/2$ and a fraction of time for the mth cross street of $1 - G/C - 2L^{(m)}/C$. Consider now the signal strategy shown in figure 3.7.

In comparing various strategies, it is helpful, as discussed in part c, to consider the times at which vehicles would pass intersection m^* and/or

In comparing various strategies, it is helpful, as discussed in part c, to consider the times at which vehicles would pass intersection m^* and/or leave the system if all signals except m^* were removed. Since the introduction of more signals cannot reduce the total trip time of all vehicles and the G and C are considered as given, we can interpret "delay" as any increase in the total trip time caused by the introduction of other signals. From figure 3.7 it is clear, for example, that if the intervals $G/2$ at intersections $m < m^*$ are fully utilized at flow s_1 (thus also the green interval G at intersection m^*), then all vehicles will arrive at m^* earlier than the times

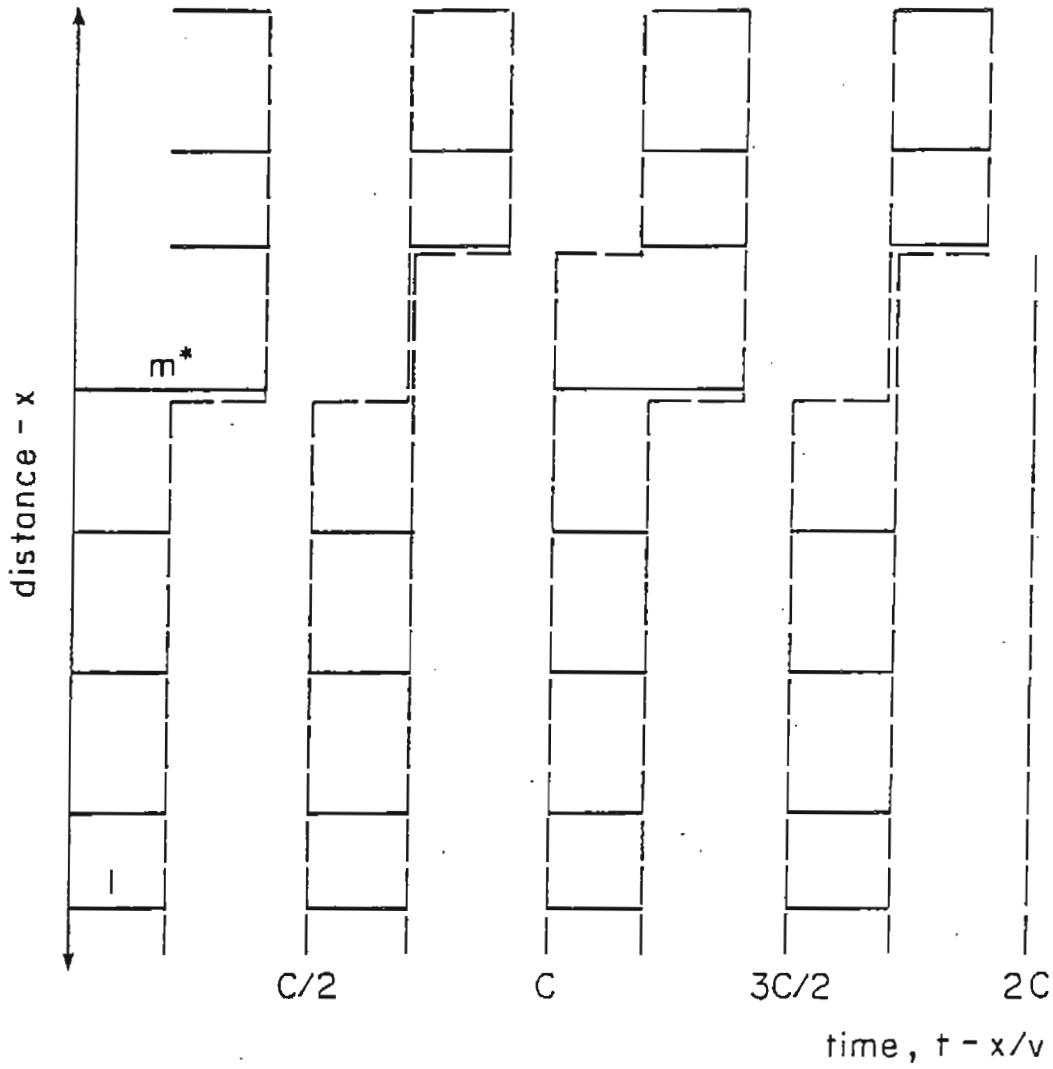


Fig. 3.7 - Coordination with two cycles at each intersection to one cycle at m*.

at which they would leave without the signals $m < m^*$. Thus the signals $m < m^*$ do not change the departure times from m^* or cause any (additional) delay.

at which they would leave without the signals $m < m^*$. Thus the signals $m < m^*$ do not change the departure times from m^* or cause any (additional) delay.

They will, however, cause some vehicles to be stopped at both intersections 1 and m^*.

More generally, even if the green intervals are not fully utilized, it is possible to partition the times C into any number of (unequal) subcycles so that vehicles from each subcycle arrive at m^* just as the signal at m^* finishes serving vehicles from the previous subcycle. We expect, however,

that the G and C would be chosen so that the signal at m* is close to saturation. Although we might also be able to operate the signals m < m* on subcycles of duration approximately C/3, C/4, etc., it is not likely that one would choose to do so. If one operated the signals m < m* on some cycle times between C/2 and C (for example, 2C/3), one could establish a periodic pattern on some multiple of the time C(2C, for example) but this would cause an increase in travel time during some cycles. There are, indeed, many possible strategies, but that shown in figure 3.7 for m < m* is certainly the most appealing.

If m* were the last intersection, or all vehicles exited the arterial at m* , or one operated all signals m > m* on a cycle time C with appropriate off-sets, there would be no further delays of stops beyond intersection m* . The trade-offs between the strategies in figures 3.2 or 3.7 are that the former has fewer stops (but the same total delay) on the arterial, but the latter gives less delay on all cross streets m < m* (typically only about half as much). The increased number of stops is independent of the number of cross streets, but the total saving in delay on the cross streets is additive with respect to all cross streets. Certainly, for a "sufficiently large" number of cross streets m < m* , the latter strategy would be preferred regardless of how one might weigh the cost of stops relative to the cost of delay.

ferred regardless of how one might weigh the cost of stops relative to the cost of delay.

If there are more intersections downstream of m* , one may choose to operate these signals on a cycle time C , or return to the cycle time C/2. If the interval G is fully utilized, the delays downstream of m* will be independent of any strategy for m < m* . One cannot, however, break the flow during G into separate periods of duration G/2 without causing additional delays to (at least) about half the vehicles. Indeed the best one can do is as shown in figure 3.7, which delays about half the vehicles for a time

$(C - G)/2$, giving an average delay per vehicle of $(C - G)/4$ (the same as the average delay per vehicle at the first intersection if vehicles arrive at a constant rate and utilize most of the green time $G/2$ at flow s_1). This also causes about half the vehicle to be stopped at intersection $m+1$, the half which were not previously stopped at intersection m^*.

Whether or not one should use a cycle time $C/2$ downstream of m^* is a balance between the additional delays and stops to the arterial traffic and the savings in delays to all cross street traffic. Again, the former penalty is independent of the number of cross streets whereas the latter is additive. Even if there is another critical intersection downstream of m^* which requires a cycle time C, it may be advantageous to use a cycle time $C/2$ between the two critical intersections if there is enough total traffic on all cross streets between them.

It is, of course, a simple exercise to evaluate the savings in stops and delays for both the arterial and the cross streets and to determine under what conditions use of a cycle time $C/2$ will result in a net saving, for any choice of penalties. On the other hand, one can also assign penalties so as to achieve any conclusion one wishes. There seems to be some implied bias that interrupting the arterial traffic is more annoying (in some sense) than delaying the cross traffic, but use of shorter cycle times at noncritical intersections seems to be one of the most effective ways of reducing delays.

than delaying the cross traffic, but use of shorter cycle times at noncritical intersections seems to be one of the most effective ways of reducing delays.

Vehicles turning on and off the arterial do not substantially alter the above conclusions; they even give further advantage to the shorter cycle times. The average number of turning vehicles in a time $C/2$ is half that in a time C. On some cross streets a cycle time C may necessitate a separate turn phase (if there are more than about 1-1/2 left-turn vehicles per cycle) whereas a cycle time $C/2$ may permit just a two-phase signal. Also, the delay at the $(m + 1)$th intersection due to turning traffic at

the mth intersection will typically be less for two cycles of duration $C/2$ than for one cycle of duration C (and any appropriate strategy of off-sets).

h. Unequal splits with turning traffic

In part c we argued that with no turning traffic and $s_1^{(m)} = s_1$, there would be no reason to choose unequal green intervals $G^{(m)}$, but in sections d to g we assumed that the green intervals $G^{(m)} = G$ would be chosen to be equal even with turning traffic. The primary consideration, however, is to keep all intersections undersaturated, if possible. If we are to use a common cycle time C, it should satisfy (3.2.2), but it does not follow from this that there is also a common green interval G satisfying both conditions in (3.2.1) for all m if the $q_1^{(m)}$ (and/or the $s_1^{(m)}$) are unequal.

There may be arbitrarily many intersections along the arterial but suppose that among these there are two (or three) intersections which barely satisfy condition (3.2.2). At least one of them (at intersections m_1 and also m_3) has a heavy cross traffic which places an upper bound on the $G^{(m_1)}$ and $G^{(m_3)}$, but another intersection m_2 has a relatively large arterial flow (or small $s_1^{(m_2)}$) which places a lower bound on $G^{(m_2)}$. Suppose that $G^{(m_2)} > G^{(m_1)}$ and $G^{(m_3)}$. If $m_1 < m_2 < m_3$ and $s_1^{(m)} = s_1$, there must be considerable traffic turning onto the arterial between intersections m_1 and m_2 so as to create the large arterial flow at m_2, and considerable traffic leaving the arterial between intersections m_2 and m_3 so as to allow intersection m_3 to accommodate a high cross traffic.

Figure 3.8 illustrates a possible signal strategy and some possible vehicle trajectories with $s_1^{(m)} = s_1$. The signals between these critical intersections are not shown, but the figure does show some trajectories originating between m_1 and m_2 (entering from cross streets) and others

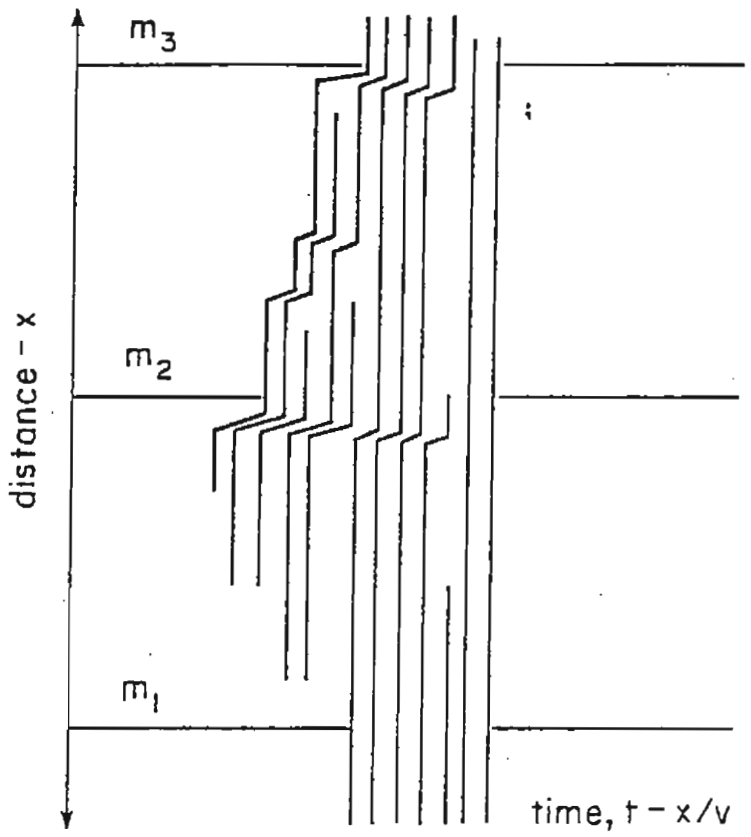


Fig. 3.8 - Coordination with turning traffic.

terminating (exiting) between m_2 and m_3. The off-sets at m_1, m_2, and m_3 are chosen so that a hypothetical vehicle passing m_1 at the end of the green will pass intersections m_2 and m_3 also at the end of the green, without interruption. We assume that the green intervals at these three intersections are (nearly) fully utilized at flow s_1 and that other signals, without interruption. We assume that the green intervals at these three intersections are (nearly) fully utilized at flow s_1 and that other signals, at intermediate intersections, can be set so that each vehicle arrives at any of the three intersections in time so it can leave at the times shown.

From the arguments of section d, it should be clear that these are the optimal off-sets. With a flow at the end of the green close to s_1, advancing the off-set at m_2 or m_3 would cut off the end of the platoon (unless it were known that the last vehicle in the platoon exited the arterial) or, equivalently, displace the lead vehicle by one carried over from the previous

green, after a delay of $C - G^{(m)}$. To retard the off-sets would just delay the whole platoon.

For any intermediate intersection, the end of the arterial green should clearly not interrupt the rear of the platoon, but, if there is any excess of signal time, there is no reason for the green to continue after the platoon has passed. Obviously one should choose the off-sets so that all greens end "simultaneously" (actually with off-set equal to the uninterrupted trip time). On the other hand, there is no reason why any vehicle should arrive at intersection m_2 or m_3 before its scheduled departure time.

The assumption here is that there is some excess capacity at these intermediate intersections. Thus, if one were to give all the excess time to the arterial green at intersection m , it would be sufficient to accommodate all vehicles arriving from upstream, including any which might have turned into the arterial at intersection $m - 1$. But if one gave the arterial barely enough green to accommodate the arrivals at an output flow s_1 , this would also guarantee that any vehicle arrives at m_2 or m_3 in time to leave at its appointed time. There is no reason to give the arterial more green time than it needs except that, with a pretimed signal pattern one does not know precisely how many vehicles may have turned on or off the arterial, and one may wish to give some excess time to accommodate fluctuations (of not know precisely how many vehicles may have turned on or off the arterial, and one may wish to give some excess time to accommodate fluctuations (of course, an excess of vehicles which are allowed to pass intersection m may not be able to clear m_2 or m_3 .) If there is a positive arrival flow on the cross streets even after the queue is discharged on the cross street (as would be true for uniform arrivals on the cross street), assigning any excess green time to the cross street would reduce delays on the cross street.

If the $s_1^{(m)}$ are not all equal, the theory is a bit more complex. With equal values of the $s_1^{(m)}$ we knew that the flow in a platoon approaching

the mth intersection would not exceed $s_1^{(m)}$ and the only issue was whether or not one should delay the lead vehicle in the platoon so as to compress the platoon back to flow s_1. If, however, $s_1^{(m)} < s_1^{(m-1)}$, for example, the flow in the platoon may be larger than $s_1^{(m)}$ at some times and less at other times. The lower value of $s_1^{(m)}$ may be due to a reduced number of lanes, narrower lanes, pedestrian conflict, parked cars, etc.

One can still construct arguments similar to those in section d and figures 3.4, 3.5, but it is now possible that a queue will grow even during the arterial green when the arrival rate exceeds $s_1^{(m)}$. The last vehicle in the platoon approaching the mth intersection may join a queue regardless of the choice of off-sets. The typical conclusion now is that (if the mth intersection is undersaturated) the off-sets should be retarded at least to allow the last vehicle in the platoon to pass during the green, even though it may be delayed by vehicles ahead of it.

It is rather tedious to derive formulas or give illustrations for all situations, but it is fairly straightforward to sketch some possible trajectories and/or cumulative arrival curves, identify what options are available, and choose appropriate off-sets. For unequal $s_1^{(m)}$ one must make special note of the fact that if some $s_1^{(m-1)}$ is larger than $s^{(m)}$ and a platoon arrives at m with flow $s_1^{(m-1)}$, some vehicles will be delayed at intersection m regardless of the off-sets. But then the front of the platoon leaving m at flow $s_1^{(m)}$ may be compressed to a higher flow at subsequent intersections if it is delayed by turning vehicles or must be squeezed through a shorter green interval. Since it is assumed that vehicles cannot recover losses by traveling faster than the design speed, the effect of unequal $s_1^{(m)}$ will be a net (nonnegative) loss.

Having introduced the possibility that one may need to use unequal green intervals $G^{(m)}$ and also the possibility described in part g that one might

operate some signals on a cycle time $C/2$ or $C/3$, we should also consider the possibility of doing both. Also, even if it is possible to accommodate the traffic with a common green interval G (or $G/2$ for half cycles), would it be advantageous to use unequal greens? The number of possible strategies is now becoming rather large, but we do not wish to use some complicated strategy which drivers may find annoying, dangerous, or confusing just to achieve some very modest reduction in total delay.

The available options are determined mostly by capacity considerations. The cycle time C is constrained by (3.2.2) and, for any choice of cycle time at an individual intersection (even if not equal to C), the split is constrained by (3.2.1). Since, for flows close to capacity, the minimum cycle time at an intersection is very sensitive to the flows, there is a strong possibility that the cycle time required by the most critical intersection is nearly or at least twice that of any other intersection or, if there are two (or a few) intersections that require comparable cycle times at least twice that of any other, there are long sequences of adjacent signals which could operate on a cycle time $C/2$.

The use of a cycle time $C/2$ instead of C at some intersection will typically reduce the cross street delays at that intersection by about a half. Clearly this is a much larger gain that one is likely to achieve by giving the cross street any excess green time not needed by the arterial (i.e., by choosing unequal $G^{(m)}$ for a fixed C). Having chosen a cycle time $C/2$, one still has the option of using unequal half-cycle green times $G^{(m)}/2$. It is still true, with unequal $G^{(m)}$, that one can typically switch from a cycle time $C/2$ to C with no increase in total delay (but an increase in the number of stops) on the arterial, but a switch from the cycle time C back to $C/2$ will cause a delay of about $(C - G)/2$ to half the arterial vehicles.

i) Stochastic effects, no turning traffic

Most of the above theory is based on a blatant disregard for "stochastic effects." Actually, there are two types of stochastic effects. One relates to the fact that not all drivers travel at the same speed, or even the same driver at different times. The other relates to the fact that the number of vehicles which arrive at an intersection during any cycle or the number of vehicles turning on or off the arterial in any cycle will vary from cycle to cycle. Consequences of the former type of behavior will be discussed in section 3.3. For now, we will be concerned only with the latter.

To develop a "general theory" for the stochastic behavior of a traffic signal system is out of the question. Even if it were possible, such a theory would be so complex as to be of no practical value. There are, however, some important issues that must be considered. For example, the choice of the cycle time C involves trade-offs between deterministic and stochastic queueing (as was the case even for the isolated intersection). Also, in the above discussion of off-sets, it was, in effect, assumed that the number of turning vehicles in any cycle could be treated as a continuous variable (the average number) even though this number might be only a fraction of a vehicle. Finally, we should consider the question as to whether or not some type of traffic-responsive strategy would be significantly better than a pretimed Finally, we should consider the question as to whether or not some type of traffic-responsive strategy would be significantly better than a pretimed strategy.

As a preliminary introduction to some possible issues, it is convenient first to consider a (rather hypothetical) situtation with no turning traffic and a signal strategy as illustrated in figures 3.2 or 3.7. There may be variations in the headways between vehicles as they pass intersection 1 and consequently also variations in the number of vehicles which can pass intersection 1 during successive green intervals of duration G or $G/2$. It may,

however, be reasonable to assume that a driver who has a short (long) headway at intersection 1 would also have a short (long) headway at other intersections. Thus, any vehicle which can pass intersection 1 during a green interval will also be able to pass all other intersections in the corresponding green intervals. A queue may form behind intersection 1 but there would be none on the arterial at any other intersections.

Stochastic queueing at intersection 1 would be associated primarily with the fact that more vehicles may arrive during some cycle times than can be served during that cycle even though the average number of arrivals per cycle is less than the average number that can be served. If the average arrivals at intersection 1 were uniform over the cycle, the queues and delays would be as described in section 2.3 for an isolated intersection, particularly equations (2.3.1) and (2.3.7).

We could, of course, use unequal green intervals $G^{(m)}$ as discussed in part c, but there would be some intersection m^* with $G^{(m^*)} = \min G^{(m)}$. Presumably m^* would be the "critical intersection" having the largest value of $q_2^{(m)}/s_2^{(m)}$ for the cross traffic. If we made $G^{(m)} > G^{(m^*)} = G$ for $m < m^*$, more than $s_1 G$ vehicles might pass the first intersection but the excess would cause a queue to form at m^*, if not at some intermediate intersection. In the extreme case, we could remove all intersections $m < m^*$. Since the arrival process to intersection 1 and the departures from intersection m^* would be essentially independent of the timing of any intermediate signals (provided they do not prevent vehicles from reaching m^* before their scheduled departure time), the existence of intermediate signals merely changes the location of the queue but not its total size. Thus, nothing is gained by giving more green time $G^{(m)} > G$ for $m < m^*$; one might as well give the excess to the cross street.

Some of the cross streets may, themselves, be part of a coordinated signal strategy in the cross direction and have only small delays at the intersection with the arterial. If one applies the same argument to the cross street as for the arterial, however, one could say that the queues at the entrance to the cross street progression are the same as if the queue were at the arterial. Perhaps other cross streets which are not parts of a cross street progression would behave as if they had uniform (but stochastic) arrivals at the intersection with the arterial

For the signal strategy of figure 3.2, queueing would be essentially the same as for a hypothetical single fixed-cycle signal serving the arrival flow at intersection 1 during one signal phase and all cross street flows "simultaneously" during the other phase. In section 2.3 we discussed "optimal" strategies for serving one or two streams simultaneously, but we tended to disregard the effects of the flow q_4 if $q_2/s_2 > q_4/s_4$, and tried to choose the cycle times and splits so as to balance the stochastic and deterministic queues (only) in directions 1 and 2. The analogous procedure here would be to identify the cross street m^* with the largest $q_2^{(m)}/s_2^{(m)}$, and to disregard the queues at all other cross streets. With possibly a large number of cross streets, however, one cannot very well neglect the total delay at all other cross streets.

Since we have formulas for the delay (at least for uniform arrivals) delay at all other cross streets.

Since we have formulas for the delay (at least for uniform arrivals) as a function of C, G, $L^{(m)}$ and $q_2^{(m)}/s_2^{(m)}$ for each intersection, there is no problem in adding them all together and minimizing the sum (numerically) with respect to G and C. Of course, one cannot very well give an explicit formula for the optimal G and C to see how it depends on all the $q_i^{(m)}$, $s_i^{(m)}$, etc., but the qualitative conclusion would be fairly simple. Unless there is some traffic intersection with a $q_2^{(m)}/s_2^{(m)}$ very close to the largest $(m = m^*)$, the stochastic queueing should be small (perhaps

negligible) at all intersections except $m = m^*$. The "optimal" choice of G and C would, therefore, involve a balance between the stochastic queues associated with the intersection $m = m^*$ (which, however, may be located elsewhere) and the combined deterministic queues at all intersections. The result of minimizing the total delay would be a cycle time C considerably smaller than one would choose if intersection m^* were the only intersection. The C would be reduced in an attempt to reduce the deterministic queues for $m \neq m^*$.

The "excess" signal time $C\left(1 - q_1^{(m^*)}/s_1^{(m^*)} - q_2^{(m^*)}/s_2^{(m^*)}\right) - L^{(m^*)}$, however, would be partitioned between the arterial and the cross street approximately as for the isolated signal, to balance the stochastic queues for the two traffic directions associated with $m = m^*$.

The details of this are somewhat academic. Even if the model were realistic (no pedestrian constraints, no turning traffic, etc.) and the "optimal" C were in an acceptable range, it is not obvious that one would be willing to permit large stochastic queues for the two traffic streams in order to reduce the deterministic queues at all the noncritical intersections. Not only would one typically tend to favor the arterial traffic, but one is not likely to trade possibly large (and uncertain) delays to some vehicles to achieve possibly smaller reductions in delays to a larger number of other (cross street) vehicles.

hicles to achieve possibly smaller reductions in delays to a larger number of other (cross street) vehicles.

The strategy in figure 3.7 may partially resolve this conflict. At the expense of causing additional stops on the arterial and also some additional delays if the cycle time is changed from C back to $C/2$, the deterministic queues at the noncritical intersections can be reduced by about a half. It is also likely now that either some other intersection is close to saturation at the cycle time $C/2$ (otherwise one might have used a cycle time of $C/3$), or one is not willing to use a cycle time less than $C/2$ (for reasons

unrelated to the delays). In the former case, the formal optimization now has another stochastic queueing term and reduced deterministic queues, both of which tend to favor larger values of C. In the latter case, one no longer has an optimization problem because the objective is unclear.

Although it may not be entirely clear what society prefers, certainly one can propose some rational choices of signal coordination (in this idealized situation) that is more logical than the popular procedure of arbitrarily taking C to be 60 seconds or 90 seconds.

We have argued here that if $s_1^{(m)} = s_1$ is the same at all intersections downstream of the critical intersection, $G^{(m)} \geq G^{(m^*)}$ for $m > m^*$, and there is no turning traffic, then any vehicle which can pass m^* during the time $G^{(m^*)}$ should be able also to pass intersection m, $m > m^*$, during the corresponding interval $G^{(m)}$, regardless of possible fluctuations in headways of vehicles passing m^* and independent of the arrival process at m^*. If, as in figure 3.2, the intersections $m > m^*$ operate on a cycle time C with off-sets so as not to intercept the through brand from m^*, there will be no queueing (stochastic or deterministic) on the arterial for $m > m^*$. If, as in figure 3.7, the intersections $m > m^*$ operate on a cycle time $C/2$ and green intervals $G^{(m)}/2 > G^{(m^*)}/2$, there will be some delay at intersection $m^* + 1$ caused by splitting the band into two parts, but again any vehicle which passes m^* in a time $G^{(m^*)}$ should pass m in the same cycle with no stochastic queueing (except possibly that the number of vehicles delayed into the second half cycle band may vary from cycle to cycle).

This point is emphasized here because in nearly all analytic models or computer programs for "optimal" signal coordination, it is postulated that the "stochastic" queueing at <u>every</u> intersection is the same as would exist if each intersection was an isolated intersection with the same flows $q_i^{(m)}$, splits, etc., and a random arrival pattern (such as Poisson arrivals),

statistically independent of the pattern at any other intersection.

Such a postulate grossly overestimates the stochastic queueing (which in the present idealized situation should be zero for $m > m^*$). Because of this postulate, the $G^{(m)}$ at each intersection are usually chosen to balance the (possibly realistically modeled) stochastic queue on the cross street with the (exaggerated) stochastic queue on the arterial, leading to the same recipe for selecting the splits at each intersection that one would use if the signals were at an isolated intersection. The exaggerated stochastic queueing on the arterial also leads to an "optimal" cycle time considerably larger than necessary.

We have also argued here that the total delay on the arterial is the same as if all signals upstream of the critical intersection were removed, provided that the signals $m < m^*$ do not prevent any vehicles from arriving at m^* before it would have left without the signals $m < m^*$. This is true regardless of the arrival pattern of vehicles to intersection 1 and any "deterministic" or "stochastic" queueing. It would be valid for any choice of $G^m \geq G^{(m^*)}$ for $m < m^*$ with off-sets chosen so the signals $m < m^*$ do not prevent a vehicle which is "scheduled" to pass m^* at the end of the green interval from arriving at m^* in time for its appointment. It is also valid if one uses a cycle time $C/2$ as in figure 3.7 for $m < m^*$.

For fixed $G^{(m^*)}$, the strategy with a single cycle time C for all intersections which minimizes the total system delay among those with $G^{(m)} \geq G^{(m^*)}$ is that illustrated in figure 3.1 or 3.2 with $G^{(m)} = G^{(m^*)}$ since it gives the minimum delay on the arterial and also the minimum delay (maximum green interval) for the cross streets. Also, the strategy illustrated in figure 3.4 gives the minimum total delay among all strategies $G^{(m)} \geq G^{(m^*)}$ with a double cycle.

It may, nevertheless, be advantageous to have $G^{(m)} > G^{(m^*)}$ for $m < m^*$ for two possible reasons. First, if $G^{(1)} = G^{(m^*)}$, any (stochastic) queueing on the arterial will occur at or upstream of intersection 1. Perhaps one does not have enough space upstream of intersection 1 to store the entire queue without blocking some intersection there. If, however, one were to choose $G^{(1)} > G^{(2)} > G^{(3)} \ldots > G^{(m^*)}$, the queue, which is actually caused by the intersection m^*, will be distributed among all the intersections 1 to m^* (at no increase in the total delay or the queue on the arterial). Note that any intersection m with $G^{(m-1)} \leq G^{(m)}$ will pass all vehicles which can pass $m - 1$, with no delay. There would be no reason (with no turning vehicles) for choosing $G^{(m-1)} < G^{(m)}$. Second, if the cross street traffic is light at intersection 1, for example, drivers on the arterial may be annoyed if they see that the cross street is receiving an excessive amount of green time while they are being delayed. They may prefer to keep moving toward m^* despite the fact that they will be stopped anyway at m^* or earlier (perhaps they do not realize that this will happen).

Suppose we were to choose $G^{(1)} > G^{(2)} > \ldots G^{(m^*)}$ and the off-sets were chosen so that the $m - 1$th signal does not prevent any vehicle from leaving the mth intersection ($m \leq m^*$) at the same time it would if all intersections upstream of m were removed. (If there were some intersections m with $G^{(m-1)} \leq G^{(m)}$ they can be disregarded and the remaining intersections renumbered so that the above inequalities do apply).

If we let $W_j^{(m)}$ be the total delay experienced by any jth vehicle by the time it passes the mth intersection, then this is also the same as if all intersections upstream of m were removed. It could be evaluated for any arbitrary arrival process at intersection 1 as a function of the signal strategy at m (only), as if m were an "isolated" signal, provided that the queue at intersections downstream of m do not back up over intersection m

preventing vehicles from leaving m at such times as they would if all intersections downstream of m were also removed. If now we let $\Delta^{(m)}$ be the delay to the jth vehicle at intersection m itself, then

$$\Delta_j^{(m)} = W_j^{(m)} - W_j^{(m-1)} \quad \text{for} \quad 2 \leq m \leq m^* \tag{3.2.6}$$
$$\Delta_j^{(1)} = W_j^{(1)} ,$$

for any process of arrivals at intersection 1.

If, for example, the arrivals at intersection 1 were equally spaced with headways $1/q_1$, the delays $W_j^{(m)}$ would be the same as predicted in the "deterministic approximation", as described in section 2.1. Of course, the $W_j^{(m)}$ for any particular jth vehicle will depend on when this vehicle would arrive at intersection m relative to the start of the red interval at intersection m if intersections upstream of m were removed, and it depends on how many other vehicles may have arrived ahead of the jth vehicle since the start of red. The $\Delta_j^{(m)}$, in turn, depends on these parameters for both the mth and $(m-1)$th intersections (and thus on the off-set between $m-1$ and m), but not on the settings of intersections upstream of $m-1$ (provided these signals satisfy the conditions specified above). It is, in principle, straightforward to evaluate the $\Delta_j^{(m)}$ for equally spaced arrivals, for each j, as a function of the signal settings at $m-1$ and m, also principle, straightforward to evaluate the $\Delta_j^{(m)}$ for equally spaced arrivals, for each j, as a function of the signal settings at $m-1$ and m, also for any other specified arrival process even if the queue fails to clear these intersections during the green intervals.

If we take the average of (3.2.6) over all vehicles j and let Δ^m, $W^{(m)}$ denote these averages of the $\Delta_j^{(m)}$, $W_j^{(m)}$ respectively, then

$$\Delta^{(m)} = W^{(m)} - W^{(m-1)} \tag{3.2.7}$$
$$\Delta^{(1)} = W^{(1)}$$

with $W^{(m)}$ given by the appropriate formula for the average delay at an isolated signal as described in sections 2.2 and 2.3. The formulas for the $W^{(m)}$, however, do not depend on the off-sets between $m-1$ and m (provided that intersection $m-1$ does not interfere with the departures from m). Of course, (3.2.7) does not describe the distribution of the $W_j^{(m)}$ or the queue lengths. These distributions do depend on the off-sets between $m-1$ and m in possibly a complicated way. The time average queue length at m, however, is given by $q_1\Delta^{(m)}$.

Equations (3.2.6) and (3.2.7) apply for any arrival pattern at intersection 1. If the arrival pattern should vary from cycle to cycle, one might model the arrivals as a stochastic process as described in section 2.3. The $W^{(m)}$ in (3.2.7) would then consist of two parts, the deterministic delay and the stochastic delay due to the possible overflow queue given by the various approximate formulas described in section 2.3. But since (3.2.7) applies for the equally spaced arrivals at intersection 1, i.e., the "deterministic" part of the $W^{(m)}$ alone, it follows that (3.2.7) applies also if we interpret the $\Delta^{(m)}$ and $W^{(m)}$ as just that part of the delay due to the overflow. Also, if we multiply (3.2.7) by q_1, $q_1\Delta^{(m)}$ would be the equilibrium average residual queue at intersection m at the end of the green interval and $q_1W^{(m)}$, $q_1W^{(m-1)}$ would represent the residual average queues that would exist at m and $m-1$ respectively if all signals upstream of m or $m-1$ were removed (as given by the formulas of section 2.3). The residual queues $q_1W^{(m)}$, $q_1W^{(m-1)}$ would depend on the degree of saturation at m and $m-1$ (i.e., the $G^{(m)}$ and $G^{(m-1)}$) but not on the off-sets.

Since the $W^{(m)}$ is very sensitive to the degree of saturation at m, it would require only a rather small difference between $G^{(m-1)}$ and $G^{(m)}$ to cause the $W^{(m-1)}$ to be appreciably less than $W^{(m)}$. In particular,

prevent most of the overflow queue from occurring at m^* .

What these formulas are trying to describe "physically" is the following phenomena. If during some cycle the actual number of arrivals at intersection 1 should exceed approximately $s_1 G^{(m^*)}$, then the number which pass intersection 1 during that cycle is the smaller of the number which arrive or $S_1 G^{(1)}$ (if there was no residual queue at intersection 1 from the previous cycle). If it is the latter, there will be a residual queue at the end of the cycle. If it is the former, the platoon proceeds downstream until it reaches an intersection m such that the number in the platoon exceeds $s_1 G^{(m)}$ and a residual queue will form there. Since $s_1 G^{(m+1)} < s_1 G^{(m)}$, the $s_1 G^{(m)}$ vehicles which do pass m probably cannot pass $m + 1$, so a residual queue also forms at $m + 1, m + 2, \ldots , m^*$. The total residual queue at all intersections, however, is the same as would occur at m^* if the intersections $m < m^*$ were removed, because any vehicle which overflows the green at m would not have passed m^* in that cycle anyway.

j. Stochastic effects with turning traffic

The evaluation of stochastic queueing for any specified pretimed signal strategy is much more complicated with traffic turning on and off the arterial. In part h we saw that, with turning traffic, one may need to use unequal green intervals $G^{(m)}$ at different intersections to satisfy capacity constraints but, contrary to the preliminary conclusions in part h, with the green intervals $G^{(m)}$ at different intersections to satisfy capacity constraints but, contrary to the preliminary conclusions in part h, with the inclusion of stochastic queueing we will no longer give all of the excess green time at noncritical intersections to the cross street.

The total stochastic queueing at all intersections will certainly be larger than would exist at any single intersection if all other intersections were removed. Suppose again we define the critical intersection m^* as the one with the largest value of $q_1^{(m)}/s_1^{(m)} + q_2^{(m)}/s_2^{(m)}$. In contrast with the situation described in part i where all vehicles entered the arterial at

some intersection 1, we might now consider the opposite extreme in which the arterial is arbitrarily long and all vehicles approaching the critical intersection entered the arterial from cross streets upstream of m^*. The times when the vehicles entered the arterial, of course, depends on the off-sets, but we are not so much concerned here with the delays caused by the off-sets as with the stochastic queueing caused by an excess number of vehicles entering the arterial in any cycle.

If, for now, we disregard possible delays to vehicles from the time they enter the arterial until they reached the critical intersection, the number of vehicles which wish to pass this intersection, in a single cycle is the number which turned onto the arterial upstream of m^*, destined for points downstream of m^* in the appropriate signal cycle. It is reasonable to assume (as we did in section 2.10) that the probability of any particular cross street vehicle turning onto the arterial is independent of whether or not any other vehicle turned. The actual number which turn at any particular intersection in a single cycle would, therefore, have a binomial distribution with a variance nearly equal to the mean number. Correspondingly, the total number of vehicles which wish to pass m^* in any cycle, the sum of those which entered from all intersections upstream, will also have a variance comparable with the mean. Thus, in the absence of any delays at other intersec-

which entered from all intersections upstream, will also have a variance comparable with the mean. Thus, in the absence of any delays at other intersections, the delays at m^* due to an overflow queue would be nearly the same as if the arrivals were a Poisson process of rate $q_1^{(m^*)}C$ per cycle.

The queue of vehicles which wishes to pass m^* in some cycle will not necessarily all be located at m^*; vehicles may have been delayed at intersections before they reached m^* (in part i, the queue could be distributed among intersections 1 to m^*, but it is caused by the constraint at m^*). The issues, however, regarding the choice of the cycle time C and the split at m^* are similar to those described in part i.

The choice of C involves a trade-off between determininistic and stochastic queueing; the choice of the split at m^* involves a trade-off between stochastic queueing on the arterial and stochastic queueing on the cross street (particularly at m^*). Drivers are perhaps more tolerant of (deterministic) delays caused by a long cycle time than stochastic delays on the arterial. They would like to believe that, once they enter the arterial, they can travel without interruption. For flows close to saturation, however, there would be stochastic queueing on the arterial even for $C = \infty$, so one cannot avoid it simply by choosing a large C. For a degree of saturation (for $C = \infty$) less than about 0.6, one could nearly eliminate stochastic queueing by choosing a large (perhaps one minute or more) cycle time. This is what is commonly done, but it is not obvious that it is based on any rational argument. Certainly if there is no stochastic queueing, it is beneficial to reduce the cycle time until there is some. In the absence of pedestrian constraints, there is no obvious reason why one could not operate a signal progression on a cycle time of 30 seconds for light traffic.

The main difference between the situation here and that described in part i is that in part i we assumed that any vehicle which could pass intersection 1 would also pass intersection m^* (and any intermediate intersection) in the "same" cycle (displaced by the transit time). Here the vehicles which might pass intersection m in any cycle differ from those section) in the "same" cycle (displaced by the transit time). Here the vehicles which might pass intersection m in any cycle differ from those which pass intersection $m - 1$, because some vehicles may enter the arterial at intersection $m - 1$ and others may turn off. The mean number of vehicles $q_1^{(m)}C$ may vary with m, as discussed previously, but also there will be a variance associated with the number which pass m, even a conditional variance given the number which passed $m - 1$.

If one has an excess (above the mean) of entering vehicles at intersections $m - 1$ and one has given intersection m barely enough green time

to serve the mean number of arrivals, then there will be an overflow queue at intersection m. One might argue that, if we had given some extra time to $G^{(m)}$, the overflow vehicles might have passed intersection m without delay, but then (for $m < m^*$) they would probably be delayed at m^* anyway. It makes no difference where the vehicles are delayed if they can reach the critical intersection in time to leave at their appointed time, the time they would leave with no upstream delay.

The problem is that some vehicles may not reach the critical intersection at the appointed time. If, for example, one has less than the average number of vehicles turning onto the arterial at intersection $m^* - 1$ but an excess of vehicles turning on at intersection $m^* - 2$, the queue may vanish at intersection m^* while there is still a queue at intersection $m^* - 1$. Thus, the delays at intersection $m^* - 1$ prevent vehicles from reaching intersection m^* at the time when they would otherwise have left.

A detailed theory of queue behavior for a situation such as described here does not exist. It would be very complex, involve a large number of parameters, and describe many types of possible interactions between the queues at different intersections. One can, however, describe a number of issues qualitatively without a detailed theory.

We already have a lower bound for the total system delay. For any choice of C and $G^{(m^*)}$, the total delay must certainly be at least as

We already have a lower bound for the total system delay. For any choice of C and $G^{(m^*)}$, the total delay must certainly be at least as large as the delay at intersection m^* if there were no other intersections and the arrivals on the arterial were a Poisson process of rate $q_1^{(m)}C$ per cycle. Furthermore, the total delay must be at least as large as this delay at m^* plus the sum of the delays to cross street vehicles at all other intersections if the $G^{(m)}$ for $m \neq m^*$ were assigned the minimum values needed to accommodate the arterial flow, i.e., any excess green time is assigned to the cross street. One can also add to this lower bound any

(minimum) delays to vehicles downstream of m^* and any delays to vehicles which exit the arterial upstream of m^*.

On the other hand, to obtain an upper bound we note that the presence of signals upstream of any mth intersection will generally reduce the variance in the number of arrivals per cycle at the mth intersection, particularly positive fluctuations. The stochastic queueing at the mth intersection should, therefore, be less than if the arrivals per cycle had a Poisson distribution with mean $q_1^{(m)}C$. Thus, for any choice of the $G^{(m)}$, the total stochastic delay should be less than the sum of such delays at all intersections which would exist if the arterial arrivals in each cycle were Poisson distributed at each intersection. The minimum total delay in the system must, therefore, also be less than the minimum of this upper bound with respect to any of the parameters, particularly the $G^{(m)}$.

The "deterministic" part of the delay on the arterial depends on the off-sets and signal control for turning movements, but, as described in part h, is nearly independent of how one splits any of the excess times between the cross street and the arterial at intersection m; for given $G^{(m^*)}$. Thus, for fixed C and $G^{(m^*)}$, the choice of the $G^{(m)}$ for $m \neq m^*$ will primarily affect only the deterministic queueing on the cross street and the stochastic queueing on both the cross street and the arterial.

~~The conventional engineering recipe for choosing the $G^{(m)}$ is either~~

the stochastic queueing on both the cross street and the arterial.

The conventional engineering recipe for choosing the $G^{(m)}$ is either to split the available time at intersection m between the arterial and the cross street in the ratio of $q_1^{(m)}/s_1^{(m)}$ to $q_2^{(m)}/s_2^{(m)}$, or to choose it so as to minimize the delay which would exist at the mth intersection if the mth intersection were isolated serving Poisson arrivals in all directions at rates $q_i^{(m)}$. Neither of these recipes is very rational, except possibly for the choice of the $G^{(m^*)}$ as explained already in section i in the case with no turning traffic.

To prevent excessive queueing on the arterial, it is important (a) to keep the bottleneck as busy as possible and (b) to limit the queueing downstream of the bottleneck.

One way to guarantee that the bottleneck is kept busy is to provide a generous amount of green time to the arterial upstream of m^* so that most positive fluctuations in entering traffic will pass the intersections $m < m^*$ without delay and queue at the bottleneck. This may not be the most desirable strategy, however. To prevent blocking of intersections, it may be advantageous that the queue be spread over several intersections. Of course, to prevent excessive queueing downstream of the bottleneck, it also suffices to give the arterial a generous green interval for $m > m^*$.

To guarantee that the bottleneck is kept as busy as possible, it is sufficient, however, to make sure that whenever the queue vanishes at m^* it is unlikely that there is a queue at intersections $m^* - 1$, $m^* - 2$, etc., or conversely that, if there is a queue at $m^* - 1$, $m^* - 2$ ... , then it is unlikely that the queue vanishes at m^*.

Whenever there is a queue at intersection m, the number of vehicles leaving m in one signal cycle should be approximately $s_1^{(m)}G^{(m)}$. The corresponding number of arrivals at intersection $m + 1$ would be this plus the number of vehicles which entered the intersection m in that cycle, less those which exited. For most signal cycles when there is a queue at m , the number of vehicles which entered the intersection m in that cycle, less those which exited. For most signal cycles when there is a queue at m , the number of arrivals at $m + 1$ per cycle should be at least as large as $s_1^{(m)}G^{(m)}$ plus the mean net number entering and leaving minus about one standard deviation of the latter. The standard deviation in question is approximately equal to the square root of the (absolute) sum of the number of vehicles which enter or leave per cycle at m (a number which typically is small compared with $s_1^{(m)}G^{(m)}$). To guarantee that it is unlikely for the queue to vanish at intersection $m + 1$ when there is a queue at intersection

m , it suffices to have this approximate lower bound for the number of arrivals at $m + 1$ exceed the number $s_1^{(m+1)}G^{(m+1)}$ which can leave $m + 1$ in one cycle.

As applied to intersection $m^* - 1$, for example, one can virtually guarantee that the queue will not vanish at intersection m^* when there is a queue at intersection $m^* - 1$, if the green interval at intersection $m^* - 1$ is large enough to supply intersection m^* with a mean number of arrivals per cycle (including turning vehicles) of $s_1^{(m^*)}G^{(m^*)}$ plus one standard deviation of the number of <u>turning</u> vehicles per cycle at intersection $m^* - 1$. In particular, if there is very little turning traffic at intersection $m^* - 1$, it would suffice to choose $s_1^{(m^*-1)}G^{(m^*-1)}$ only slightly larger than $s_1^{(m^*)}G^{(m^*)}$. The average queue at intersection m^* would, in this case, be small compared with the queue at intersection $m^* - 1$ or somewhere upstream.

One would also like to choose the green interval at intersection $m^* - 2$ so that, whenever there is a queue at intersection $m^* - 2$, it is unlikely that the queue would vanish at m^*. With the above choice of $G^{(m^*-1)}$, however. the queue could vanish at m^* only if the queue also vanishes at $m^* - 1$. It suffices, therefore, to choose the $G^{(m^*-2)}$ so that it is unlikely that the queue vanishes at $m^* - 1$ if there is a queue at $m^* - 2$. If one continues this argument to successive intersections upstream, the green interval at intersection m would be chosen so as to keep intersection m^* busy whentinues this argument to successive intersections upstream, the green interval at intersection m would be chosen so as to keep intersection m^* busy whenever there is a queue at m , despite possible (negative) fluctuations in the net total number of vehicles turning or off the arterial between intersections m and m^* . Actually one cannot continue this procedure indefinitely. If, for some sufficiently large $m^* - m$, most vehicles passing m^* turned onto the arterial between m and m^* , then the queue at m^* is nearly independent of what one does at intersection m . Nevertheless, it is clear that little can be gained by giving more green time to the arterial than

described above at intersections $m^* - 1$, $m^* - 2$, ... m until a significant fraction of the vehicles at m^* turned onto the arterial between m and m^*.

A similar (but perhaps more obvious) situation exists downstream of m^*. To prevent queueing at intersection $m^* + 1$, it suffices to choose $s_1^{(m^*+1)}G^{(m^*+1)}$ only large enough to serve the $s_1^{(m^*)}G^{(m^*)}$ vehicles which leave m^* (when there is a queue at m^*) plus the net average number of vehicles entering or leaving per cycle plus about one standard deviation of the latter. (Note that we are assuming here that if there are statistical fluctuations in the number of departures from m^* and $m^* + 1$ when $G^{(m^*)}$ is fully utilized, then there is a high correlation between them. In the absence of turning traffic, any vehicles which can pass m^* in a time $G^{(m^*)}$ can also pass $m^* + 1$ in the corresponding time $G^{(m^*+1)}$). Similarly, to prevent queueing at intersection m further downstream, it suffices to provide enough excess green time to accommodate fluctuations in the number of vehicles entering and/or leaving the arterial between intersections m^* and m. Again, we see that, for sufficiently large $m - m^*$ so that nearly all vehicles passing m entered the arterial between m^* and m, the queue at m will be nearly independent of the queue at m^* no matter what one does.

If we could choose the $G^{(m)}$ for $m < m^*$ so as virtually to guarantee no matter what one does.

If we could choose the $G^{(m)}$ for $m < m^*$ so as virtually to guarantee that each vehicle which passes m^* leaves m^* at the same time as it would without any other intersections, we can evaluate approximately the average (overflow) queues at intersections m^*, $m^* - 1$ ... by methods similar to those used at the end of section i.

In order that the above conditions be true, it would also be necessary that vehicles leave intersections $m^* - 1$, $m^* - 2$, ... , at the same time they would if there were no intersections upstream from these points. In

the absence of intersections upstream of m^* (or $m^* - 1$, $m^* - 2$, ...), however, the number of arrivals at these intersections per cycle would have a Poisson distribution with mean $q_1^{(m^*)}C$ (or $q_1^{(m^*-1)}C$, ...). As in section i, the average stochastic delay per vehicle which passes intersection m, $W^{(m)}$, can be evaluated from the formulas for the overflow queue $Q^{(m)}$ at an isolated intersection with Poisson arrivals $W^{(m)} = Q^{(m)}/q_1^{(m)}$ and the formulas for $Q^{(m)}$ as discussed in section (2.3).

With the existence of intersections $m^* - 1$, $m^* - 2$ (which do not affect the average delay per vehicle passing m^*), the delay per vehicle passing m^* can also be written as the average time $\Delta^{(m^*)}$ which a vehicle spends in the queue at m^* itself, plus the average delay it experiences before passing $m^* - 1$, provided the vehicle did not enter the arterial at $m^* - 1$. If we let $p^{(m^*)}$ be the fraction of vehicles which pass m^* which entered the arterial at $m^* - 1$, the generalization of (3.2.7) for the stochastic part of the delay is

$$W^{(m^*)} = \Delta^{(m^*)} + (1 - p^{(m^*)})W^{(m^*-1)} \tag{3.2.8}$$

or

$$\Delta^{(m^*)} = W^{(m^*)} - (1 - p^{(m^*)})W^{(m^*-1)}.$$

The average (overflow) queue length at m^* is $q_1^{(m^*)}\Delta^{(m^*)}$.

We expect that $p^{(m^*)}$ will be small, but in a hypothetical situation The average (overflow) queue length at m^* is $q_1^{(m^*)}\Delta^{(m^*)}$.

We expect that $p^{(m^*)}$ will be small, but in a hypothetical situation in which all vehicles passing m^* entered at $m^* - 1$, $\Delta^{(m^*)} = W^{(m^*)}$, and the delays at m^* would be independent of the delays at $m^* - 1$. If, however, all vehicles passed $m^* - 1$, $p^{(m^*)} = 0$, and $\Delta^{(m^*)} = W^{(m^*)} - W^{(m^*-1)}$ as in section i.

These formulas are derived under a hypothesis of "point queues," i.e., there are no storage constraints. If the queue at intersection m^* backs up beyond intersection $m^* - 1$ (but vehicles are not allowed to block the

intersection for the cross traffic) the formula for $W^{(m^*)}$ is still valid under the same conditions (certainly intersection $m^* - 1$ would not prevent vehicles from leaving m^* at their appointed time). The formula for $W^{(m^*-1)}$ would not mean much, however, because the queue at m^* would prevent vehicles from leaving (or reaching) $m^* - 1$ at the time they would otherwise. One can use the estimate of $\Delta^{(m^*)}$, (3.2.8), to assess whether or not the queue at m^* would likely back up to intersection $m^* - 1$.

The details of these formulas are not as important as some of the qualitative implications. Presumably, intersection $m^* - 1$ is less congested than m^*; otherwise, it would have been identified as the critical intersection. One can certainly choose $G^{(m^*-1)}$ so that intersection $m^* - 1$ can supply an average of $s_1^{(m^*)}G^{(m^*)}$ vehicles per cycle to intersection m^* when there is a queue at $m^* - 1$. If the fraction of vehicles turning on or off the arterial at $m^* - 1$ is small, it would require only a small additional green time at $m^* - 1$ to compensate for the possible (negative) fluctuations in the arrivals at m^* caused by turning vehicles. This additional green time would not likely cause excessive queueing on the cross street at $m^* - 1$ unless intersections m^* and $m^* - 1$ were almost equally congested. The more likely situation is that this choice of $G^{(m^*-1)}$ will give so much time to the cross street that there is hardly any stochastic queueing on the cross street and one has the option, at little cost, of choosing the $G^{(m^*-1)}$ even larger.

If intersection m^* is quite congested ($Q^{(m^*)}$ is large), a reduction of the degree of saturation at $m^* - 1$ even slightly below that at m^* would make $W^{(m^*-1)}$ significantly less than $W^{(m^*)}$ (even a reduction enough to compensate for fluctuations of only one or two turning vehicles per cycle). The effect of intersection $m^* - 1$ as described by (3.2.8) is that it absorbs some of the larger fluctuations in the arrivals from up-

is less queueing at intersection m^* itself. It has little effect on the delay to vehicles passing m^* and probably little effect on the number of stops. The only obvious penalty to the arterial traffic is that vehicles which exit at $m^* - 1$ must suffer a delay $W^{(m^*-1)}$ (or more, if the queue from m^* backs up over intersection $m^* - 1$).

If one were to choose $G^{(m^*-1)}$ to minimize the total delay, for fixed C and $G^{(m^*)}$, the competition would be between the stochastic delay on the arterial only for the vehicles exiting at $m^* - 1$ and the delay to the cross traffic at $m^* - 1$. If an increase in $G^{(m^*-1)}$ would cause the queue at m^* to back up over intersection $m^* - 1$, then there would be little benefit to the turning traffic. In any case there would be no reason to make $G^{(m^*-1)}$ so large as to eliminate the arterial queue caused by the intersection $m^* - 1$. The details of this are somewhat academic because there typically is little to be gained by choosing $G^{(m^*-1)}$ any larger than necessary to keep intersection m^* busy. (Certainly one should not split any excess green time so as to make the degree of saturation equal on the arterial and the cross street at $m^* - 1$).

Similar arguments can be applied to intersections $m^* - 1, m^* - 2, \ldots, m$, but at each stage one will be giving more and more green time to the arterial in an attempt to keep intersection m^* busy despite possible fluctuations in the increasing number of entering and exiting vehicles between m^* and m. Sooner or later one may encounter some intersection with sufficiently heavy cross traffic that one may need to make some nontrivial trade-offs between the cross street and the arterial delays.

So far, we have discussed the choice of the $G^{(m)}$ for intersections close to the critical intersection. The same type of arguments, however, apply also to intersections adjacent to any other (less) congested intersection.

Suppose, for example, that intersection m' is the most congested intersection upstream of m^*. For any choice of $G^{(m')}$ there is no reason to choose $G^{(m'+1)}$ any larger than necessary to serve the $s_1^{(m')}G^{(m')}$ vehicles from a fully utilized green interval at m' plus the average net number of vehicles entering or leaving per cycle at m' plus about one standard deviation of the latter. This would be sufficient virtually to guarantee that there is no arterial queue at $m' + 1$ (One can just as well imagine that intersection $m' + 1$ is not there.) A similar argument applies to intersections $m' + 2$, $m' + 3$, etc.

Upstream of m', there is little advantage to choosing $G^{(m'-1)}$ any larger than necessary to assure that intersection m' will be kept busy whenever intersection $m' - 1$ is busy, i.e., $G^{(m'-1)}$ should be able to supply at most $s_1^{(m')}G^{(m')}$ vehicles per cycle less the net number of vehicles entering and leaving the arterial at $m' - 1$ plus about one standard deviation of the latter. This, of course, does not guarantee that there is no queue at intersection $m' - 1$, but it does virtually guarantee that intersection $m' - 1$ has no effect on the delay to vehicles passing m'. (A queue at $m' - 1$ will, however, delay vehicles exiting at $m' - 1$). Similar arguments apply to intersections $m' - 2$, $m' - 3$, ...

For the purpose of selecting the $G^{(m')}$ itself, one can now argue that the contribution of this intersection to the total stochastic delay on the

For the purpose of selecting the $G^{(m')}$ itself, one can now argue that the contribution of this intersection to the total stochastic delay on the arterial is certainly less than the stochastic delay to all vehicles passing this intersection that would exist if the arrivals were a Poisson process of rate $q_1^{(m')}$, but it must be at least as large as this multiplied by the fraction of vehicles which exit the arterial between intersections m' and m^*. It is likely to be closer to the latter. Since the stochastic queueing is quite sensitive to the degree of saturation, the stochatic queues at m'

are likely to be considerably smaller than at m^* . If we then weight the queue on the arterial at m' by the fraction of turning vehicles between m' and m^* , one is not likely to admit much stochastic queueing on the m'th cross street.

We have already discounted most of the deterministic queueing to the cross streets for $m \neq m^*$ in selecting $G^{(m^*)}$ and C . If one could operate intersection m' on a cycle time of $C/2$ (or $C/3$), one could still argue that the contribution of this intersection to the total stochastic delay on the arterial is less than if this intersection were isolated with Poisson arrivals, but at least as large as this multiplied by the fraction of vehicles which exit the arterial between intersections m' and m^* . If the stochastic delays caused by intersection m' are not excessive for a cycle time $C/2$, one could operate all intersections $m < m'$ on a cycle time $C/2$ and reduce the deterministic queueing for the cross streets at all of these intersections by about 1/2. In deciding what is excessive, however, one should consider the fact that the stochastic delays on the arterial at m' should be weighted approximately by the fraction of vehicles exiting the arterial between intersections m' and m^* .

One arrives at similar conclusions for intersections downstream of m^* . If m'' is the most congested intersection downstream of m^* , the arrivals at this intersection will be a mixture of vehicles which entered the arterial between intersections m^* and m'' , and those which passed m^* . The latter, however, will have a variance due to the random number of vehicles which exit the arterial between intersections m^* and m'' .

A queue at m'' will delay vehicles passing m'' independent of their origin and one should choose the green interval at m'' to balance the stochastic delays on the arterial and the cross street. The queue on the arterial will certainly be less than if the arrivals on the arterial were Poisson but, for

small fractions of turning vehicles, it would be comparable with this multiplied by the combined fraction of vehicles which either turned on or off the arterial between intersections m^* and m''. This would be true, also, if one chose to operate all signals $m > m^*$ on a cycle time $C/2$.

One could try to evaluate reasonable choices of the $G^{(m)}$ for $m \neq m^*$ from calculations of delays, but one can also "fine-tune" any choices in the field. For intersections downstream of the critical intersection, one can observe the stochastic (overflow) queues on the arterial and the cross street and adjust the $G^{(m)}$ to achieve some reasonable balance between them. This should be done in order of increasing m from m^* (for given values of $G^{(m^*)}$ and C) because any adjustment of $G^{(m^*+1)}$, for example, may affect the queue at $m^* + 2$. If there is never a stochastic queue on the arterial, one should reduce $G^{(m)}$ and give the extra time to the cross street. For intersections upstream of the critical intersection, one should recognize that any reduction of the arterial queue at one intersection may only shift the queue to downstream intersections with no net benefit to the through traffic. One should consider the delays on the arterial only for the vehicle exiting the arterial, but otherwise choose the $G^{(m)}$ only large enough to keep the critical intersection busy. Starting from values of $G^{(m)}$ which are larger than necessary, one could successfully adjust them in the order $m^* - 1$, $m^* - 2$, ...

k) Oversaturated intersections

One cannot always design intersections to accommodate the peak demand. If one increases the cycle time, one can increase the capacity of the critical intersection, but not very much. In any case, there is a practical limit as to how large a cycle time drivers will tolerate. If an intersection becomes oversaturated, there is not much one can do to accommodate the vehicles, except to reroute them (if possible). We will discuss here some relevant issues,

If there is some critical intersection, we might first imagine that one could measure or predict the (average) cumulative number of vehicles $\bar{n}_1^{(m^*)}(t)$, $\bar{n}_2^{(m^*)}(t)$ (see figure 2.5), that would arrive at this intersection by time t in directions 1 and 2 if there were no delays at other intersections. Suppose, also, that the slopes $q_i^{(m^*)}(t) = d\bar{n}_i^{(m^*)}(t)/dt$ are such that, during some time period

$$\frac{q_1^{(m^*)}(t)}{s_1} + \frac{q_2^{(m^*)}(t)}{s_2} > 1 - \frac{L}{C} .$$

If there is no blocking due to intersections downstream of m^* and intersections upstream of m^* can keep intersection m^* busy, the rate of departures from m^* in directions 1 and 2 will be independent of how the vehicles arrive. The total delay to all vehicles which pass m^* in direction i (excluding possible delays downstream of m^*) will be the area between the curves $\bar{n}_i^{(m^*)}(t)$ and the corresponding cumulative departure curve.

Whether vehicles entered the arterial from cross streets or some "first" intersection, the curve $\bar{n}_1^{(m^*)}(t)$ will presumably be some smooth curve independent of any signal strategy (unless the signals cause drivers to change routes). Except for some possible residual queues that may exist when intersection m^* first becomes oversaturated in direction i, the cumulative de-

routes). Except for some possible residual queues that may exist when intersection m^* first becomes oversaturated in direction i, the cumulative departure curves for direction i will depend on the $G^{(m^*)}$ but be essentially independent of how the other signals are set.

If one were concerned only with the delays to those vehicles passing m^*, these delays would be essentially the same as if m^* were an isolated intersection serving the cumulative demands $\bar{n}^{(m^*)}(t)$. Issues regarding how one should split the cycle time between directions 1 and 2 would be the same as discussed previously in section 2.4. If one increases $G^{(m^*)}$, one will decrease the delays on the arterial but increase the delays on the

cross street m^*. It is not clear how one would like to do this, but certainly one would not expect either traffic stream to suffer all of the delays.

The most important conclusion here is that, for any choice of the $G^{(m^*)}$ (possibly time-dependent), the total delay to all vehicles passing m^* on the arterial is nearly independent of anything one does to the signals upstream of m^* as long as these signals do not prevent vehicles from reaching m^* so as to cause the queue to vanish at m^* while there are vehicles queued upstream.

The total delay is also independent of the order in which vehicles pass m^*. A vehicle which turns onto the arterial at intersection $m^* - 1$, for example, may be delayed in reaching this intersection by the queue on the cross street; but once it reaches the intersection, it can usually find some empty pavement on the arterial to squeeze in. It can, therefore, by-pass most of the queue on the arterial which is waiting upstream of $m^* - 1$. The point here is that vehicles need not pass m^* in the order of their expected arrival times at m^* (the time they would arrive if not delayed). The total delay to all vehicles passing m^*, including any possible delays to vehicles on the cross streets which will turn onto the arterial and pass m^*, is independent of the order in which they pass m^*, as long as these delays do not affect the cumulative departure curve from m^*. Whatever one m^*, is independent of the order in which they pass m^*, as long as these delays do not affect the cumulative departure curve from m^*. Whatever one driver may gain from by-passing part of the queue, someone else loses.

Note that the choice of $G^{(m^*)}$ would be based mostly on the queues on the cross street and the arterial at intersection m^*. There is not likely to be a large queue on the cross street at $m^* - 1$ unless this intersection is also oversaturated. But even if it were, this queue would be considerably shorter than the cross street queue at m^*; otherwise, $m^* - 1$ would have been identified as the critical intersection. In the absence of queues on the cross streets for $m \neq m^*$, the vehicles which turn onto the arterial

at $m^* - 1$, in effect,have "priority" over most of those which turn on at $m^* - 2$, which have priority over those from $m^* - 3$, etc.

If there are no alternative routes for drivers who wish to use the arterial, the consequences of a high degree of oversaturation can be very severe. Drivers turning onto the arterial at intersections $m^* - 1$, $m^* - 2$,.. will pass m^* without much delay, but they will consume some of the capacity of intersection m^* . As the queue propagates upstream, the through traffic at intersection m will be blocked by the queue at intersection $m + 1$. The average flow past intersection m which is destined for m^* or beyond cannot exceed the residual capacity at m^* left after the turning vehicles from $m^* - 1$, $m^* - 2$, .. $m+1$ have been served. If the queue propagates upstream far enough, all traffic passing m^* could come from turning vehicles and there would be no capacity left for through traffic from upstream. Thus, the arterial sufficiently far upstream would become a parking lot of stationary vehicles waiting for all the turning vehicles to be served. The vehicles upstream would not be served until almost the end of the rush hour when the queues dissipate. In addition, of course, any vehicles which were supposed to exit the arterial at intersection m^* , $m^* - 1$, .. would be caught in the queue caused by those which are destined for m^* .

Effects analogous to the above actually happen on freeway approaches to bridges or other bottlenecks where there is no suitable alternative

Effects analogous to the above actually happen on freeway approaches to bridges or other bottlenecks where there is no suitable alternative route. Traffic engineers may meter or close ramps in an attempt to provide comparable opportunities for all drivers but the freeway may still become a large parking lot during the rush hour peak. For the typical arterial, however, one would expect that drivers could find some other route rather than stubbornly wait as other vehicles sneak in from the cross streets.

As noted previously in section 1.5, use of the total delay as an objective function is mathematically very convenient because of the insensitivity of the total delay to the order of service or where vehicles are

delayed. On the other hand, because of this, the total delay does not give a proper representation of what society likes. Despite the fact that the total delay is insensitive to vehicles turning onto the arterial and by-passing the queue, no one would believe that a small fraction of travelers should suffer most of the inconvenience of the queue.

Unfortunately there is not much the traffic engineer can do to provide equitable service. He could try to avoid queueing on the arterial by increasing $G^{(m^*)}$, at the expense of the cross traffic at m^*. If the cross street m^* is also a major arterial, however, one will have the same problems there. He could try to restrict the number of vehicles turning onto the arterial at some intersections $m^* - 1$, $m^* - 2$, .. , but if a queue forms in the turn bays of the cross street, it may block the approaches for the through traffic on the cross street (at intersections which may not otherwise be congested). To ban left turns at these (uncongested) intersections would likely produce a less congested flow pattern (if these vehicles can use some alternative routes), but it would be rather difficult to convince political groups that this is an equitable solution.

If an arterial becomes oversaturated, one must either find a place to store the queue (usually on the arterial itself) or find some alternative routes. If the alternative routes pass through residential neighborhoods, one may get some complaints even if there is adequate capacity.

store the queue (usually on the arterial itself) or find some alternative routes. If the alternative routes pass through residential neighborhoods, one may get some complaints even if there is adequate capacity.

Intersections downstream of m^* will typically not create any problem. The flow from m^* is constrained by the capacity output rate, independent of any queues upstream. Since m^* was the critical intersection, one can probably serve at $m^* + 1$, $m^* + 2$, .. any flow which can pass m^*. There is a possibility, however, that the turning traffic onto the arterial at intersections $m^* + 1$, $m^* + 2$, .. could accumulate enough traffic to cause some other intersection to be oversaturated despite the constraint on the output

3.3. Pretimed signals, with platoon spreading

In the previous section we assumed that the trip time between intersections was the same for all vehicles, independent of the position in the platoon, whether or not a vehicle was stopped, or who the driver was. Obviously this is not a very realistic postulate. Unfortunately, we do not have any theory of vehicle dynamics sufficiently reliable to describe in detail how the signal coordination should be modified to correct for the obvious deficiencies. One can, however, make some qualitative description of the consequences of certain effects and at least describe how a desirable choice of off-sets and splits might be related to things which one can observe directly in the field.

a) Uninterrupted flow

In all our discussions so far, it has been assumed that there were saturation flows $s_i^{(m)}$. These are the flows that exist after the first few vehicles have passed the intersection during the green when there is a queue. It is also generally assumed that these are the highest (average) flows that can pass the intersection under any conditions, although there is some evidence to suggest that higher flows can be achieved if vehicles are decelerating as they pass the intersection than if they are accelerating. Lacking sufficient data about the dependence of flows, velocities, etc., on the manner in which vehicles approach the intersection, we will adopt the conventional theme that there exists some empirical relation between speed v and flow q (speed and density, or flow

the intersection, we will adopt the conventional theme that there exists some
empirical relation between speed v and flow q (speed and density, or flow and density) of the type shown in figure 3.9.

Drivers would like to believe that once they enter a signal progression with off-sets designed for an average speed v^* (as in figure 3.1), they can travel at a constant speed v^* through many successive signals (except possibly for some delays caused by vehicles turning on or off the arterial). According to figure 3.9, however, the flow through the intersection at speed $v^* \geq v_m$ (v_m is the speed which gives the maximum flow s), is less than the saturation

put

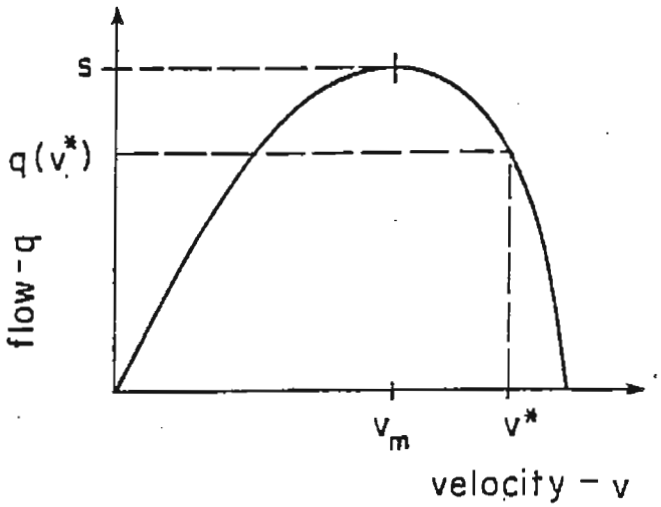


Fig. 3.9 - A hypothetical flow-velocity relation.

flow s . Thus, only about $q(v^*)G < sG$ vehicles can pass the intersection at speed v^* during an effective green time G . Although the curve of q vs. v is flat for v close to v_m , the desired progression speed v^* is typically appreciably larger than v_m and the $q(v^*)$ could be significantly less than s .

If there were some first intersection as in figure 3.1, the flow past the first intersection would be s_1 until the queue vanished. If drivers knew the progression speed v^* , however, they would accelerate from the speed v_m they had when passing intersection 1 and try to adjust to the speed v^* . This will cause the platoon to "spread." Presumably, the flow will stabilize at the value $q(v^*)$ provided that for each mth intersection, the green interval $G^{(m)}$ is large enough that

cause the platoon to "spread." Presumably, the flow will stabilize at the value $q(v^*)$ provided that for each mth intersection, the green interval $G^{(m)}$ is large enough that

$$q_1^{(m)}(v^*)G^{(m)} > q_1^{(m)}C \ . \qquad (3.3.1)$$

This is a necessary condition for the platoon to pass all intersections without any interruptions; it is not a necessary condition for the signals to be "undersaturated."

To satisfy (3.3.1), one must either choose G sufficiently large and/or v^* sufficiently small. If one increases G , one will increase the delay to the cross traffic and even run the risk that the cross street will be oversatu-

rated. If (3.3.1) is to hold and the mth cross street is undersaturated, it would be necessary that

$$\frac{q_1^{(m)}}{q_1^{(m)}(v^*)} + \frac{q_2^{(m)}}{s_2^{(m)}} < 1 - \frac{L^{(m)}}{C} . \qquad (3.2.2)$$

However, one should also choose the C sufficiently large that stochastic queueing is quite rare, because any queues on the arterial would tend to disrupt the smooth flow one is trying to achieve. To eliminate most queueing, the cycle time should probably be about three times the minimum value which satisfies (3.2.2).

One might argue that the $L^{(m)}$ in (3.2.2) should be less than the usual value for stopped traffic because a moving vehicle approaching the intersection just as the signal turns green will avoid some of the start-up time loss. If there is any traffic turning onto the arterial, however, one cannot guarantee that the lead vehicle in a platoon will be able to pass without interruption.

The theory here is not sufficiently precise that one can seriously propose some "optimal" values of the splits and the C. Neither is it clear what the "objective function" is. The point is that for moderate traffic on the arterial one may wish to choose a cycle time appreciably larger than that proposed in the previous section so that traffic can flow smoothly at some proposed progression speed v^*.

gression/speed v^*.

It is not even clear how one should choose the progressive speed for moderate traffic. If one could increase v^* and drivers would accept the increased speed, everyone would have a shorter trip time and thus less "delay." In fact, the traffic engineer will choose a v^* which is reasonable and acceptable, not according to some hypothetical criteria of minimum delay. He would not choose the v^* as the <u>average</u> speed drivers would like but a speed which <u>most</u> drivers could maintain.

It is important to notice that the larger the v^*, the smaller is the

$q(v^*)$. Thus, as the demands $q_1^{(m)}$ increase, it may be necessary to reduce the progression speed v^* if one wishes to maintain an (nearly) uninterrupted flow $q_1(v^*)$ during the green interval.

b. Closely spaced intersections

To illustrate further the points described in part a, suppose that intersections are closely spaced (arbitrarily close), there is no turning traffic, and all signals operate on the same cycle time and splits. If the progression speed v^* were less than the speed at which drivers would like to travel and (3.1.1) was satisfied, a driver could try to travel at a speed larger than v^*, as illustrated in figure 3.10 by the trajectory (a), but he would gain nothing.

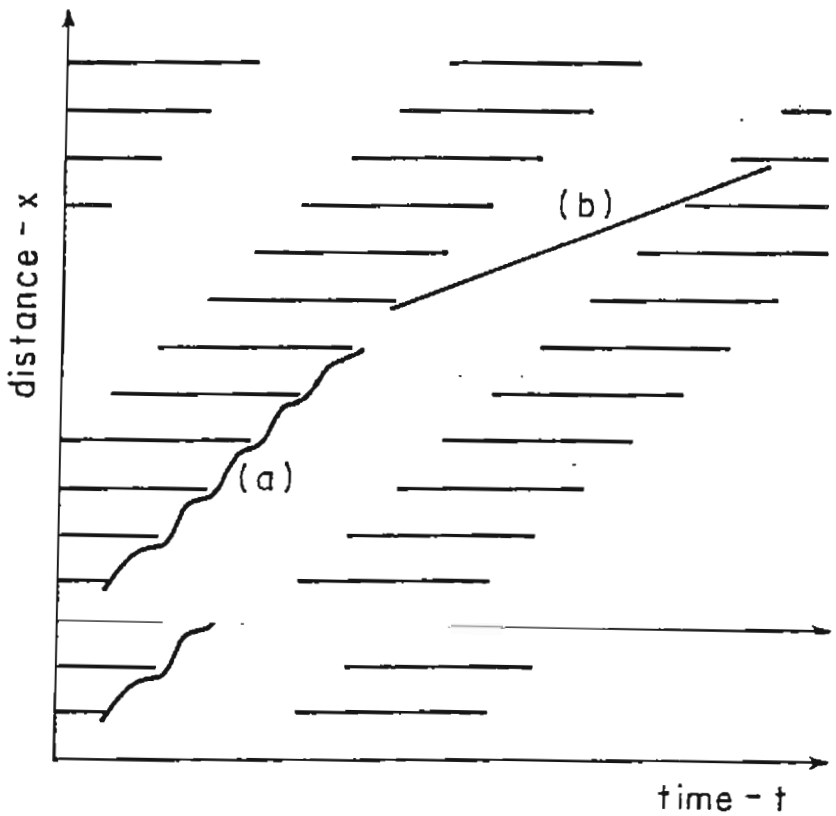


Fig. 3.10 - Vehicles which wish to travel at a speed other than the progression speed.

He might as well just accept the speed v^*. Since the total trip time (delay) is proportional to $1/v^*$, it would be advantageous to increase v^* until either drivers do not wish to drive that fast or the flow cannot pass during the green.

If one were to choose v^* larger than drivers wish to drive, they would travel at their desired speed $v < v^*$ as illustrated in figure 3.10 by trajectory (b). A driver who passes one intersection at the start of the (effective) green and quickly reaches the speed v, will travel a time T such that

$$Tv = (T - G)v^* \quad , \quad T = G/(1 - v/v^*)$$

before being stopped by a red signal and having to wait a time $C - G$. This pattern will then repeat. The long time average speed of the vehicle is therefore

$$\frac{vT}{T + C - G} = \frac{v}{1 + \left(\frac{C - G}{G}\right)\left(1 - \frac{v}{v^*}\right)} .$$

The important feature of this is that this average speed is a decreasing function of v^*, i.e., it is not advantageous to choose a progression speed larger than the speed at which drivers wish to travel.

A simpler algebraic expression can be derived if we define

δ = deviation of the off-set of the signal per unit distance of travel from the trip time

$$= \frac{1}{v^*} - \frac{1}{v} .$$

In terms of this, one can show that

$$v^* \quad v$$

In terms of this, one can show that

$$\begin{matrix}\text{time to travel} \\ \text{unit distance}\end{matrix} = \frac{1}{v} + \begin{cases} \left(\frac{C - G}{G}\right)|\delta| & \text{if } \delta < 0 \\ |\delta| & \text{if } \delta > 0 \end{cases} \qquad (3.2.3)$$

i.e., the "delay" per unit distance of travel as a function of δ has the form shown in figure 3.11.

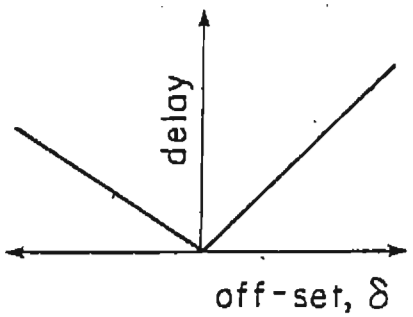


Fig. 3.11 - A shape for the delay vs offset.

We arrive at similar conclusions if the speed of travel is constrained by the flow which can pass during the green interval rather than the speed travelers would like. Suppose that for specified values of q_1, C, and G we define a speed v' as the speed such that

$$q_1 C = q(v')G \ , \tag{3.2.4}$$

i.e., the maximum speed at which vehicles can travel and still pass the intersection during the green interval G. We assume that v' is less than the desired speed v in (3.2.3).

If we choose a progression speed $v^* < v' < v$, vehicles could pass the intersection during the green at the speed v^*, but they could not travel any faster. Obviously it would be advantageous to increase v^* at least until it reaches v'.

any faster. Obviously it would be advantageous to increase v^* at least until it reaches v'.

If we choose a progression speed $v^* > v'$ some vehicles will be cut off from the end of the platoon. But this implies that there will be vehicles waiting at the start of green. If vehicles cannot pass each other at the intersections, a stopped vehicle will delay all vehicles in an approaching platoon. In effect, the average platoon speed is reduced. If intersections are sufficiently close together, the average platoon speed during the green should stabilize to a nearly constant value such that exactly $q_1 C$ vehicles can pass

during the green interval (provided the signals are undersaturated, $q_1 C < sG$), i.e., to the average speed v' .

The evaluation of the average speed of vehicles is now essentially as in (3.2.3) except that the v is replaced by v' . We thus conclude that the optimal choice of v^* (for fixed G , q_1 , and C) is the v' of (3.2.4)

To implement this strategy, it is not necessary that one evaluate a speed-flow relation and do all the subsidiary calculations. As is true of any of the strategies described so far, the conclusion is that the off-sets should be chosen so that a typical driver at the end of the platoon can pass all intersections without being stopped. It is generally easier to infer the appropriate offsets from direct observation of the transit time for vehicles at the end of the platoon than to try to evaluate it from formulas of questionable accuracy. Note also that the travel time of the last vehicle in a platoon should be relatively insensitive to variations in behavior of the lead driver (if he jumps out fast and then stops), or to small fractions of turning traffic.

In essence, we can still apply many of the same arguments as described in section 3.2 but with the speed v in section 3.2 replaced by the (flow dependent) speed v' . One should recognize, however, that one may be able to increase the v' by increasing G and C (at the expense of delay to the cross streets).

increase the v' by increasing G and C (at the expense of delay to the cross streets).

c. Platoon spreading between intersections

The theory described in part a and b was based on the premise that the lead driver of a platoon and all followers, when faced with a known progression speed v^* , would accept this speed and try to travel at a nearly constant speed to match the v^* . The opposite premise would be that a platoon leaving some intersection would behave independent of the setting of any downstream signals until the platoon reaches the downstream signal. The former hypothesis is perhaps closer to the truth for closely spaced intersection and/or

if the progression speed is posted. The latter may be better for intersections far enough apart or for drivers who are unaware of the progression speed.

There have been many experimental observations of "platoon spreading" presumably for intersections arbitrarily far apart, and people do commonly base signal off-sets on the premise that the arrival times at a downstream intersection are the same as they would be if the downstream intersection were not there. There does not seem to have been much (any?) experimental investigation of how drivers react to changes in the signal strategy. This would, indeed, be a very tedious thing to study.

Experimental observations typically show that the time it takes for a platoon to pass some point a distance x downstream of an intersection increases with x, at least for sufficiently large x. There are, however, some observations which suggest that this time may decrease at first and then increase, i.e., the platoon flow reaches a maximum at some nontrivial distance (maybe 200 ft) downstream of the intersection. No one seems to have seriously studied the possible implications of this latter effect (should one try to "shoot" vehicles at the critical intersections from intersections upstream?)

In section 3.2d we discussed some "local optimal" strategies for serving a more or less arbitrary periodic pattern of arrivals at some intersection. (The "delays," however, were interpreted as the delays until the vehicles passed the intersection at flow $s_1^{(m)}$ if there were a queue. It did not include possible delays downstream.) The arrival patterns were assumed to be generated by vehicles turning on or off the arterial at upstream intersections, but the shape of arrival patterns caused by turning traffic is not much different from that caused by platoon spreading. The conclusion in section 3.2d was that one would not typically cut off vehicles from the end of the platoon unless the flow at the end of the platoon was well below the average flow over

a whole cycle $q_1^{(m)}$. This conclusion applies here also. One should time the signal for the last vehicle in a platoon (not a straggler) and force the lead vehicle to slow down and compress the platoon to a higher flow so that it can pass during the green interval.

Whether or not the behavior of the lead (and following) vehicles in a platoon depends on the off-sets of signals is quite critical. As happens so often in transportation, the "user optimal"strategy for individual drivers does not necessarily lead to the "social optimal." The lead drivers of a platoon have no incentive to drive faster than the progression speed between intersections. Indeed, his effort is minimized if he travels at the speed of the progression rather than driving at his (higher) desired speed and then having to stop. The social optimal, however, is the latter strategy. The faster the lead vehicle travels, the faster the last vehicle can travel (after some time lag). If we choose the off-sets so that the last vehicle clears during the green, then the faster the lead vehicle travels, the faster one would choose the progression speed. For the drivers of the lead vehicle in one platoon to drive fast does not give any benefit to himself but to the lead vehicles of succeeding platoons. One can increase the progression speed only if all (or most) lead vehicles travel fast.

On the other hand, maybe the platoon spreading on a signalized arterial is fast.

On the other hand, maybe the platoon spreading on a signalized arterial is much less than on a highway downstream of a single intersection. Also, the effective flow-velocity relation may be quite different for signalized arterials than for comparable unsignalized highways. On a signalized arterial drivers near the end of a platoon should be motivated to travel at the progression speed even if they must travel at short headways to do so. Thus the average speed may be much less sensitive to the flow for signalized arterials. Indeed there is some evidence to suggest that, for reasonable choices of v^* , the drivers will travel at the speed v^* , nearly independent of the flow up to flows quite close

to the saturation flow, and that the platoon spreading is negligible.

So far, we have considered the possible effects of platoon spreading only for $G^{(m)} = G$. If, however, as in section 3.2j, there is some critical intersection m^*, the existence of platoon spreading may affect the choice of the $G^{(m)}$ for $m \neq m^*$. In section 3.2j we argued that there was no reason to choose $G^{(m^*+1)}$ any larger than necessary to accommodate the maximum output from intersecton m^* plus any effects of turning vehicles. It is still true that the off-sets should be chosen so that the typical last vehicle in a platoon from intersection m^* can pass $m^* + 1$ without delay. If, however, the choice of $G^{(m^*+1)}$ causes the lead vehicle to be delayed at $m^* + 1$ so as to compress the platoon, this disturbance will eventually propagate to the last vehicle and cause the speed of the last vehicle to decrease (usually at some point well downstream of $m^* + 1$). This, in turn, will mean that one should delay the off-sets at some intersection downstream of $m^* + 1$. The end result, of course, is a reduction in the average speed, i.e., a delay.

If there is some excess green time to distribute at intersection $m^* + 1$, then there is some advantage to choosing the $G^{(m^*+1)}$ and the offsets so that neither the first nor the last vehicle in the platoon is delayed. The consequences of doing this at intersections $m^* + 1$, $m^* + 2$, ... is that the platoon will continue to spread at intersections $m^* + 1$, $m^* + 2$, ... Eventually

quences of doing this at intersections $m^* + 1$, $m^* + 2$, ... is that the platoon will continue to spread at intersections $m^* + 1$, $m^* + 2$, ... Eventually one will need to compress the platoon at some intersection, m'' say, in order to accommodate the cross traffic. Meanwhile, however, one has allowed vehicles in the platoon to spread and travel more or less as fast as they can.

On the other hand, if one must compress the platoon at some intersection m^* or m'', there is likely to be little penalty for compressing it already at some point upstream of these intersections, at $m^* - 1$ or $m'' - 1$. Delaying the lead vehicle in the platoon at intersection $m^* - 1$ in anticipation

that it will be delayed anyway at intersection m^* will not affect the motion of the last vehicle in the platoon until some point downstream of $m^* - 1$, possibly not until after the last vehicle has passed intersection m^* anyway. Indeed there may be some advantage partially to compress the platoon at $m^* - 1$ so that the lead vehicle in the platoon must only slow down somewhat to pass m^* but not come to a complete stop.

For any actual arterial one could adjust the green intervals and off-sets, and observe how drivers respond. There is no theory, however, sufficiently reliable to describe these effects in anything but a qualitative way. These are not even any experimental studies yet that shed much light on these issues.

d. Light traffic, closely spaced intersections

In most of the discussion above, it was implied that the platoon spreading (if any) was due primarily to the tendency of drivers to follow each other with increasing headways as the speed increases. For light traffic and intersections sufficiently far apart, however, drivers may pass each other between intersections and there will be little interaction between drivers except possibly at the intersections.

As an idealization of this situation, one might imagine that (for sufficiently light traffic) there is essentially no interaction between vehicles. Each driver responds to the signals independent of any other vehicles. If the signals were sufficiently close together and the signals set at some progres-

Each driver responds to the signals independent of any other vehicles. If the signals were sufficiently close together and the signals set at some progres-sion speed v^*, a driver who is willing to drive at least as fast as v^* would be able to travel an arbitrarily long distance without ever being stopped. If he should drive faster (or slower) than speed v^* for a while due to random disturbances, inaccurate knowledge of v^*, etc., he would observe that he was arriving at successive intersections earlier (or later) in the green intervals and he could make adjustments in his speed. He would not just be making corrections for deviations in his speed from v^*, he would be correcting for deviations in his

arrival time at successive intersections. If we thought of a "platoon" as some hypothetical superposition of noninteracting trajectories, such a behavior would give essentially no "platoon spreading." There would be some distribution of the times vehicles passed each intersection but there would be no tendency for this distribution to spread (if v^* is slower than any drivers wishes to drive).

The choice of v^* may be made on the basis of the speed limit or some other socially acceptable criteria. One could, however, base the choice of v^* on the speeds drivers would like to travel. Not all drivers like to drive at the same speed so, for any choice of v^*, there will be some drivers who would like to travel faster than v^* (but can't) and other drivers who wish to and do drive slower than v^*.

If the intersections were sufficiently close together that most drivers with $v < v^*$ would travel several intersections before being stopped by a red signal, we could interpret the average travel time (including stops) of any driver who wishes to travel at a speed v as in (3.2.3). If there is a probability density $f_V(v)$ of drivers who wish to travel at speed v, the average time to travel unit distance (for all drivers) would be the average of (3.2.3) over the distribution of v, i.e.,

$$\text{average time to travel unit distance} = \int^{\infty} \frac{1}{v} f_V(v)dv$$

over the distribution of v, i.e.,

$$\text{average time to travel unit distance} = \int_0^{\infty} \frac{1}{v} f_V(v)dv$$

$$+ \int_0^{v^*} \left(\frac{C - G}{G}\right)\left(\frac{1}{v} - \frac{1}{v^*}\right) f_V(v)dv + \int_{v^*}^{\infty} \left(\frac{1}{v^*} - \frac{1}{v}\right) f_V(v)dv \qquad (3.2.5)$$

$$= \frac{1}{v^*} + \left(\frac{C}{G}\right) \int_0^{v^*} \left(\frac{1}{v} - \frac{1}{v^*}\right) f_V(v)dv \ .$$

If, for given values of C and G, we choose v^* so as to minimize (3.2.5), we obtain the condition that v^* should satisfy

$$F_V(v^*) = G/C \qquad (3.2.6)$$

with

$$F_V(v^*) = \int_0^{v^*} f_V(v)dv \ ,$$

the fraction of drivers with desired speed v less than v^*, i.e., the v^* should be equal to the G/C fractile of the velocity distribution. Thus, the larger the fraction of time G/C one gives to the arterial, the faster is the optimal speed v^*. Since the velocity distribution is typically rather narrow compared with mean and the range of acceptable values of G/C is not very large, the optimal v^* is not expected typically to be very much different from the mean desired speed of all drivers.

The more interesting aspect of (3.2.5) is that it depends on the ratio G/C but not otherwise on G or C separately. If for fixed values of G/C one were to increase both G and C, a driver with $v < v^*$ would be able to travel further during the green interval before being stopped, but the time he is stopped $C - G$ would also increase proportionally. Thus his average speed (including stops) would remain the same, although the number of stops per unit distance of travel would decrease.

It follows from this that, for any given value of G/C and thus fixed value of the delay on the arterial, the value of G (or C) itself affects

It follows from this that, for any given value of G/C and thus fixed value of the delay on the arterial, the value of G (or C) itself affects

only the cross street delay. If the traffic on all cross streets is also light ($q_2^{(m)}/s_2^{(m)}$ small compared with 1) but uniform in time, the average delay per cross street vehicle (as in section 2.2) is approximately

$$\text{delay per vehicle} = \frac{(G + L)^2}{2C} = \frac{(G/C)}{2}(G + 2L + L^2/G) \ , \qquad (3.2.7)$$

the fraction of stopped vehicles $(G + L)/C$ times the average delay $(G + L)/2$ for each vehicle which is stopped.

For fixed value of G/C, the delay per vehicle on the cross street will be minimized with respect to G if G is chosen so that

$$G = L \quad . \tag{3.2.8}$$

Note that G here is some appropriate "effective" green interval and the L is the total effective lost time per cycle, but if there is at most one vehicle leaving the signal per cycle on either the arterial or any cross street, one may need to define the G and L somewhat differently than in section 1.4 and (2.2.4). Regardless of the precise interpretation of G and L, however, the interesting conclusion is that the "optimal" arterial green interval depends only on the effective lost time per cycle, independent of the choice of C, the progression speed v^* or any weights attached to delay on the cross street relative to the arterial. This optimal green interval (with no pedestrian constraints) is considerably less than used in practice! It is likely to be only about ten seconds. Of course, for G close to the optimal value, the delay is quite insensitive to the choice of G so one need not satisfy (3.2.8) very accurately.

With the progression speed v^* determined from (3.2.6) and the green interval from (3.2.8), there is still one final parameter to determine; namely, the split G/C. The choice of this definitely will depend on the competition between the arterial and cross street delays. The cross street delay per vehicle, for fixed G, is proportional to G/C but the average arterial delay per vehicle per mile the arterial and cross street delays. The cross street delay per vehicle, for fixed G, is proportional to G/C but the average arterial delay per vehicle per mile of travel is proportional to C/G.

We have assumed here that G/C would be the same at all intersections along a very long arterial or at least nearly constant over distances comparable with a typical trip length on the arterial. The delay (3.2.7) is the average delay per vehicle on the cross street whereas (3.2.5) gives the average delay per vehicle per unit distance of travel on the arterial. If there were some nearly constant cross street flow <u>per unit length</u> of arterial (actually a "flux") q_2^* (i.e., the sum of the cross street flows on all cross streets per unit length of arterial),

we could define a cross street delay per unit time and unit length of arterial as q_2^* times (3.2.7). Similarly, if there were a constant flow q_1 on the arterial, the delay on the arterial per unit time and unit length of arterial would be q_1 times (3.2.7).

We can thus define a total delay per unit time and unit length of arterial as the sum of contributions from the arterial and the cross streets; namely,

$$q_2^* \frac{(G/C)}{2}(G + 2L + L^2/G) + \frac{q_1}{v^*} + \left(\frac{C}{G}\right) q_1 \int_0^{v^*} \left(\frac{1}{v} - \frac{1}{v^*}\right) f_V(v)dv \qquad (3.2.9)$$

or if we choose $G = L$, (3.2.9) gives

$$2q_2^* L(L/C) + \frac{q_1}{v^*} + \left(\frac{C}{L}\right) q_1 \int_0^{v^*} \left(\frac{1}{v} - \frac{1}{v^*}\right) f_V(v)dv \ . \qquad (3.2.10)$$

For any v^* this now appears to be a well defined function of L/C which can be minimized with respect to L/C, but there is a restriction. Certainly we must have $C > G + L = 2L$, but if a formal minimization of (3.2.10) gives L/C larger than 1/2 (negative cross street green), this implies that we should actually give the cross street barely enough time to serve one vehicle.

A formal minimization of (3.2.10) with respect to L/C gives

$$\left(\frac{L}{C}\right)^2 = \frac{q_1}{2q_2^* L} \int_0^{v^*} \left(\frac{1}{v} - \frac{1}{v^*}\right) f_V(v)dv \ . \qquad (3.2.11)$$

$$\left(\frac{L}{C}\right)^2 = \frac{q_1}{2q_2^* L} \int_0^{v^*} \left(\frac{1}{v} - \frac{1}{v^*}\right) f_V(v)dv \ . \qquad (3.2.11)$$

but to determine both L/C and v^* one must solve (3.2.11) and (3.2.6) simultaneously. Although it is fairly straightforward to solve these equations numerically (if the resulting $L/C < 1/2$) it is not easy to interpret, except to note that the split $L/C = G/C$ will increase with the flow q_1 on the arterial and decrease with the flow on the cross street (but only proportional to the square root of these flows).

e. Light traffic, large spacing between intersections

If, as in part d, we assume that each driver has a certain desired speed v that he will maintain at all times except when stopped, there is an interesting phenomena which might exist for sufficiently large spacing between intersections. The key assumption here is that if a driver with desired speed v is stopped at some intersection, he will return to (exactly) the same speed v that he had before, independent of any signal coordination.

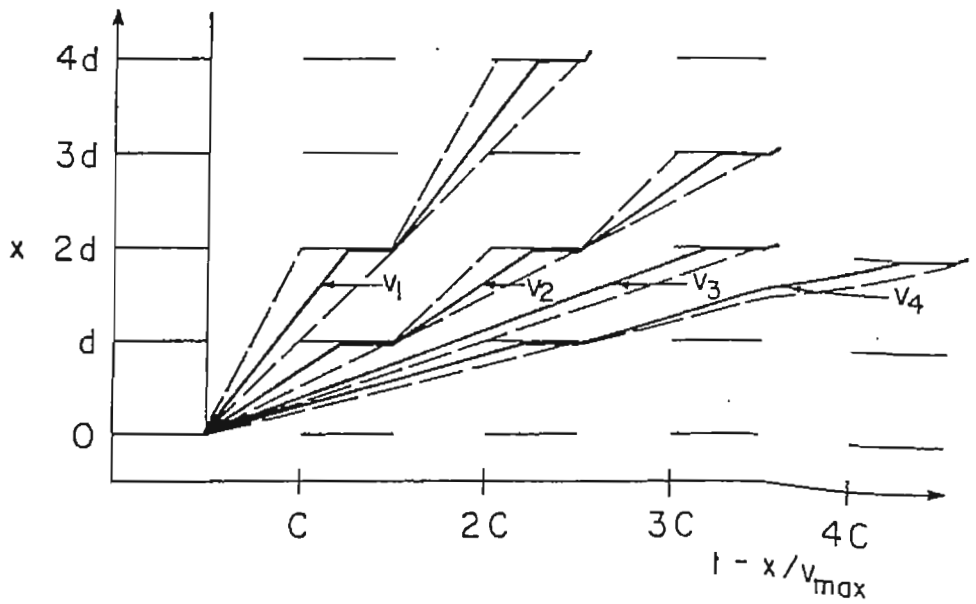


Fig. 3.12 - Trajectories of vehicles which retain the same speed.

Figure 3.12 shows a hypothetical sequence of equally spaced signals at spacing d. The x-coordinate can be scaled in any units. Although in the

Figure 3.12 shows a hypothetical sequence of equally spaced signals at spacing d. The x-coordinate can be scaled in any units. Although in the figure it may appear that the signals are "close together," they are actually "far apart" (possibly several miles). The x coordinate is, in effect, measured in multiples of the d, whatever it may be. We have also drawn the "time-axis" as in figure 3.2 but relative to some fast vehicle traveling at speed v_{max}. A driver traveling at speed v_{max} would have a vertical trajectory in figure 3.12, but any vehicles with $v < v_{max}$ would have a trajectory with a positive slope. The time coordinate is measured in multiples of the cycle

With this scaling of the time-space diagram, the "time" it takes a vehicle at some typical speed $v \leq v_{max}$ to travel between intersections is $d/v - d/v_{max}$. For sufficiently large d, it is possible that this "time" could be comparable with the cycle time C, maybe several times C. The typical spread in velocities is likely to be at most about 20 percent, so the typical value of $d/v - d/v_{max}$ is perhaps about $d/5v_{max}$. For this to be comparable with C means that we are considering spacings d of the order of $5\, v_{max} C$, i.e., about five times the distance a vehicle can travel in a cycle time (typically in the range of at least one to five miles, depending on the v_{max} and C).

If we imagine a "platoon" as a hypothetical superposition of noninteracting trajectories of vehicles with various speeds v, then the situation we are concerned with here is one in which the spreading of a platoon starting at the beginning of the green from one intersection exceeds the cycle time by the time the vehicles reach the next intersection.

Figure 3.12 shows some typical trajectories leaving the first intersection at the start of green at velocities $v_{max} > v_1 > v_2 \ldots > v_4$. The signals are coordinated for a progression speed v_{max}, but if the arrival times of vehicles at the second intersection are spread more or less uniformly over several cycles, the average delay to vehicles at the second intersection will be nearly independent of the off-sets. If $G = C/2$, about half the vehicles will be stopped at the second intersection and the average delay per vehicle will be approximately $C/8$.

at the second intersection and the average delay per vehicle will be nearly independent of the off-sets. If $G = C/2$, about half the vehicles will be stopped at the second intersection and the average delay per vehicle will be approximately $C/8$.

The delays at subsequent intersections are not the same as at the second intersection. If vehicles which are stopped at the second intersection return to exactly the same speed they had before and the offset between the second and third intersections is the same as between the first and second, then any vehicle stopped at the second intersection will also be stopped at the third intersection

and every intersection thereafter. If about half the vehicles behave in this way, they will themselves contribute an amount 1/2 to the average number of vehicles stopped per intersection and C/8 to the average delay per vehicle per intersection.

For those vehicles which are not delayed at the second intersection, half of them will be stopped at the third intersection, a sixth of them at the fourth intersection, etc. Each vehicle stopped at the third intersection would be stopped also at intersections 5, 7, 9, etc. Each vehicle stopped for the first time at the fourth intersection would subsequently be stopped again at intersections 7, 10, 13, etc. All together the fraction of vehicles stopped per intersection will be

$$\frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{12} \cdot \frac{1}{3} + \ldots \cong 0.67$$

and the average delay per vehicle per intersection would be about 0.174C. This result is quite insensitive to the progression speed.

The scheme of coordination illustrated in figure 3.12 is obviously not very efficient. If we had chosen the offsets "at random" and independent of each other, the probability of a vehicle being stopped at any intersection would be 1/2 independent of any previous events. The average number of vehicles stopped per intersection would therefore be 1/2 and the average delay per vehicle per intersection would be C/8 = 0.125C. Thus, the customary scheme of stopped per intersection would therefore be 1/2 and the average delay per vehicle per intersection would be C/8 = 0.125C. Thus, the customary scheme of progression for any fixed progression speed is worse than "no progression."

One can readily see why the coordination scheme of figure 3.12 is inefficient. Vehicles which manage to pass the second intersection without delay during any cycle will be spread over the whole green interval and, by the time they reach the third intersection, they will be spread over a whole cycle time. The delay to these vehicles at the third intersection would be nearly independent of the offset between the second and third signals, so there is nothing one can do to help them anymore. The vehicles which were stopped at the second intersection,

however, will leave this intersection at the start of the green and will be spread only over a time G when they reach the third intersection. Obviously it would be advantageous to set the third signal in just the opposite phase to figure 3.12 so that these vehicles which were stopped at the second intersection can pass the third intersection without being stopped again.

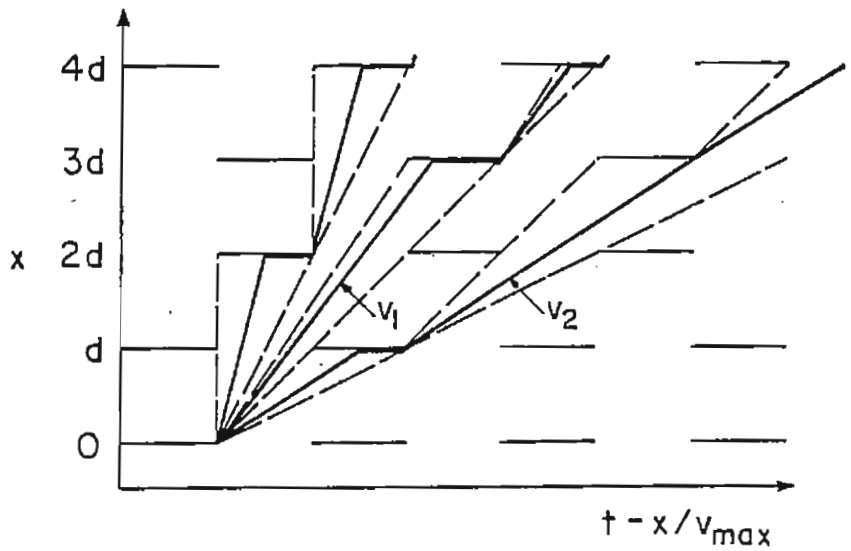


Fig. 3.13 - A strategy for widely spaced intersections.

Figure 3.13 shows a more efficient coordination plan in which the jth signal is set so as to pass these vehicles which were stopped at the j - 1th

Figure 3.13 shows a more efficient coordination plan in which the jth signal is set so as to pass these vehicles which were stopped at the j - 1th signal. (If the offsets between intersections one and two were zero, this scheme is commonly described as "double alternate," but it is used on two-way streets for a quite different reason than described here.). The average delay per vehicle per intersection is only about 0.105C for the plan in figure 3.13 as compared with 0.125C for random coordination and 0.175C for the coordination of figure 3.12.

The above arguments apply only if a significant fraction of the vehicles have a trip time between adjacent intersections which differs from the mean trip

time by approximately G so that vehicles which pass one intersection without delay are spread over approximately a cycle time when they reach the next intersection. Otherwise, as described in part d, only a small fraction of the vehicles are stopped at any single intersection and the efficient scheme of coordination is as in figure 3.12 for some suitable progression speed. The evaluation of delays and strategies for intersection spacings intermediate between the extremes described in parts d and e is rather complex. Obviously there must be some spacings d at which one will abandon the usual progression strategy and switch to a pattern like in figure 3.13 or possibly some "intermediate" strategy.

The theory described here may seem somewhat academic because traffic engineers do not usually worry about signal coordination for signal spacings large enough for the scheme of figure 3.13 to apply (maybe they should). We have already questioned the use of models for platoon spreading for closely spaced intersections on the grounds that the spreading may be considerably less on a signalized arterial with closely spaced intersections than predicted from observations on an arterial with signals far apart. But if the signals are far apart, it is not sufficient to consider only the magnitude of the spreading. One must also consider the possibility that the speeds of vehicles at various positions in the platoon may be different. Even if the platoon should spread so much as to overlap platoons from other signal cycles so as to create a nearly constant flow, one may still find some periodic behavior in the velocity distributions. If one were to interrupt this flow with another signal, the velocity distributions would be relevant to the pattern of coordination downstream of that signal. Two consecutive traffic signals behave like a "velocity filter" which stops vehicles with particular velocities.

We have assumed here that drivers retain their velocities exactly, which implies that the oscillations of the velocity distribution created by a signal

will persist forever. This oscillation will, however, eventually decay because an individual driver cannot maintain his desired speed exactly. It will decay in a distance comparable with that needed for the uncertainty of the trip time of an individual driver with specified desired speed to be approximately a cycle time. The uncertainty of the trip time for an individual driver, however, is expected to be considerably less than the difference in trip times between different drivers with different desired speeds. The distance needed for the velocity oscillation caused by a traffic signal to decay is not known but one might expect it to be of the order of ten miles or more.

3.4. Time Varying Signal Strategies

In most of the theory described so far, it has been assumed that the average number of vehicles to arrive per cycle at any intersection is (nearly) stationary. The desired choice of the cycle time, splits, and offsets, however, were shown to depend on the flows $q_i^{(m)}$ in various directions at each intersection and these flows will vary throughout the day. We do not expect the (average) flows to change very much in a single cycle time so we would expect that any of the proposed strategies based on stationary flows would be applicable during some reasonable period of time (many cycles). There is still some question, however, as to how often one should change the signal coordination and how one should do it.

In practice one should change the signal coordination and how one should do it.

In practice, signal systems are usually designed to operate with only a few different plans, at least one for the morning peak, one for the afternoon, and one for the off-peak. When the signals change from one plan to another, the individual signals will typically lock into a new plan with the next signal switch after initiation of the change. Vehicles which pass a signal in a cycle when the pattern changes are likely to experience some extra delay.

It is obvious from figures 3.1 or 3.2 that one can change the cycle time at will on a one-way arterial provided that one retains the same progression

speed and that any changes in the cycle time propagate downstream from intersection 1 with a time lag x/v . The same would be true for any of the more complex patterns such as in figures 3.6, 3.7, or 3.8. As traffic increases during the rush hour, it would be advantageous to allow the cycle time to increase gradually with a small increase in each cycle to match the increase in flow. One could keep the cycle time constant for many cycles and then change it abruptly, but, if one can do it gradually, one should do so.

As traffic increases, the platoon speed may decrease and one may also wish to decrease the progression speed. Typically one wishes to choose the offsets so as to match the expected trip time of the last vehicle in a platoon. Obviously some complications arise if one tries to make sudden and substantial changes in the offsets. One cannot easily patch together a pattern as in figure 3.1 with another pattern having a different progression speed.

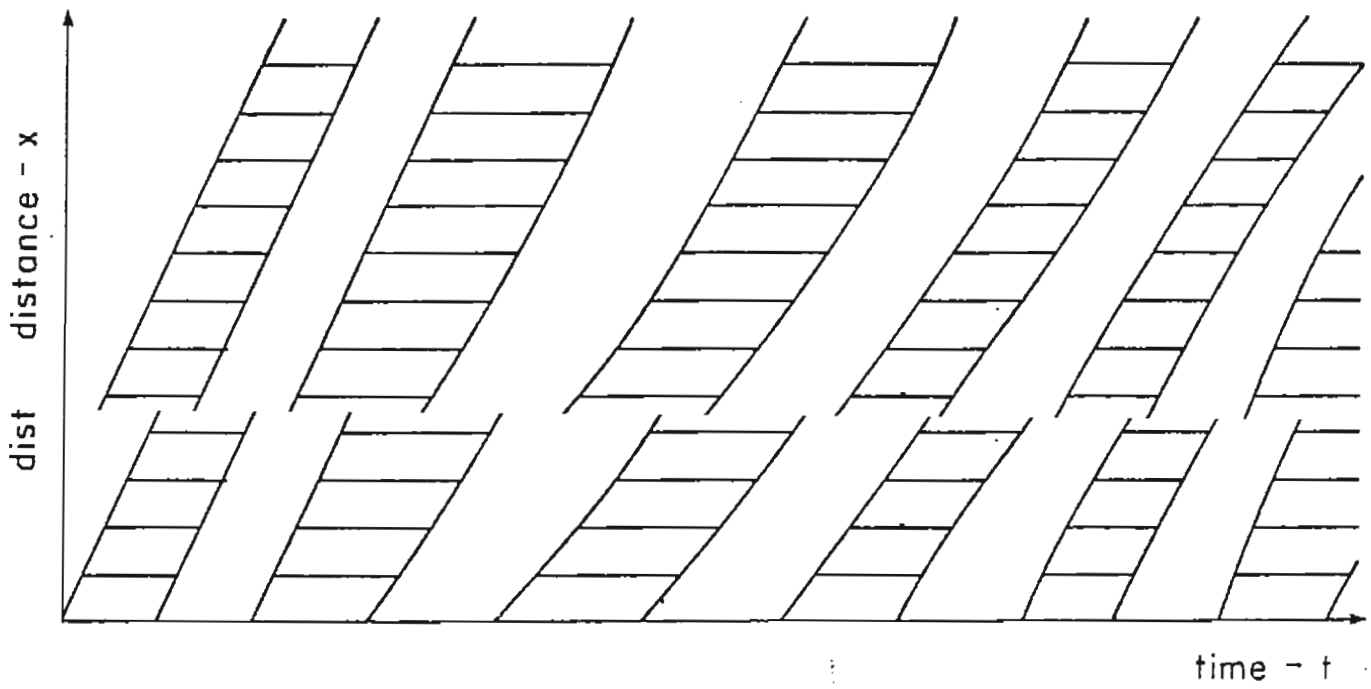


Fig. 3.14 - A continuously changing signal plan.

If all signals are connected to some central control, one could, in principle, switch the signals according to any predetermined time-dependent pattern. Suppose, as illustrated in figure 3.14 (in an exaggerated way), one

were gradually to increase the cycle time at intersection 1 as the traffic increases and then decrease the cycle time as the flow decreases. The ends of the greens at intersections downstream of intersection 1 are set so as to follow the expected trajectory of the last vehicle in each platoon. The starts of the green are chosen to maintain some predetermined split at each intersection.

As the cycle time and the length of the platoon increase, the speed of the platoon may decrease. The trajectories of the last vehicles in succeeding platoons will start to diverge, which means that the cycle times downstream of intersection 1 will be even larger. The longer cycle times downstream, however, will lead to a decrease in the platoon flow and, therefore, an increase in the platoon speed. Thus, the cycle times downstream will not continue to increase but will stabilize. Certainly there is no reason why vehicles in succeeding platoons should continue to move further apart.

Similarly, as the cycle time at intersection 1 decreases as the flow decreases,the length of the platoon decreases and its speed increases. The trajectories of the last vehicles in succeeding platoons will start to converge causing the cycle time downstream of intersection 1 to decrease even faster. This, however, will compress the platoon and cause it to travel slower. Thus, again, the cycle times downstream tend to stabilize. There is no way that an average vehicle in one platoon can overtake a vehicle in another platoon.

again, the cycle times downstream tend to stabilize. There is no way that an average vehicle in one platoon can overtake a vehicle in another platoon.

It is obvious from figure 3.14 that if the flow increases or decreases by only a small amount in each cycle that one can devise a plan of signal coordination in which there is virtually no extra penalty caused by a time varying cycle time and progression speed.

3.5. Traffic Responsive Control

If traffic were sufficiently light on both the arterial and all cross streets that signals were necessary only to avoid occasional conflicts, then

a conventional type of vehicle actuated (V-A) signal would give less delay than any pretimed signal. A V-A signal could be designed so as to give a green light to any approaching vehicle provided that no other vehicle in a cross direction wanted to pass at nearly the same time. A driver on the arterial could travel at any speed he wishes and cause the signals to respond accordingly. A conflict with a cross street vehicle may occasionally cause a delay to vehicles on the arterial but one could, if desirable, program the signal so as to give some preference to the arterial vehicles.

If the flow on the arterial exceeds about four vehicles per minute, however, it is expected that at least one arterial vehicle will wish to pass a signal during the time the arterial flow is interrupted to serve any cross street vehicle. At least two vehicles will pass a signal on the arterial in some cycles. If a V-A signal is designed to respond to vehicles individually, it is possible that the pattern of signal switches could be rather chaotic. Rather than trying to understand how a signal system will respond to two or three vehicles per cycle, it is convenient to consider the opposite extreme of heavy traffic in which the number of vehicles served per cycle is moderately large (perhaps five or more).

a. No turning traffic

In sections 2.6 and 2.7 we saw that it was typically possible to design a

a. No turning traffic

In sections 2.6 and 2.7 we saw that it was typically possible to design a vehicle-actuated control for an isolated signal so as to give considerably less delay than any fixed cycle strategy. We would, of course, hope that one could achieve similar success with some type of traffic responsive control on an arterial or a network. Unfortunately, the strategies that were used for isolated signals cannot, generally, be extended to multiple signals no matter what information one might gather from vehicle detectors, or how one uses these data.

In the idealized geometry of figures 3.1 and 3.2 one could, in principle, switch the signal at intersection 1, or any other intersection, according to

any arbitrary pattern. In particular, at any time t one could switch the signal or not based on any prior information obtained from detectors located anywhere (but not on the times of events in the future), or any historical data such as the average flows $q_i^{(m)}(t)$ observed over many days. Any strategy based only on the $q_i^{(m)}(t)$ would, however, be interpreted as a "pretimed strategy" because one would use the same strategy on all (similar) days. The issue is whether or not any current information from detectors regarding stochastic properties, i.e., how the traffic today differs from the average over many days, can be of any use.

For an isolated signal the saving in delay for a V-A signal, as compared with a F-C signal was due primarily to a combination of two effects. First, one would continue the green (particularly for the direction with the largest s_i) as long as the queue was discharging. It is not useful to pay a penalty for switching the signal if one cannot gain a higher flow. Secondly, one would terminate the green as soon as the queue vanished and the flow dropped from s_i to q_i. To continue the green was, in effect, equivalent to having a larger lost time per cycle. We did see in section 2.7 that there were some potential problems for two-way traffic if the queue in direction 2 should vanish before that in direction 4. If $q_2/s_2 > q_4/s_4$, however, the typical strategy would be to continue the signal for direction 2 until the queue vanishes in direction 2, which should usually be after the queue vanishes in

strategy would be to continue the signal for direction 2 until the queue vanishes in direction 2, which should usually be after the queue vanishes in direction 4.

If we think of the strategy for a signal progression as in figure 3.2 as analogous to a single intersection serving all cross street "simultaneously," the analogue of the above V-A strategy would be to terminate the cross street green when the queue vanishes at the m*th cross street (the busiest), which should usually be later than on any other cross street. The problem with this, however, is more obvious from figure 3.1 than figure 3.2 because "simultaneous"

in figure 3.2 really means with a time lag x/v . Thus, one would like to terminate the cross street green at intersection 1 at such time t that the cross street queue at intersection m^* will vanish at time t plus the uninterrupted travel time from intersection 1 to m^* .

If intersection 1 is the critical intersection ($m^* = 1$), there is no problem. One can operate intersection 1 like an isolated V-A signal and switch all signals downstream accordingly with offsets x/v (but with unequal consecutive cycle times). Otherwise, one must try to predict at time t , from detector records prior to time t or any other relevant information, what will happen at intersection $m^* > 1$ a trip time later. If the critical intersection is a distance x from intersection 1, one would need detectors a distance comparable with x from the arterial on the cross street, and even then, these detectors would not likely give very reliable information regarding when the queue will vanish at the arterial on the cross street.

It is possible to salvage some of the advantageous features of the V-A signal at an isolated signal. To illustrate some possible strategies, we consider first a very idealized situation. Suppose that none of the cross streets are themselves part of a signal progression in the cross direction. They all have uniform arrival rates at flows $q_2^{(m)}$, $q_4^{(m)}$ with $q_2^{(m)}/s_2^{(m)} > q_4^{(m)}/s_4^{(m)}$. Suppose, also, that the arterial flow $q_1^{(1)}$ approaching intersection 1 is stationary, perhaps with a Poisson arrival process, and there is no turning traffic.

have uniform arrival rates at flows $q_2^{(m)}$, $q_4^{(m)}$ with $q_2^{(m)}/s_2^{(m)} > q_4^{(m)}/s_4^{(m)}$.

It is useful here to recall some of the arguments in section 3.2c and 3.2g, that the total delay on the arterial is dictated by the times at which vehicles pass m^* , provided that there is no further delay downstream of m^* . Thus, the delay is insensitive to any change of strategy which does not change the times at which vehicles pass m^* . The existence of signals upstream of m^* certainly cannot reduce the delay on the arterial, so the arterial delay must

be at least as large as if there were no signals upstream of m^*. For a F-C strategy at m^*, we also saw that there were many strategies for the control of signals upstream of m^* which would have no effect on the total arterial delay, including some strategies with a cycle time C at m^* and $C/2$ or $C/3$ upstream of m^*.

If there were no signals upstream of m^*, one could, of course, operate the signal at m^* as an isolated signal, use a V-A strategy, and achieve considerably less delay than for a F-C signal. Furthermore, give the times at which vehicles would leave m^* in the absence of signals upstream of m^*, the presence of signals upstream of m^* would not cause any additional delay if all vehicles arrive at m^* prior to the time they would have left if there were no upstream signals. With this flexibility in strategy, it might seem that one should be able to control the signals upstream of m^* so that vehicles leave m^* as they would if the signal at m^* operated as an isolated V-A signal.

To see what the problem is, suppose that we imagine that all signals between intersections 1 and m^* are removed. The presence of these signals certainly cannot reduce the delay on the arterial. They would only make things more complicated. Also, we anticipate that if we know how to control the signal at intersection 1, the signals between 1 and m^* can be set so as to cause negligible additional delay on the arterial. We might wish to operate the signal at intersection 1, the signals between 1 and m^* can be set so as to cause negligible additional delay on the arterial. We might wish to operate the signal at intersection 1 with 1, 2, or 3 cycles at intersection 1 for every cycle at intersection m^*. The same problems exist with any of these, but for the purpose of illustration, suppose that we wish to choose two cycles at 1 for every cycle at m^*.

Suppose that we had somehow managed to operate the signals prior to some time t_0 at intersection 1 so that a vehicle leaving intersection 1 at time t_0 reaches m^* at time $t_0' = t_0 + x/v$ just as the signal at m^* switches

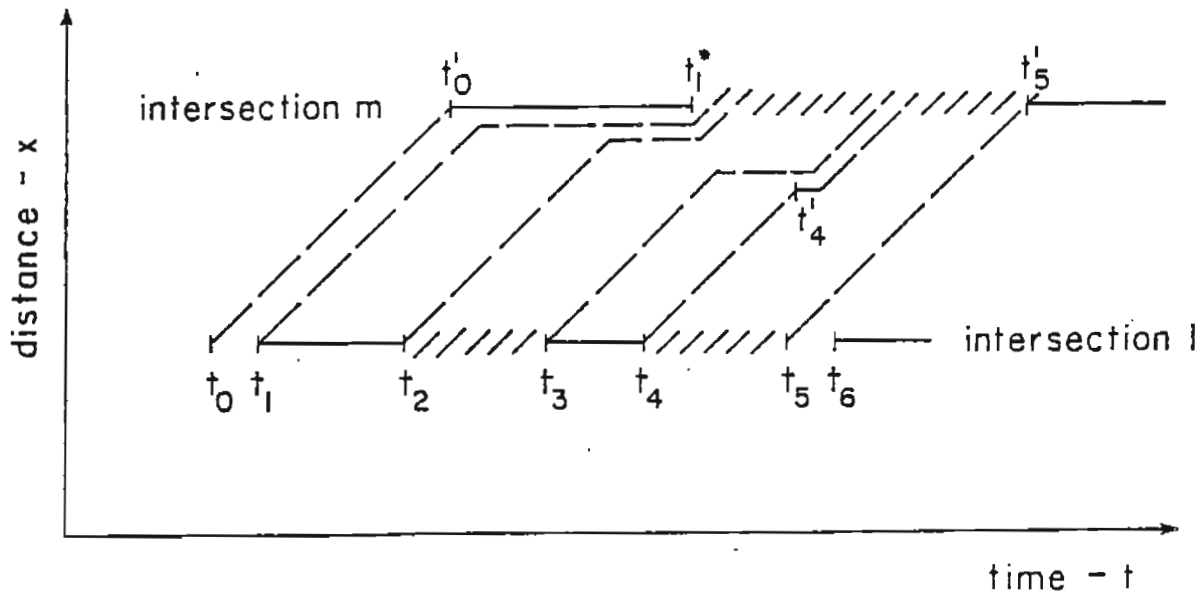


Fig. 3.15 - A traffic responsive signal strategy.

to red, as illustrated in figure 3.15. Actually, one cannot generally identify the time t_0 until one has observed the last vehicle in the platoon pass m* at time t'_0, but presumably the arterial queue at intersection 1 would have vanished by time t_0 since otherwise the green interval at intersection m* would have been extended to accommodate the flow at rate s_1. If one had a vehicle detector at m*, the signal at m* could be programmed to switch when the flow approaching m* drops much below s_1. Thus, one could correctly deter-

detector at m*, the signal at m* could be programmed to switch when the flow approaching m* drops much below s_1. Thus, one could correctly determine the time t'_0 when it occurs, even though one may not be able to predict its value soon enough to identify the time t_0 before the latter occurs.

One must now specify some times t_1, t_2, t_3, t_4, and t_6 when one will switch the signal at intersection 1 based on prior information from detectors at intersection 1 or any other available data. There is some flexibility in these choices, but there are some constraints. In order to keep the signal at m* busy, the time $t'_4 = t_4 + x/v$ should occur not later than the time t'_3 when

the last vehicle to leave intersection 1 before time t_3 passes intersection m* (after being delayed at m*). Also, if one wishes to return at time t_5 to the same type of situation that existed at time t_0, then the time t_6 should occur after time t_5, which in turn should occur after the arterial queue at intersection 1 has vanished. But, of course, one cannot estimate the time t_5 accurately until time t_5' which is expected to occur after one has already chosen the time t_6.

If capacity constraints permitted one to operate signal 1 with two cycles for every cycle at m*, this would mean that the total time $(t_2 - t_1) + (t_4 - t_3)$ needed by the cross street at 1 plus the extra time lost in switching the signal is, on the average, less than the time $t_1^* - t_0'$ needed by the cross street at m*. (Otherwise one would have chosen to operate intersection 1 with one cycle for each cycle at m*, which is certainly possible if m* is the critical intersection). If one were to operate the signal at intersection 1 as an isolated V-A signal, the mean cycle time would, therefore, be less than half the minimum average cycle time at m*. One has some extra time to waste at intersection 1, but, if one were to operate the signal at 1 as a V-A signal, one would use all this extra time in switching losses, at the expense of losing all possibility for coordination with the signal at m*. If by increasing the average cycle time slightly at intersection 1 (at the expense of some increase in delay to the cross street at 1), one can maintain some co-

by increasing the average cycle time slightly at intersection 1 (at the expense of some increase in delay to the cross street at 1), one can maintain some co-

ordination on the arterial, it would obviously be advantageous to do so. Relative to figure 3.15, the time t_6 should be sufficiently large that successive times $t_1 - t_0$, $t_6 - t_5$, etc., do not decrease so as to cause $t_6 - t_5$ or its counterpart in later cycles to become negative and terminate the flow at t_5' prematurely.

Many attempts have been made to program computers to respond to information obtained from vehicle detectors and generate a traffic responsive network

coordination. Most of these programs are designed to minimize the expected delay to vehicles at a single intersection or a single intersection plus its immediate neighbors over some finite time period (usually one or two cycles). Results of such attempts have been rather inconclusive; many have clearly led to a net increase in total delay as compared with some pretimed strategies. The reason for such failures is quite clear even in the present idealized situation. Clearly, the "optimal" strategy for control of intersection 1 does not depend just on its neighboring intersections; it depends mainly on the traffic conditions at m^* which may be far away. Also, what is "optimal" over a short time, one or two cycles, may lead to poor signal coordination over a longer time.

If intersections 1 and m^* are sufficiently close together, it would be possible to devise some very efficient "feed-back" control strategies. The key problem is to estimate the time t_5 so that one can devise a strategy at signal 1 which is compatible with what one would like to happen at m^*. If one knows the past history of arrivals at intersection 1, it is actually possible to estimate the time t_5 at any time after t_1^* (rather than t_5'). At any time after time t_0' one could keep an accurate count of the number of arterial vehicles which have arrived at intersection 1 since time t_0. Since all these vehicles should pass m^* by time t_5' at an average flow s_1 starting at time t_1^*, one can estimate the time $t_5' - t_1^*$ needed for the queue to vanish at m^*, and from should pass m^* by time t_5' at an average flow s_1 starting at time t_1^*, one can estimate the time $t_5' - t_1^*$ needed for the queue to vanish at m^*, and from this estimate the time t_5, for any value of t_1^*.

In order for this estimate of t_5 to be of much use, however, its value should be larger than t_1^* or, at least, any proposed choice of t_6 should be larger than t_1^*, because one cannot take any action based on the estimate until one has made the estimate at time t_1^*. Essentially, the condition for this to be so is that the typical green interval at m^* is larger than the trip time from 1 to m^*. If true, then one can choose the time t_6 based on the observed

traffic on the cross street prior to t_1^* and the observed traffic on the arterial prior to t_5. In particular, one could choose the time t_1^* so that the cross street queue vanishes at m^*. Vehicles would then leave m^* as if it were an isolated V-A signal.

One still has some options in choosing the times t_1 to t_4. There is some excess time to waste at intersection 1 and it can be distributed between the cross street and the arterial. Ideally, one should give all the excess time to the cross street because this would reduce the delay on the cross street with presumably little effect on the arterial delay. The benefit to the cross street, however, would be quite small and to implement this one would want to choose $t_6 = t_5$ and $t_1 = t_0$, and control the time t_4 so that the arterial queue vanishes at time t_5. Actually, it would suffice to choose t_4 so that the first vehicle to leave intersection 1 after time t_4 would arrive at m^* just as the last vehicle to leave intersection 1 at time t_3 passes m^*. The time t_5 could then be identified as the time at which the arterial queue vanishes at intersection 1. One cannot estimate the time t_4, however, until one has identified the time t_1^*. This is, of course, a much more severe restriction than the previous condition that t_5 be larger than t_1^*. Since little is to be gained by giving the excess time to the cross street, a safer strategy for $t_5 > t_1^*$ would be to switch the signals when the cross street or arterial queue vanishes at times t_2, t_3, and t_4 but then after time t_4 extend the arterial $t_5 > t_1^*$ would be to switch the signals when the cross street or arterial queue vanishes at times t_2, t_3, and t_4 but then after time t_4 extend the arterial green after the queue vanishes until the estimated time t_5.

Since the situation described here is so idealized, we will not try to give further details of how one would implement the above strategies. The main point is that in order to devise an efficient signal strategy which is responsive to the fluctuations in the cross street traffic at m^*, one must be able to feed information from the cross street back to intersection 1 early enough so that one can take some effective action. What one is trying to do is to increase or

decrease the cycle time in response to any stochastic fluctuations in the number of vehicles to arrive on either the cross street at m^* or the arterial at intersection 1.

If the distance between intersection 1 and m^* is so large that the trip time exceeds the typical green interval at m^*, any observations of stochastic fluctuations in the cross street traffic at m^* occur too late to permit any immediate control of the signal at intersection 1. There are, however, a variety of strategies involving a delayed response, depending on the distance between intersection 1 and m^*. If, for example, there should be an excess of cross street arrivals at intersection m^* during some cycle, one may choose either to serve them immediately at the expense of delaying the arterial platoon which is already approaching intersection m^*, or force the excess vehicles to wait in queue.

In subsequent cycles one is equally likely to have a deficiency or another excess of arrivals on the cross street. In the former case the previous disturbance will dissipate with no further damage, but in the latter case the situation will get worse. It would, therefore, be advantageous to increase the cycle time at intersection 1 to provide some excess capacity so as to guard against the latter possibility even if the response is delayed (unless the delay is so large that the damage has already been done before one can take any corrective action). Strategies of this type may, however, be too complicated and too restrictive to be of much practical use (even where they might apply).

Another possible type of strategy with no feed back control would be to respond to the observed fluctuations in the arterial arrivals at intersection 1, give the cross street some excess green to absorb fluctuations in the traffic at m^*, but then drive the signals downstream of 1 with predetermined off-sets. In figure 3.15, for example, for any choice of $t_2 - t_1$, one could choose t_3

as the time when the arterial queue vanishes. The expected number of cross street vehicles to arrive at m^* during a time $t_3 - t_2 + L$ is $q_2^{(m^*)}(t_3 - t_2 + L)$ and the expected time needed to serve them is $q_2^{(m^*)}(t_3 - t_2 + L)/(s_2^{(m^*)} - q_2^{(m^*)})$.

If one gave the cross street at m^* barely enough time to serve the average number of arrivals, the queue at m^* would behave approximately as (actually worse than) a saturated F-C signal, so one should give the cross street a little extra time as one would for an isolated F-C signal. This would then specify the time $t_4 - t_3$. The t_5 is now chosen as the time when the arterial queue vanishes again, etc.

At intersection m^* one now chooses the $t_1^* - t_0$ as $(t_4 - t_3) + (t_2 - t_1) + L$, $t_0' = t_0 + x/v$, and $t_5' = t_5 + x/v$. Presumably, the time $t_5' - t_1^*$ is now sufficient to accommodate any arterial vehicles which pass intersection 1 between times t_2 and t_3 or between t_4 and t_5 at flow s_1 , and $t_1' - t_0'$ is sufficient to accommodate the expected number of arrivals on the cross street at m^* during the previous arterial green plus some fluctuations.

The above strategy may not be the "optimal" strategy, but it certainly could be designed to give less delay than any pretimed strategy. For a pretimed strategy one must choose a cycle time to balance the deterministic queues against the stochastic queues on both the arterial (at intersection 1) and the cross street at m^* . For this modified strategy, there is no stochastic queue against the stochastic queues on both the arterial (at intersection 1) and the cross street at m^* . For this modified strategy, there is no stochastic queue on the arterial, but the green interval for the cross street must be chosen to balance the deterministic queues against the stochastic queue on the cross street at m^* . The main improvement comes from the fact that whenever there is a deficiency of arterial arrivals at intersection 1, one does not waste time continuing the arterial green after the queue is gone. Also, the cycle time will increase if there is an excess of arterial arrivals. One does not, however, respond to fluctuations in the cross street traffic.

b. Variations in trip time

Most of the discussion in part a dealt with strategies to respond to fluctuations in the number of vehicles to arrive at an intersection during any cycle. It was implied, however, that the trip times of vehicles between intersections were the same for all vehicles, and known. There will also be stochastic effects resulting from the fact that the trip times are not exactly predictable. A traffic responsive signal system is capable of responding to variations in the trip times, but it is not obvious that it should.

In section 3.3 we discussed some modifications in strategy that one might employ in response to platoon spreading, which might be the result of "stochastic effects" in the sense that the spreading may be due to differences in the speeds of different drivers. We are not so much concerned now with the question of how one should respond to a predictable behavior of the platoon (including possible platoon spreading), but with how one should respond to differences in the platoon behavior from one cycle to the next.

As a platoon of vehicles moves downstream from some intersection, one certainly should not cater to the desires of the lead vehicle in the platoon; one should look at the last vehicle. The trip time of the last vehicle is not expected to vary very much from one cycle to the next, but if the last vehicle should travel faster than expected, this would mean that the flow past some intersection (particularly m^*) would drop sooner than expected.
should travel faster than expected, this would mean that the flow past some intersection (particularly m^*) would drop sooner than expected.

If the (average) cycle time has been chosen so that there is usually a residual queue on the cross street at m^*, it would certainly be advantageous to terminate the arterial green at m^* as soon as the last vehicle in the platoon passes, because any extra time could be used to reduce the cross street queue. Whether or not one should do so at other intersections is less important. If there is no overflow queue on the cross street, giving more time to the cross street is only marginally worthwhile.

It is not obvious, however, that one should extend the arterial green interval if the last vehicle in the platoon is traveling slower than expected. If drivers know the predetermined progression speed, they should try to keep up with it under the threat that they will be cut off if they don't. If a driver knows that the green will be extended, even if he makes only a minimum effort to stay close to the vehicle ahead, he has no incentive to stay close. If, however, the last vehicle in a platoon is slow because the platoon is too long, it may be advantageous to shorten the platoon by pushing the last vehicle into the next cycle.

c. Turning traffic

In part a we discussed the possibility of controlling a signal at m^* in response to fluctuations in the arterial flow entering the arterial at intersection 1. The strategy was based primarily on the assumption that any vehicle which could pass intersection 1 during some green interval could also pass intersection m^* in the same length of green interval, but it exploited the fact that one could vary the cycle time at intersection 1 at will and drive all signals downstream accordingly.

If, however, one has a long arterial, many of the vehicles which pass m^* may have turned onto the arterial from cross streets upstream of m^*. Whereas in part a we had difficulty varying the cycle time at intersection 1 in response to fluctuations in the cross street traffic at m^*, it now becomes difficult in part a we had difficulty varying the cycle time at intersection 1 in response to fluctuations in the cross street traffic at m^*, it now becomes difficult to vary the cycle time in response to fluctuations in either the cross street traffic or the arterial traffic. In order to maintain any coordination of the signals, it would seem that the best one could do is to allow individual signals to make certain deviations from some pretimed strategy.

There must be a first intersection somewhere. Even though it may be so far from the critical intersection m^* that most of the vehicles passing m^* entered the arterial from cross streets between 1 and m^*, the choice of a cycle

time at intersection 1 affects the traffic behavior at intersection 2, which affects that at intersection 3, etc. One must choose a cycle time at intersection 1 based upon the average traffic behavior downstream. In particular, the (average) cycle time C at m* should be chosen so that intersection m* is undersaturated (if possible) and the (average) cycle time at intersection 1 should be chosen as C, C/2, or C/3 . Although one would like the vehicles to pass m* at the times they would pass if intersection m* operated as an isolated V-A signal, this is impossible. By the time one has observed the fluctuations in the number of vehicles per cycle approaching m* , it is too late to change the periodicity of the approaching platoons.

Presumably, having chosen the cycle time at intersection 1 (based on the expected traffic at m*), there is some extra time to distribute between the cross street and the arterial. There may be vehicles approaching intersection 1 after the queue vanishes in either direction. How one partitions any excess time is somewhat arbitrary because it will not have much effect on the average delays in either direction.

In the pretimed signal strategies discussed in sections 3.2j and 3.3c, one may have given extra green time to the arterial at intersections 2, 3, etc., to accommodate a certain amount of platoon spreading and/or turning traffic, if these intersections also had ample excess time to distribute between the ar-

accommodate a certain amount of platoon spreading and/or turning traffic, if these intersections also had ample excess time to distribute between the ar-

terial and the cross street. Of course, there is no reason to give the arterial extra green time which no one can use. If the pretimed strategy allowed for some <u>average</u> amount of turning traffic or platoon spreading, but it did not actually occur, then there is no reason to start the arterial green before any vehicles arrive in the platoon. Also, there is no reason to continue the arterial green after the last platoon vehicle passes if this occurs before the scheduled time to terminate the green. Giving any unused time to the cross street may be of some benefit to the cross street traffic even if there is no overflow queue on

On the other hand, if more than the expected number of vehicles turned onto the arterial at intersection $m - 1$ and were waiting for the signal at intersection m, or the lead vehicle in the platoon from intersection $m - 1$ arrives at m earlier than expected, one might wish to start the arterial green at m a bit earlier than planned. The benefit from doing this, however, may be quite small because sooner or later the platoon will be compressed anyway when it reaches some more congested intersection m' or m^*. Also, at the end of the green, it is not obvious that one should extend the arterial green for a straggling vehicle which is not keeping up with the platoon. Certainly one would not give any extra time to the arterial (at either the beginning or the end of the green) if this would cause an overflow of the queue on the cross street (but not on the arterial).

The strategy for control is somewhat arbitrary until the arterial platoon approaches some moderately congested intersection m' or m^* at which there may be queueing on the cross street. If one had allowed the platoon to spread upstream of m' so as to have an average flow less than s_1, one will want to compress the platoon. Although one may have made some deviations from the pretimed plan upstream of m', one probably had not allowed the arterial green to continue beyond its pretimed value and one would have terminated the green early only if the platoon had terminated early. Thus, one does know approxi-

mately the time at which the last vehicle in the platoon will reach m'.
early only if the platoon had terminated early. Thus, one does know approximately the time at which the last vehicle in the platoon will reach m'.

There may, however, be some uncertainty in how much one needs to compress the platoon in order for it to pass m' without delaying the last vehicle in the platoon, if one can.

If there is no overflow queue on the arterial or the cross street at m' and the approaching platoon can pass m' during the scheduled time at flow s_1, it will not make much difference how one uses any excess time, i.e., how

much one compresses the platoon. The important question is what one should do otherwise, i.e., if the use of the pretimed strategy will (or has) caused a queue to form on either the arterial or the cross street.

There are obvious ways in which one can make improvements on any pretimed strategy. What is "optimal", however, depends on the weights that one attaches to the delays on the arterial relative to delays on the cross streets and to long delays to a few vs. short delays to many. We will consider the merits of a few possible strategies.

d. Modification of a pretimed plan

In any pretimed signal strategy the traffic on the cross street or the arterial responds to the signal, but each is otherwise independent of the other. Whether or not there is an overflow queue on any cross street would, therefore, be independent of whether or not there was one on the arterial. The stochastic queues on different cross streets would likely be statistically independent of each other but, of course, the arterial queues at different intersections would be highly correlated.

At any intersection there is no reason to continue the arterial green after the last platoon vehicle has passed. Any extra time might be profitably used by the cross street even if there is no stochastic queueing on the cross street. At any intersection, however, at which one has compressed the platoon to a flow s_1 in order to avoid possible queueing on the cross street, one could squeeze the arterial traffic a little harder. If the flow on the arterial should drop appreciably below s_1, this could be interpreted to imply that there is no queue on the arterial at this intersection and that any following vehicles are stragglers. Particularly if there was an overflow queue on the cross street from the previous cycle, it would be advantageous to terminate the arterial green as soon as the platoon passes and give the extra time to the cross street.

If one uses this strategy at both intersections m' and m^* $m' < m^*$, one should also consider the consequences of any action at m' on the delays at m^*. If there is an overflow queue on the arterial at m^* and one should cut off any stragglers at m', there would be no penalty for doing so because the vehicle which is delayed at m' would have been delayed at m^* anyway. If there is no overflow queue at m^*, however, terminating the green early at m' would likely mean that one would also terminate the green early at m^* with potential benefits to the cross streets at both m' and m^*. Note, also, that if m' were to operate on a cycle time $C/2$ and m^* on a cycle time C, terminating the arterial green early at the time corresponding to t_3 in figure 3.15 would have no effect on the arterial delay.

If it is advantageous to terminate the arterial green early if the queue vanishes, one should also consider the advantages of terminating the cross street green early. We have already argued, however, that, if there is no overflow queue on the arterial, one should give to the cross street any time not needed by the arterial. One will give extra time to the arterial only if releasing the arterial platoon early will decrease the trip time of the last vehicle in the platoon.

Indeed, there are some risks associated with releasing the arterial vehicles too soon if the signal on the arterial is programmed to terminate the green when there is a drop in flow. If by advancing the start of the arterial green at intersection m to serve any vehicles which turned onto the arterial at $m - 1$, one will clear these vehicles before the arrival of the platoon from $m - 1$, the signal might misinterpret the drop in flow as the end of the platoon and terminate the arterial green before the platoon arrives. This would be particularly risky in the situation illustrated in figure 3.15. If one started the green early at intersection m^* at time t_1^*, there is a danger that the vehicles of the first platoon from m' will clear before the second platoon arrives.

It would not be very difficult to program signals to avoid the above risk. Any signal controller could record, for example, whether or not the arterial queue vanished during the previous green interval or even count the number of vehicles in the residual queue between the detector and the intersection. If the cross street queue should then vanish early during the following cross street green phase, the extra time could be used to serve the residual queue on the arterial.

This modification in strategy would obviously be advantageous at intersection m^*. The pretimed strategy at m', $m' < m^*$, however, is likely to be such that, if there is an overflow queue on the arterial at m', there is almost certainly one also at m^*. Giving more green time to the arterial at m', therefore, will only shift the arterial queue from m' to m^* with no net benefit. For intersections downstream of m^*, the pretimed strategy is likely to be such that there is seldom a queue downstream. If there is one, however, it would be advantageous to advance the start of green downstream of m^* whenever the start of green is advanced at m^*.

Since the above modifications can be applied to <u>any</u> pretimed signal plan, there is a very large class of strategies to consider. If, however, the pretimed strategy is such as virtually to guarantee that there is a queue at m^* whenever there is a queue upstream, and that the queues downstream of m^* are negligible, the delays in the system will be dominated by those at or caused by whenever there is a queue upstream, and that the queues downstream of m^* are negligible, the delays in the system will be dominated by those at or caused by intersection m^*. Furthermore, if most arterial vehicles passing m^* turned onto the arterial at various intersections upstream, the stochastic delays on the arterial (and the cross street) will be essentially those which would exist if the number of vehicles to arrive at m^* in a single cycle are Poisson distributed (but pulsed).

One can estimate the consequences of these modifications by recognizing

that whenever there is an overflow queue on either the arterial or the cross street (or both), the intersection will keep busy for the whole cycle time, except for the fixed lost time L for switching. In this sense the signal behaves as if it were serving just one traffic stream and one queue. Indeed, if $s_1^{(m^*)} = s_2^{(m^*)} = s$, it will serve traffic at the same rate s in either direction 1 or 2 for a combined time of $C - L$ per cycle. For a given degree of saturation for the combined traffic

$$\rho = \left[(q_1^{(m^*)} + q_2^{(m^*)})/s\right](1 - L/C)^{-1} , \tag{3.5.1}$$

the average stochastic queue (for the combined traffic) will be essentially that described in section 2.3 for a F-C signal serving arrivals which are Poisson distributed with mean $(q_1^{(m^*)} + q_2^{(m^*)})C$ per cycle. For sufficiently small $1 - \rho$, the average (combined) queue would be approximately

$$I/2(1 - \rho) . \tag{3.5.2}$$

for $I_1 = I_2 = I$.

Without this traffic responsive modification, the queues on the arterial and the cross street would depend on the pretimed split but would be statistically independent of each other. If the splits were chosen so that the degree of saturation in the two directions were equal, the queues in both directions would be approximately as in (3.5.2) for the same ρ. Thus, for $s_1^{(m^*)} = s_2^{(m^*)}$ of saturation in the two directions were equal, the queues in both directions would be approximately as in (3.5.2) for the same ρ. Thus, for $s_1^{(m^*)} = s_2^{(m^*)}$ and this split, this modified strategy reduces the combined average queue by about 1/2 (actually for $\rho < 1$, the reduction is even more than this).

As pointed out in section 2.3, however, the optimal split for a pretimed strategy is not $\rho_1 = \rho_2$. Any excess time should be distributed as in (2.3.15). If one chooses the optimal split for the pretimed strategy, the traffic responsive strategy reduces the total stochastic delay by a factor of only about

$$\left[1 + \frac{2}{(q_1/q_2)^{1/2} + (q_2/q_1)^{1/2}}\right]^{-1} .$$

For $q_1 = q_2$ the optimal split is indeed $\rho_1 = \rho_2$ and this factor is 1/2. For $q_1 \neq q_2$ the improvement is not this large, but it is not very sensitive to q_2/q_1. For $q_2/q_1 = 0$ there is no improvement (the two strategies are essentially the same), but even for $q_2/q_1 = 1/4$ (a rather extreme case) this factor is $5/9 = 0.55$, only slightly larger than 1/2.

If one were to use some strategy with a pretimed maximum green interval for both directions 1 and 2 (G_{M1} and G_{M2}), but terminated either green if the queue vanishes before the time G_{M1} or G_{M2} expires, the signal will still be kept busy whenever there is a residual queue in either direction. If $s_1^{(m*)} = s_2^{(m*)} = s$, the average combined queue would still be given by (3.5.2) independent of the G_{M1} or G_{M2}. The distribution of the queue between directions 1 and 2, however, is quite sensitive to G_{M1} and G_{M2}. If, for example, one chose G_{M1} slightly larger than $(q_1^{(m*)}/s_1^{(m*)})C$ so that most of the excess time is assigned first to direction 1, the queue in direction 1 will stay relatively short; most of the queueing will be in direction 2. The partition of the total queue between directions 1 and 2 is a rather complex function of the G_{M1} and G_{M2}.

With the queues at intersection $m*$ reduced to about half what they would be for an isolated F-C signal (for $s_1^{(m*)} = s_2^{(m*)}$), one can now reconsider the

With the queues at intersection $m*$ reduced to about half what they would be for an isolated F-C signal (for $s_1^{(m*)} = s_2^{(m*)}$), one can now reconsider the choice of C. Since the choice of C is based mainly upon a balance between deterministic and stochastic queueing at the intersection $m*$, reducing the stochastic queue by about 1/2 would mean that one could afford to reduce C. From the formulas of section 2.3 for an isolated F-C signal, one could infer that a reduction of the stochastic queue by 1/2 would lead to about a 15 percent reduction in optimal C and about a 25 percent reduction in total delay. Actually the 15 percent reduction in C does not affect the total delay very much. One should obtain about a 25 percent reduction in total delay even if one does

not change C. This represents a rather substantial improvement considering that the installation of computerized signal systems have typically given only about a 10 to 15 percent reduction in delay.

We saw in section 2.6 that a V-A signal at an isolated intersection would typically give about half the delay of an optimal F-C strategy, i.e., a 50 percent reduction in delay. As one might expect, the above modification of the F-C strategy, which eliminates wasted time when a queue vanishes early but does not allow the cycle time to drift, gives a delay about half way between that of the F-C and V-A strategies at an isolated signal.

The above conclusions are based on an implied assumption that the effective lost time L for the traffic responsive strategy would be nearly the same as for a pretimed strategy. As was discussed previously in section 2.5 for the isolated signal, for this to be true, it is necessary that vehicle detectors be located in such a position that the signal can switch from green to yellow when the last platoon vehicle is still about two seconds trip time from the intersection so as to maintain a flow close to s_1 or s_2 for about two seconds into the yellow (as it would be for most cycles in a pretimed strategy). This is particularly important here because the flow at intersection m^* is likely to drop nearly to 0 after the last platoon vehicle passes (rather than to q_1 or q_2 as at an isolated signal).

If the cross streets, but particularly the cross street at m^*, are not part of some progression system in the cross direction, one would probably choose a maximum green for the arterial G_{M1} sufficiently large that most of the stochastic queue is on the cross street (whether or not such a strategy will actually minimize the total delay). If, however, the cross street at m^* is itself part of a progression system in the cross direction, one would probably choose to distribute any stochastic queues equitably between the two directions. If m^* is the critical intersection for the arterial in question, it is probably

also a critical intersection on the cross street progression.

For the arterial there will be a common cycle time C for all intersections and, when there is no residual queue at m^* from the previous cycle, one would like to start the arterial green at m^* according to a pretimed plan compatible with the arrival time of the approaching platoon. The same would also be true of the cross street. Thus, if there is no residual queue in direction 1, the cross street green should terminate no later than some prespecified time; and if there is no residual queue in direction 2, the arterial green should terminate no later than some prespecified time.

If the cycle time has been chosen so as to balance the deterministic and stochastic queues, it will be quite rare for the queue to vanish in both directions 1 and 2 in the same cycle. When this happens, one has some time to waste and it makes little difference where the excess time is assigned. A more common event would be that the queue vanishes in direction 1 before the specified time or the queue vanishes in direction 2 before the prespecified time (but not both). In the traffic responsive modification of the pretimed plan, whenever the queue vanishes in direction 1 before the prespecified time any excess time would be assigned to direction 2 to reduce the queue in direction 2 and vice versa.

The benefit from such a strategy will still be essentially as described above (at least for $s_1^{(m^*)} = s_2^{(m^*)} = s$); namely, a reduction of the stochastic queues at m^* by approximately 1/2. The argument here is a repetition of that used previously for the arterial itself. If m^* is the critical intersection on the arterial and the signals upstream of m^* are set so that the queue will not vanish at m^* unless it also vanishes at upstream intersections, then the (residual) queue at or caused by m^* on the arterial will be the same as if there were no intersections upstream and vehicles arrived according to a Poisson process. But if the arterial is also the critial intersection on a progression system for the m^*th cross street, one can apply a similar argument for the

queue behavior for direction 2 on m^* and say that the residual queue at or caused by intersection m^* behaves as if there were Poisson arrivals in both directions. For this to be true, one may need to provide a little extra green time to the arterial at intersections upstream of m^*, and similarly on the m^*th cross street, and, if one advances the start of the arterial green at m^*, one may choose also to advance the start of green at intersections downstream of m^*.

The above estimates of the savings in delay were based on an assumption that $s_1^{(m^*)} = s_2^{(m^*)}$. If the cross street at m^* is part of a progression system, one would have chosen the pretimed split to adjust for possible differences in $s_1^{(m^*)}$ and $s_2^{(m^*)}$ (particularly if the cross street was a two-way street with queueing predominantly only in one of the two directions). The saving in delay for a traffic responsive modification of the plan would be difficult to estimate, but we expect them to be comparable with those described above. If, however, the cross street is not part of a progression system, then it is likely that $s_1^{(m^*)} > s_2^{(m^*)}$, since a one-way arterial is likely to have at least two lanes of traffic in direction 1 and the cross street may have just one lane for direction 2 and one for direction 4. In this case, one might choose the pretimed split so that most (all) of the queueing is in direction 2.

It is easy to estimate the queueing in direction 2 if there is negligible queueing in direction 1. Suppose, for example, that direction 2 received zero scheduled time but obtained any time remaining after the platoon passes in direction 1. Suppose, also, that the off-sets in direction 1 are chosen so as to compress the platoon in direction 1 to a flow $s_1^{(m^*)}$ at m^* and that the arterial green terminates promptly when the flow drops below $s_1^{(m^*)}$. If the system is undersaturated, an arterial platoon should almost certainly pass m^* within a time $C - L$, so there should be no overflow queue on the arterial and the cross street should obtain some time every cycle. The cross street

green will terminate and the arterial green will start according to the pretimed plan with fixed cycle time C.

With this strategy, the (overflow) stochastic queue on the cross street behaves as if it has Poisson arrivals at rate $q_2^{(m^*)}C$ arrivals per cycle, but a random green interval. The green interval would be $C - L$ less the time needed to serve, at rate $s_1^{(m^*)}$, a Poisson distributed number of vehicles on the arterial with mean $q_1^{(m^*)}C$ per cycle. If one neglects the variance in the departure headways (i.e., assume $I_1 + I_2 = 1$), one can show that the average queue at m^*, for ρ sufficiently close to 1, is approximately

$$\text{average queue} = \frac{\left(q_2^{(m^*)}/s_2^{(m^*)}\right) + \left(s_2^{(m^*)}/s_1^{(m^*)}\right)\left(q_1^{(m^*)}/s_1^{(m^*)}\right)}{2(1-\rho)(1-L/C)}$$

$$\cong \frac{1}{2(1-\rho)}\left[1 - \frac{q_1^{(m^*)}}{s_1^{(m^*)}}\left(1 - \frac{s_2^{(m^*)}}{s_1^{(m^*)}}\right)\right], \tag{3.5.3}$$

with

$$\rho = \frac{q_1^{(m^*)}/s_1^{(m^*)} + q_2^{(m^*)}/s_2^{(m^*)}}{1 - L^{(m^*)}/C} \quad . \tag{3.5.4}$$

For $s_1^{(m^*)} = s_2^{(m^*)}$, (3.5.3) gives the same average delay as in (3.5.2) since, in this case, the total delay is (nearly) independent of how one partitions the cycle time between the arterial and the cross street. If $s_2^{(m^*)} >$

since, in this case, the total delay is (nearly) independent of how one partitions the cycle time between the arterial and the cross street. If $s_2^{(m^*)} > s_1^{(m^*)}$, the average queue (3.5.3) is larger than (3.5.2) but for $s_2^{(m^*)} < s_1^{(m^*)}$ it is less, which confirms our expectation that one should give priority to the direction with the larger saturation flow.

Perhaps the most interesting aspect of (3.5.3) is that the average queue on the cross street for $s_1^{(m^*)} > s_2^{(m^*)}$ is even less than it would have been for a F-C strategy with degree of saturation ρ for the arterial and the cross street. This is true despite the fact that, with this modified strategy, there is no overflow queue on the arterial and the cross street

queue must absorb the fluctuations in both the cross street and the arterial traffic. The reason for this is that the average arterial green interval is actually less than it would be for a F-C strategy. The extra time which one would give to the arterial traffic in a F-C strategy to absorb fluctuations in the arterial traffic is given to the cross street.

If, for example, $q_1^{(m^*)}/s_1^{(m^*)} = q_2^{(m^*)}/s_2^{(m^*)}$ so that one would use equal green intervals on the arterial and the cross street for a F-C strategy and $s_2^{(m^*)}/s_1^{(m^*)} = 1/2$ (two lanes for the arterial for one on the cross street, in direction 2), the numerator in (3.5.3) is only 3/4, i.e., the cross street queue is only 3/4 that for a F-C strategy and the total queues on both the cross street and the arterial is only about 3/8 times that of a F-C signal.

If, despite the fact that the cross street traffic has less delay than for a F-C strategy, one objects to assigning all the delays to the cross street, one can, of course, modify other pretimed plans in which the cross street is given some guaranteed minimum green interval whenever there is a queue.

In the above discussion we assumed that there were possible overflow queues on the cross streets only at intersection m^*. If queues might exist also at the other intersections, one may wish to modify the above strategy. If a cross street queue might form at intersection m'' downstream of m^*, terminating the arterial green at m'' when the flow drops below $s_1^{(m'')}$ will, as at m^*, give more average times to the cross street than any pretimed plan. The average cross street queue at m'' should be appreciably less than at m^* (otherwise m'' would have been the critical intersection).

If there should be an overflow queue on the cross streets at both m^* and m'' in the same cycle (displaced by the transit time), the combined cross street flow at m^* and m'' when both queues are discharging is $s_2^{(m^*)} + s_2^{(m'')}$. If this is larger than $s_1^{(m^*)}$, it might be advantageous to give some preference to the cross street. Note that, if we terminate the green early at m^*, any

arterial vehicle delayed at m^* would not suffer the same delay again at m''. The competition is, therefore, between the arterial flow at m^* and the combined cross street flows at both m^* and m''. One might, in this case, wish to terminate the cross street green when the cross street queue vanishes at m'', but it would be difficult to judge at intersection m^* when one should terminate the cross street green at m^* so that the queue vanishes at m'' a transit time later. If, however, queueing at m'' is a possible issue, one could use a modified pretimed plan which gives a predetermined minimum green to the cross street (and thus a predetermined maximum green to the arterial). One would impose these restrictions at intersection m^*, however, so as to give extra time to the cross streets at both m^* and m''.

If a queue might form at an intersection m' upstream of m^*, one has more options because there is likely to be a cross street queue at m^* (a transit time later) whenever there is one at m'. If $s_2^{(m')} + s_2^{(m^*)} > s_1^{(m^*)}$, it may be advantageous to continue the cross street green at m' until the queue vanishes, even if this occurs after some pretimed end of the green. The following arterial green should then terminate no later than the prescheduled time even if this causes an overflow queue on the arterial. The cross street green at m^* will now also be extended by the same amount of time as at m' since the arterial platoon will arrive at m^* this much later. The arterial platoon at m^* which was probably cut off early at m' will likely pass m^* by the scheduled time. If not, the green should probably be terminated at the scheduled time anyway to make certain that the signal switches stay locked in to the scheduled cycle time C.

The strategies of dealing with queues at several intersections simultaneously are not expected to give major improvements over some simpler strategies. There are many possible variations on the general scheme of starting from

some pretimed signal plan but then avoiding any waste of time serving one traffic direction when the time can be profitably used in the other direction. Having done so, the total delay is actually not very sensitive to the details of the original signal plan. Also, what is "best" may involve question of equity as well as total delay.

3.6 Summary

Despite the large number of parameters associated with the description of flows and strategies for a one-way arterial, there are some fairly simple strategies which are certain to give considerably less delay than the common procedure of choosing some arbitrary cycle time for all signals and off-sets equal to the (average?) trip time between intersections, or any of the currently used attempts at a traffic responsive system.

Whether one uses some traffic responsive system or not, one should first devise an efficient pretimed plan in the following steps (for moderately heavy arterial traffic and intersection spacing typical of an urban arterial).

1. Identify the most critical intersections. Formally, these are the intersections which require the largest cycle time according to the condition (3.2.2). For any existing system, however, one should know ahead of time which ones they are. They are the ones where queues sometimes form or for which the green intervals are almost fully utilized.

2. Choose a cycle time C and split for the most critical intersection. green intervals are almost fully utilized.

2. Choose a cycle time C and split for the most critical intersection. We assume that there exists some C such that this intersection is undersaturated. If not, then there is not much one can do except to reroute the traffic (see section 3.2k). If the system is undersaturated, the choice of C and the split necessarily involves a trade-off between stochastic and deterministic queueing on both the arterial and the cross street, at least at the critical intersections. The choice here is somewhat arbitrary because the objective is not clear; but as a preliminary choice, one could choose a C comparable

with twice the minimum C which satisfies (3.2.2) and a split as one would if the critical intersection were an isolated intersection.

3. For the proposed value of C (or perhaps a somewhat larger value), identify which other intersections could operate efficiently on a cycle time of C/2. Many intersections probably can if C is large compared with L. If this is possible at all or most intersection other than the critical intersection, one should consider the possibility of operating the critical intersection at a cycle time C and others at C/2. See section 3.2 g, h. At the expense of some disruption of the arterial flow at or near the critical intersection, this may give a very substantial reduction in the delay to the cross streets and turning traffic, and a substantial net reduction of delay for the whole system.

4. Measure some typical trip times between intersections for vehicles at the rear of platoons. Regardless of how many vehicles may turn on or off the arterial, the off-sets should usually be chosen so that a hypothetical vehicle which would pass the critical intersection at the end of the arterial green will pass every other intersection upstream or downstream of the critical intersection also at the end of a green interval. If some intersections operate on a cycle time C/2 choose the off-sets for the ends of the intermediate green intervals similarly.

on a cycle time C/2 choose the off-sets for the ends of the intermediate green intervals similarly.

There may be some situations in which this is not the "optimal" choice of off-sets (see section 3.2d,3), but this choice is always reasonable.

5. Choose the arterial green intervals (and thus the starting times of the arterial green intervals) for the noncritical intersections. The choice here is also somewhat flexible (see section 3.2j) because one presumably has some extra time to partition between the arterial and the cross street. The main issue here is that there is negligible benefit to the arterial traffic from giving the arterial more green time than needed in order that the

platoon can pass the critical intersections. In particular, there is typically little benefit derived from advancing the arterial green at some mth intersection to clear out a queue of vehicles which may have turned onto the arterial at the intersection m - 1, unless the extra green time is needed for all vehicles to pass this intersection at a flow $s_1^{(m)}$.

6. Having implemented some proposed strategy as described above, one should "fine-tune" the strategy to correct for any deficiencies.

For a well designed pretimed signal strategy one expects some stochastic queueing at or caused by critical intersections. If there is none, the cycle time should be reduced. If, however, C has been chosen approximately as in step 1, the total delay will be quite insensitive to any small changes (by 20 percent or so) in C . It is not possible from casual observation to see whether or not a change in C will improve the operation.

There should be negligible stochastic queueing at noncritical intersections downstream of critical intersections. When a green interval at the critical intersection is fully utilized (there is a residual queue), the last vehicle in a platoon from the critical intersection should usually just barely clear the downstream intersection at the end of the green interval, except possibly in cycles when there is an exceptionally large number of new vehicles turning onto the arterial and/or a deficiency of vehicles leaving. Otherwise, one should not be too concerned about what happens to the lead vehicles turning onto the arterial and/or a deficiency of vehicles leaving. Otherwise, one should not be too concerned about what happens to the lead vehicles in the platoon.

There will typically be stochastic queueing at noncritical intersections upstream of the critical intersection. The key thing to check here is that whenever the queue vanishes at the critical intersection, the (residual) queue should also vanish at upstream intersection, and the last vehicle in a platoon from upstream should arrive at the critical intersection close to the end of of the green interval. This will guarantee that the critical intersection is

kept as busy as possible. There is no reason to give upstream intersections any more arterial green time than necessary to keep the critical intersection busy whenever there is a queue upstream.

If there were no stochastic queueing, a well designed pretimed signal strategy would be very effective in keeping an arterial platoon compact and moving at a desirable speed. No information which one could obtain from vehicle detectors would be of much use except perhaps that an on-line measurement of the flows over 10 muinutes or more might be useful for the purpose of gradually varying the cycle times, splits, and off-sets automatically as these flows vary throughout the day. The primary function of a traffic responsive system, however, would be to reduce the delays due to stochastic queueing.

Most of the stochastic queueing will occur at or be caused by the critical intersections. In a pretimed plan one chooses the cycle time and split at the critical intersection so as to give some extra capacity to both the arterial and the cross street in order to accommodate fluctuations. This means that the intersection will be idle or underutilized for a certain fraction of time, independently in the two directions. The objective of a traffic responsive strategy would be to salvage any unused time for one direction and give it to the other direction whenever there is a residual queue in the latter direction. By the time one has observed some idle time, however, (particularly on the cross street) it is too late to influence the arrival time of an arterial pla-. By the time one has observed some idle time, however, (particularly on the cross street) it is too late to influence the arrival time of an arterial platoon. The traffic responsive strategy must, therefore, be coordinated with a pretimed periodic arrival of successive arterial platoons.

To be most efficient, a traffic responsive strategy should salvage and reassign as much of the underutilized time as possible. The main source of saving occurs when an arterial platoon clears the critical intersection prior to the scheduled end of the green interval, because the flow should drop nearly to zero after the platoon passes. To maximize this saving, however, the start

of the arterial green should be such as to compress the platoon to a flow $s_1^{(m^*)}$. One might even delay the start of the arterial green slightly at the critical intersection as compared with what one would have in a pretimed plan in order to be reasonably certain that any slack in the platoom caused by vehicles leaving the arterial at upstream intersections is squeezed out. The detectors should also be located so as to terminate the arterial green when the last vehicle in the platoon is about two seconds trip time from the intersection, so that the flow will continue for about two seconds into the yellow.

Saving even a few seconds is very important here because the average total amount of excess time available per cycle in the pretimed plan (with an appropriately chosen C) is comparable with the effective lost time per cycle. Any saving (or loss) in time must be compared with the effective lost time per cycle, not the cycle time.

Reassigning any saving in time from the arterial to the cross street will reduce the stochastic queueing on the cross street. This, in turn, will allow one to revise the pretimed split and give a larger scheduled green interval to the arterial. One could almost eliminate stochastic queueing on the arterial at the critical intersection by allowing the green to run as long as necessary to clear the platoon, provided that any time not needed by the arterial is given to the cross street.

If one can almost eliminate stochastic queueing on the arterial at the to the cross street.

If one can almost eliminate stochastic queueing on the arterial at the critical intersection and still give most of the average excess time to the cross street, the queue will sometimes vanish on the cross street before the scheduled time for the termination of green on the cross street. There is a net excess capacity, so there will still be some (average) time to waste. If the cross street is not itself a part of a coordinated signal system in the cross direction, the arrival rate on the cross street will not drop to zero when the queue vanishes. There is a benefit to the cross street from continu- [illegible] the green on the cross street even after the queue vanishes, because this

will postpone the start of the queue growth on the cross street in the next cycle. If one terminates the green early, there is likely to be a residual queue on the cross street in the next cycle.

If there is no queue on the arterial due to an overflow in a previous cycle, there may be negligible benefit from giving any excess time from the cross street back to the arterial, even if there are some vehicles which turned onto the arterial upstream waiting for the arterial green to start. The pretimed start of the arterial green was designed to compress the approaching platoon to a flow $s_1^{(m*)}$ (with allowance for the typical number of turning vehicles). Advancing the start of the arterial green from the scheduled time may have little effect on the time the last vehicle of the approaching platoon passes the intersection. Thus, any extra time which one gives to the arterial may be dissipated by allowing a flow less than $s_1^{(m*)}$ during the arterial green. Furthermore, there is no guarantee that the arterial vehicles which are released early at the critical intersection will be able also to pass downstream intersections early. These vehicles may eventually be compressed back into a compact platoon downstream.

If one uses a strategy of giving any time not needed by the arterial platoon to the cross street at the critical intersection, there is no need to have vehicle detectors on the cross street at the critical intersection. One could also apply the same strategy at every (noncritical) intersection, i.e., start vehicle detectors on the cross street at the critical intersection. One could also apply the same strategy at every (noncritical) intersection, i.e., start the arterial green at the preset time, terminate the green when the platoon has passed, and give any excess time to the cross street. Such a strategy, however, would require vehicle detectors on the approach to every intersection on the arterial.

Since detectors and controllers are expensive, one should consider possible strategies with cheaper equipment at noncritical intersections where there is ample time to waste. At intersections downstream of a

traffic responsive signal one would already have a preset start of the arterial green, but the duration of the platoon will vary from cycle to cycle depending on how many vehicles passed the actuated intersection and how many vehicles may have entered or left the arterial in between, but mostly because of the former. If one has ample time to waste at the downstream intersection, one could simply give the arterial a preset green interval large enough to accommodate almost any size platoon that is likely to pass. Since, however, the signals are already coordinated to time the start of the arterial greens, one could have the traffic responsive signal drive the signals immediately downstream. The arterial green at the downstream intersection could terminate at some specified time (a trip time) after the green terminates at the actuated intersection. This would automatically adjust the green interval at the downstream intersection for variations in the size of the platoon passing the actuated signal. To reduce the queueing downstream, one would also need to provide some extra green time to accommodate fluctuations in the number of vehicles turning on or off the arterial, but actually one would have already chosen the start of the arterial green downstream to accommodate for some turning vehicles.

Unfortunately, one cannot use a traffic responsive signal to drive signals upstream. In the pretimed plan, however, one would presumably have given signals upstream of a critical intersection enough arterial green time virtually to guarantee that the critical intersection would be kept busy whenever there nals upstream of a critical intersection enough arterial green time virtually to guarantee that the critical intersection would be kept busy whenever there was an overflow queue upstream.

If, with this pretimed plan, one were to make the critical intersection traffic responsive as proposed above, there would be no stochastic queue at the critical intersection. Any vehicle which could pass the upstream intersection during a green interval would also pass the critical intersection in the same platoon. In effect, the upstream intersection imposes a maximum arterial green on the critical intersection, slightly varying from cycle to cycle,

however, because the critical intersection will adjust to vehicles turning on or off the arterial.

There are likely to be stochastic queues at the upstream intersections, the same as in the pretimed plan, but these queues will be small compared with the queue that would have existed at the critical intersection in the pretimed plan. One has some flexibility here. One is making a trade-off between queueing on the arterial and queueing on the cross streets (particularly at the critical intersection). One may find it desirable to even out the lengths of successive arterial platoons by imposing a maximum on the arterial green (at the critical intersection or upstream). If the last vehicle in a long platoon travels slower than the last vehicle in a short platoon, the last vehicle in a long platoon might be cut off anyway because it cannot keep up with the signal progression.

The conclusion here is that, if the cross streets (particularly at the critical intersection) are not part of a progression system in the cross direction, one can design a very efficient traffic responsive control system by having vehicle detectors only on the arterial approaches to critical intersections.

If any cross street is itself part of a signal progression in the cross direction, particularly the cross street at the critical intersection, one would want to have a preset time for the start of the cross street green as well as a preset time for the start of the arterial green. There may now be stochastic queueing on both the arterial and the coordinate cross street. An efficient traffic responsive system should now have detectors on both the arterial and the coordinated cross street designed to terminate the green in either direction whenever the flow drops before the preset time. If one can transfer all the unused time in one direction to the other direction whenever there is a queue in the latter direction, the combined average queue in both

3.7. Commentary

This chapter contains no references to the previous literature. Whereas the literature on the isolated traffic signal is fairly well developed and has evolved in a systematic way, the existing literature on the one-way arterial is so sketchy as to be of little value. The "theory" literature deals with such idealized situations as to be of questionable practical value. Most of the "engineering" literature passes over the one-way arterial with simply a comment that one should choose the off-sets equal to the design speed trip time, and then goes on to the two-way arterial or a discussion of various computer programs. The lack of a systematic theory for a one-way arterial, however, has led to all sorts of myths and ill-conceived strategies for two-way arterials and networks.

If, for a one-way arterial with all signals operating on the same cycle time and with reasonable splits, one has chosen a cycle time reasonably close to an "optimal", the total delay would be insensitive to the cycle time. Also, if one has chosen the off-sets close to some average trip time, the delay would be rather insensitive to the off-sets. It is fairly common for traffic engineers to use a cycle time which is considerably larger than necessary, but they are not likely to make a poor choice of off-sets. Aside from the possibility that the delay might be significantly reduced by decreasing the cycle time from some large value, the only strategies described here that are likely to give a substantial (maybe 20 percent or so) reduction in delay are those with most signals operating on half the cycle time of the critical intersection or certain traffic responsive strategies (neither of which are in common use). Most previous attempts to improve on simple strategies have actually gone in the wrong direction.

It has been known for at least 25 years that one should choose the off-sets between signals so that the last vehicle in a platoon will pass every intersection,

regardless of the lengths of the green intervals. Most signal engineers, however, continue to think in terms of lead vehicles and the off-sets for the start of the green intervals. It is also part of the engineering lore that one should clear out any queue before the arrival of a new platoon (even if the queue was caused by an overflow of vehicles from the previous cycle). In most cases no serious damage is done, but one could imagine a situation in which, by advancing the off-sets to clear out a queue, one caused more vehicles to overflow into the next cycle which, in turn, would cause one to advance the off-sets even more, etc.

It took traffic engineers a long time to accept the notion that there may be an "optimal" cycle time for an isolated intersection based on a balance between the deterministic and stochastic queueing. Now, however, it has become almost universal (in the most elaborate computer programs) to postulate that the stochastic queue at every intersection is the same as if each intersection were isolated, even on a coordinated arterial or network. This has led to "refinements" which, in the idealized case with no turning traffic, gives unequal arterial green times at different intersection and a larger delay than would exist with equal green intervals (although the calculated delay with incorrect formulas predicts a reduction in delay).

The existence of some recipes for calculating "delays" at oversaturated intersections seems to perpetuate a myth that there is some "optimal" strategy

The existence of some recipes for calculating "delays" at oversaturated intersections seems to perpetuate a myth that there is some "optimal" strategy for dealing with oversaturated intersections. At least government agencies periodically promote research to develop some magical recipes to do the impossible.

As regards traffic responsive strategies, most signal installations have detectors in the wrong place to acquire relevant information and traffic controllers which implement irrational recipes. What strategies have been used to respond to traffic on an arterial have done just the opposite of what they

should do. Instead of having detectors on the arterial to terminate the arterial green when the platoon passes, they have detectors on the cross street to terminate the cross street green when the queue vanishes. The latter strategy is not even guaranteed to give any improvement over a pretimed plan.

4. COORDINATION ON A TWO-WAY ARTERIAL

4.1. Introduction

We will be concerned in this chapter with an arterial highway which carries traffic in two directions (in the numbering scheme of Chapter 2, in directions 1 and 3). The arterial intersects many one or two-way streets, but all intersections are simple right angle junctions (the cross street directions being labeled as 2 and 4). There may be turning movements at any intersection, but we will assume that the flow of vehicles turning on or off the arterial at any intersection is small compared with the through traffic in directions 1 and 3. It may be necessary to have turn bays and multiphase signals at some intersections, but since turning vehicles blocking a through lane are so disruptive, we will assume that there are turn bays at any intersections where a significant number of turns may occur.

The most obvious complication that arises for a two-way arterial as compared with a one-way arterial is that any progression scheme designed to benefit the traffic movement in direction 1 is typically incompatible with a scheme designed to benefit the traffic movement in direction 3. Thus the off-sets (and cycle times) must usually be chosen so as to compromise between the benefits to directions 1 and 3. As discussed in section 3.1, however, the first (and cycle times) must usually be chosen so as to compromise between the benefits to directions 1 and 3. As discussed in section 3.1, however, the first step in the analysis of any signal system is to determine for each intersection individually whether or not that intersection can accommodate the expected (average) flows $q_{ti}^{(m)}$ and $q_{\ell i}^{(m)}$ of through traffic and (left) turning traffic in directions $i = 1, 2, 3, 4$ at intersection m.

This is typically a much more complex problem for a two-way arterial than for a one-way arterial because the most critical intersection is likely to be one which requires some multiphase signal strategy with separate turning phases.

Since the capacity of any intersection is essentially independent of any choice of off-sets or coordination, the capacities for various strategies which were discussed in sections 2.10 and 2.11 for an isolated signal apply also to any signal in a network.

If some mth intersection is oversaturated or close to saturation for some specified flows $q_{ti}^{(m)}$ and $q_{\ell i}^{(m)}$, the issue of possible rerouting of traffic is much more important for a two-way arterial than a one-way arterial. Turning vehicles on a two-way street severely limit the capacity of the intersection for through traffic and/or necessitate unreasonably long cycle times. To reduce delays on the arterial, it is clearly advantageous to try to induce drivers who have a choice of routes to make their left turns at the less congested intersections. Unfortunately, other social issues are apt to arise. The least congested intersections would be those with low flows on the cross street, i.e., minor roads. They might, for example, be roads into residential neighborhoods. The residents may not want traffic to be diverted into their neighborhood. In any case, there would be some reason why any minor street has low flow (compared with the arterial or other cross streets). It must be unattractive in some way.

It is not reasonable to assume, under congested conditions, that the $q_{\ell i}^{(m)}$ are independent of the signal strategy. In the extreme case, if one bans left

It is not reasonable to assume, under congested conditions, that the $q_{\ell i}^{(m)}$ are independent of the signal strategy. In the extreme case, if one bans left turns at some intersections, the corresponding $q_{\ell i}^{(m)}$ will certainly be zero. If one provides some positive but restrictive capacity for turning vehicles, the turning vehicles might form a queue, overflow the turn bay, and block the through lanes. As discussed previously for the isolated intersection, to avoid blocking of the through lanes, one should not only provide turn bays long enough to store any vehicles which may enter the turn bay during any single cycle, one should also serve all or most of the vehicles in the turn bay each cycle

to prevent them from accumulating. If a queue should develop on the through lanes and back up beyond the entrance to the turn bay, this would limit the rate at which vehicles can enter the turn bay. The through vehicles and the turning vehicles would then suffer nearly the same delays at the intersection.

If there are alternative routes, particularly for the turning vehicles, the queues at critical intersections should (theoretically) stabilize at some value such that any additional drivers would choose an alternative route rather than wait in the queue. Thus, the existence of queues on the arterial may induce left turn vehicles to turn at intersections other than the critical intersections (whether one wishes them to do so or not).

There is a dilemma here. If a traffic engineer should try to coordinate signals so as to eliminate queues for the through traffic on the arterial (possibly by storing any excess vehicles at entrances to the system), it becomes more difficult to regulate the flow of left-turn vehicles. If a signal clears the queues in the turn bay each cycle, there is little inducement for the left-turn vehicles to turn at less critical intersections. On the other hand, if the signal does not clear the queue in the turn bay, there is a risk that the turn bay will overflow and block the through traffic, unless the delay to vehicles in the turn bay itself is sufficient to induce left turn vehicles to turn elsewhere. If the traffic engineer bans left turns (during certain hours) hicles in the turn bay itself is sufficient to induce left turn vehicles to turn elsewhere. If the traffic engineer bans left turns (during certain hours) at some critical intersections, this may cause some other intersection to become oversaturated due to the turning traffic or it may cause excessive traffic on streets where it is not wanted. Thus, it is typically very difficult (impossible) for a traffic engineer to control the routing of traffic in such a way as to utilize fully the capacity of a two-way arterial and also keep the traffic moving smoothly.

Most of this chapter will deal with the choice of off-sets and cycle times

for specified flows $q_{ti}^{(m)}$ and $q_{\ell i}^{(m)}$. These flows are assumed to be sufficiently low as to guarantee the existence of some signal strategy for which each intersection is undersaturated (for all directions). We do not explicitly consider the fact that the choice of the signal strategy may affect the flows. Any "optimal" strategy will be an optimal for fixed values of the flows.

4.2. Coordinate transformations

For any pretimed signal strategy on a two-way arterial the trajectories of vehicles traveling in any direction depend on the setting of the traffic signal. The choice of signal strategy will depend on the $q_{ti}^{(m)}$ and $q_{\ell i}^{(m)}$, but, for any given choice of signal strategy, there is no direct interaction between vehicles traveling in directions 1, 2, 3, and 4, in particular between vehicles in directions 1 and 3 (except possibly for turning vehicles which, with no special turn signal, are allowed to filter through gaps in the opposing traffic stream). Much of the theory described in Chapter 3 relating delays, stops, etc., to the signal timing on a one-way arterial, therefore, applies also to two-way arterials for directions 1 and 3 respectively.

Most of the complications associated with the theory for a two-way arterial relate to the fact that the signal timing as seen by the vehicles in directions 1 and 3 are severely constrained because the signal permits only certain combinations of traffic movements (through or turning traffic) simultaneously. One cannot choose the signal timings in directions 1 and 3 independently. Also for a two-way arterial the signal timing affects the through traffic and turning traffic differently, so one must consider delays to the through and turning traffic separately.

Since delays, i.e., trip times relative to some ideal "uninterrupted" trip time, depend only on the actual times that vehicles pass various points, but not on the detailed motion of the vehicles between intersections, we

found it convenient, in dealing with the one-way arterial, to replace actual trajectories of vehicles by some hypothetical trajectories for which vehicles travel either at some design speed between intersections or are stopped at an intersection. If the design speed should vary from one intersection to the next, we further recognized that the delays do not depend on the scale of distances. If one measured "distance" from some reference point as the uninterrupted trip time from that reference point, then the "speed" of an uninterrupted vehicle would always be 1. An idealized trajectory would then have slope of either 1 if the vehicle is moving or 0 if stopped. We also found it useful to imagine that an observer at each intersection would measure time relative to the passing of some hypothetical uninterrupted vehicle, so that he would measure "time" as $t - x/v$. An uninterrupted vehicle would then have a vertical trajectory , and any deviation of a vehicle from a vertical trajectory would measure its delay.

For a two-way arterial, it is convenient to use a slightly different type of schematic trajectory representation to illustrate how the signal affects the delays to vehicles in <u>both</u> directions at the same time. The trajectory of a vehicle traveling in direction 3 will, of course, have a negative slope, and the typical (or design) speed in direction 3 may be different than in direction 1, i.e., the trip time $\tau^{(m)}$ from intersection $m - 1$ to m in direction 1 the typical (or design) speed in direction 3 may be different than in direction 1, i.e., the trip time $\tau^{(m)}$ from intersection $m - 1$ to m in direction 1 may be different from the trip time $\tau'^{(m)}$ from intersection m to $m - 1$ in direction 3.

Whereas for a one-way arterial we chose to measure "time" at some intersection relative to the passing time of some uninterrupted reference vehicle so as, in effect, to make all uninterrupted vehicles have vertical trajectories, we could measure "time" relative to the passing time of <u>any</u> moving object. Suppose, for example, that an observer at intersection m starts his clock a

time $(\tau^{(m)} - \tau'^{(m)})/2$ later than the observer at intersection $m - 1$, i.e., relative to the passing of some object that takes a time $(\tau^{(m)} - \tau'^{(m)})/2$ to travel from intersection $m - 1$ to m. The "trip time" of an uninterrupted vehicle traveling from $m - 1$ to m, measured as the difference in the clock times as seen by the observers at $m - 1$ and m is now

$$\tau^{(m)} - (\tau^{(m)} - \tau'^{(m)})/2 = (\tau^{(m)} + \tau'^{(m)})/2 ,$$

the average of the trip times in the two directions. Similarly, the "trip time" of an uninterrupted vehicle traveling from m to $m - 1$ will be

$$\tau'^{(m)} + (\tau^{(m)} - \tau'^{(m)})/2 = (\tau^{(m)} + \tau'^{(m)})/2 .$$

Thus, the "trip times" in the two directions are equal.

Suppose now, in combination with this, we measure "distance" so that the "spacing" between intersections $m - 1$ and m is $(\tau^{(m)} + \tau'^{(m)})/2$. An interrupted vehicle traveling in either direction 1 or 3 will now travel a "distance" $(\tau^{(m)} + \tau'^{(m)})/2$ in a "time" $(\tau^{(m)} + \tau'^{(m)})/2$. Thus, a vehicle traveling in direction 1 will now, in effect, have a trajectory with slope (speed) 1, and a vehicle traveling in direction 3 will have a trajectory with slope -1.

With this simple transformation of space and time coordinates, we can map slope -1.

With this simple transformation of space and time coordinates, we can map any problem involving space dependent design speeds and/or unequal speeds in directions 1 and 3 into a corresponding problem with a constant speed of 1 in direction 1 and -1 in direction 3.

The above transformation depends only on the uninterruped trip times between intersections. It has nothing to do with the signal settings themselves or how drivers may react to the signals. An observer at intersection m will, however, record all events, such as the actual times when vehicles

pass him or the times at which the signal changes phase, relative to his clock. If in the transformed coordinates we should choose a signal off-set $\delta^{(m)}$ between the signals at intersections $m-1$ and m, the off-sets relative to absolute time would be $\delta^{(m)} + (\tau^{(m)} - \tau'^{(m)})/2$. "Delays" to vehicles in the transformed coordinates would be measured as displacements from a hypothetical trajectory of slope 1 or -1 in directions 1 or 3 respectively. These delays are the same, however, in either coordinate system since each observer measures delay in the same units (seconds), as the time displacement between the actual trajectory and an uninterrupted trajectory.

Hereafter in this chapter we will assume that, if the design speed varies with location or direction, the problem has been transformed into an equivalent problem with a speed which is independent of location or direction. We also assume that the uninterrupted trip time between intersections $m-1$ and m (or m and $m-1$) for a vehicle which may turn left at intersection m (or $m-1$) is the same as for a vehicle continuing through intersection m (or $m-1$).

4.3. Schematic trajectories

A schematic representation of some hypothetical (transformed) trajectories for a two-way arterial are shown in figure 4.1. In contrast with a one-way arterial which permitted vehicles to turn off the arterial at any time during an arterial green phase of a signal, we must now recognize for a two-way arterial terial which permitted vehicles to turn off the arterial at any time during an arterial green phase of a signal, we must now recognize for a two-way arterial that a signal may have up to three separate phases for the 1, 3 directions, chosen, however, from four possible combinations plus as many as three phases from a possible four for the 2,4 directions. Whereas for a one-way arterial we could represent the signal by drawing a solid line segment during the effective red time and no line during the effective green (just two states of the signal), for a two-way arterial with a multiphase signal, it would be advantageous to use some color codes with possibly eight different colors to represent each of the eight possible combinations of traffic movements.

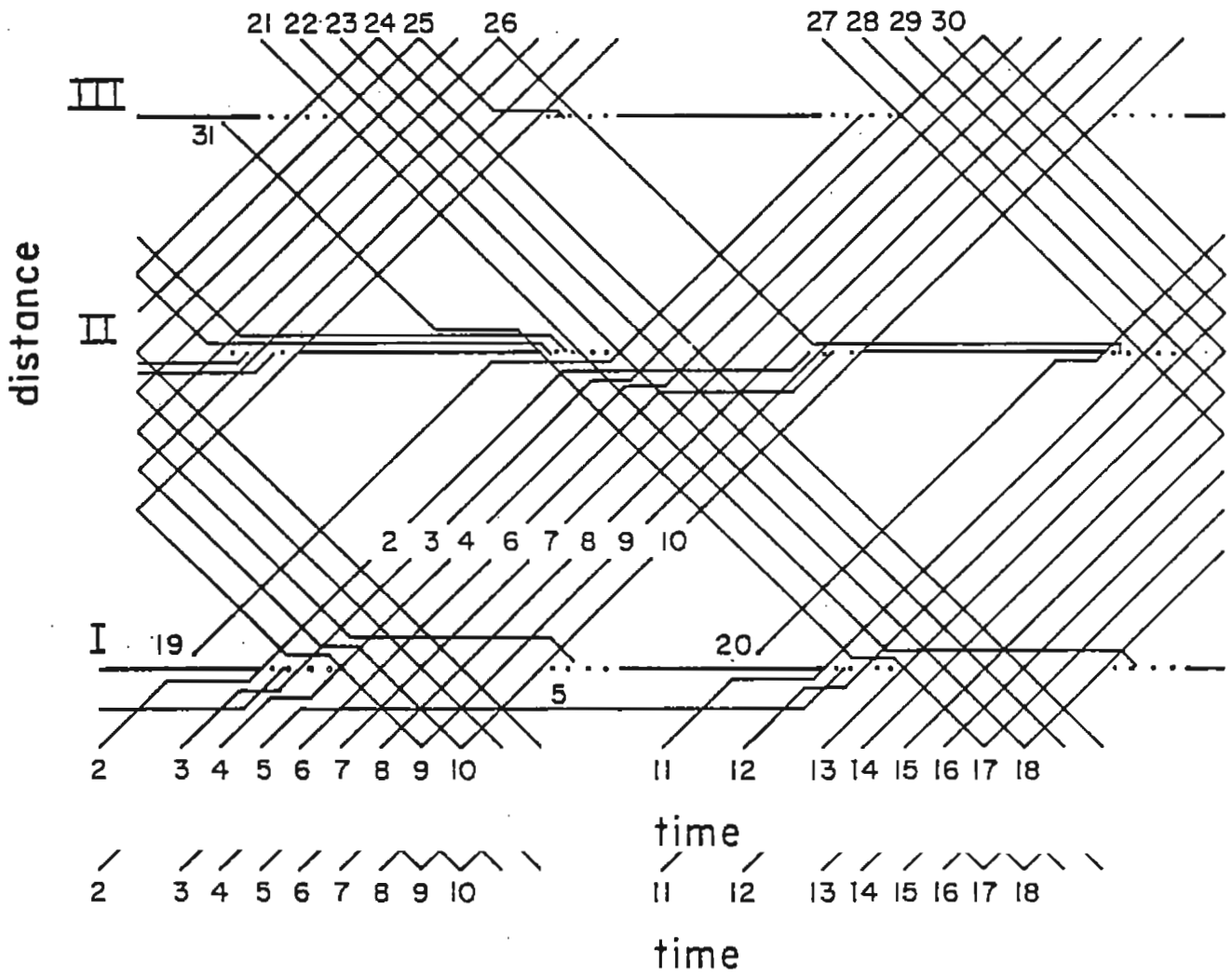


Fig. 4.1 - A schematic representation of vehicle trajectories for two-way traffic.

The phases for directions 2 and 4 will determine when vehicles may turn onto the arterial from the cross streets, but we do not expect that the times when vehicles turn onto the arterial will significantly affect the strategy for timing the signals for directions 1 and 3. Although we should include the trajectories for turning vehicles in our analysis, we have not coded the separate phases for directions 2 and 4 in figure 4.1. We have simply drawn a solid line segment over those (effective) times when there is no traffic movement for either directions 1 or 3.

The phase sequences in directions 1 and 3 will be important because the time needed to serve the through traffic in direction 3 will not generally be the same as the time needed to serve the through traffic in direction 1, and the phase sequences will affect the off-sets of the signals for the two through traffic movements. It might be advantageous to identify three possible phases for directions 1 and 3, t1 plus $\ell 1$, t1 plus t3, and t3 plus $\ell 3$ (in some order) as in figure 2.12 b or c, but we will use only two codes. We will draw no line when the signal permits through travel in both directions 1 and 3, and a dotted line when the signal permits through travel in only one direction, i.e., for either the t1 plus $\ell 1$ or the t3 plus $\ell 3$ phases. For any particular signal one must, however, specify whether the dotted line means that travel is permitted in direction 1 or 3, i.e., whether the signal has a leading or lagging left in direction 1 or 3. In figure 4.1, for example, intersections I is permitted in direction 1 or 3, i.e., whether the signal has a leading or lagging left in direction 1 or 3. In figure 4.1, for example, intersections I and III have a leading left for direction 1 and a lagging left for direction 3, whereas for intersection II they are reversed. If one chooses to use simultaneous left turn phase as in figure 2.12d, this time can be absorbed into the effective red (solid line segment), since it does not permit through traffic in either directions 1 or 3.

Figure 4.1 is not meant to illustrate some potentially optimal strategy, only to show how one can visualize the consequences of some arbitrary strategy.

The scale of coordinates has been chosen as described in section 4.2 so that vehicle trajectories have slope 1 or -1 in directions 1 or 3 respectively. We do not specify in which lanes vehicles may be traveling but there are turn bays which allow through vehicles to pass turning vehicles which are stopped.

Not all vehicles in figure 4.1 are numbered, but vehicles 1 to 20 are traveling in direction 1. Vehicle 1 starts in the turn bay and its trajectory terminates in this figure when it leaves the arterial during the leading left phase at intersection I. Vehicles 2, 3, and 4 arrive at intersection I before they can leave and are delayed. Vehicle 5 will turn left, but it barely misses the turn phase and is delayed nearly a whole cycle (one of the disadvantages of a leading left phase). Vehicles 6 to 10 pass intersection I with no delay. Vehicles 11 to 18 follow the same pattern as vehicles 1 to 10. Vehicles 19 and 20 turn onto the arterial from the cross street at some time during the arterial red and are delayed at intersection II.

Vehicles 2 and 6 turn left at intersection II during a lagging left phase. Vehicles 7, 8, 9, and 10 experience no delay at any of the intersections I, II, or III. Vehicle 19 leaves the arterial at intersection III with no delay during a leading left.

For the opposite direction, vehicle 25 turns left during the lagging left (for direction 3) of intersection III after a short delay. Vehicle 26 turns left in

For the opposite direction, vehicle 25 turns left during the lagging left (for direction 3) of intersection III after a short delay. Vehicle 26 turns left in the leading left at intersection II after a much longer delay. Vehicle 31 enters the arterial at intersection III but does not delay vehicle 21 which arrives at intersection II just after vehicle 31 has passed the intersection. Vehicle 27 arrives after the signal has been green for some time (unused time).

4.4 Two-way progression

Suppose, for now, we neglect turning traffic and assume that $q_{t1}^{(m)} = q_{t1}$, $s_{t1}^{(m)} = s_{t1}$, $q_{t3}^{(m)} = q_{t3}$, and $s_{t3}^{(m)} = s_{t3}$ are the same at every intersection.

The fraction of time needed to serve the traffic is q_{t1}/s_{t1} in direction 1 and q_{t3}/s_{t3} in direction 3 (independent of the cycle time C). If one could somehow compress the traffic in directions 1 and 3 into periodic platoons of flow s_{t1} and s_{t3} respectively, and these vehicles could travel without interruption, the vehicle trajectories would appear as illustrated in figure 4.2a and b for relatively small or large values of q_{ti}/s_{ti} respectively.

In this figure "distance" is measured in units of the uninterrupted trip time as described in section 4.2 so that all trajectories have slope +1 in direction 1, -1 in direction 3. For now, we also disregard stochastic effects and assume that all platoons traveling in the same direction have the same number of vehicles giving "bandwidths" of $(q_{t1}/s_{t1})C$ and $(q_{t3}/s_{t3})C$ in directions 1 and 3 respectively. But if we measured time and "distance" in units of C, the geometry of figure 4.2 would be independent of C. Except for a vertical or horizontal translation of the origin of the coordinate system or a scale factor, figures 4.2a, b depend only on q_{t1}/s_{t1} and q_{t3}/s_{t3}.

If there were no intersections, or traffic in directions 2 and 4 had to yield to traffic in directions 1 and 3, the traffic in directions 1 and 3 could travel without interruption. One could also insert traffic signals at any location along the arterial for the sole purpose of keeping the platoons compact, or to displace any vehicles which may turn onto the arterial from cross streets into the next arterial platoon. The issue here is whether or not, having done to displace any vehicles which may turn onto the arterial from cross streets into the next arterial platoon. The issue here is whether or not, having done this, it is possible to accommodate the cross street traffic during the time when there is no arterial flow in either direction.

We have not yet specified in figure 4.2a, b where the cross streets may be located. In doing so, one still has the choice of two as yet unspecified parameters, the cycle time C which specifies the scale of figure 4.2 and the location of some reference point for the vertical coordinate; for example, the location of some critical intersection. We also have the freedom of choosing a time origin (a reference point for the horizontal coordinate) but, obviously,

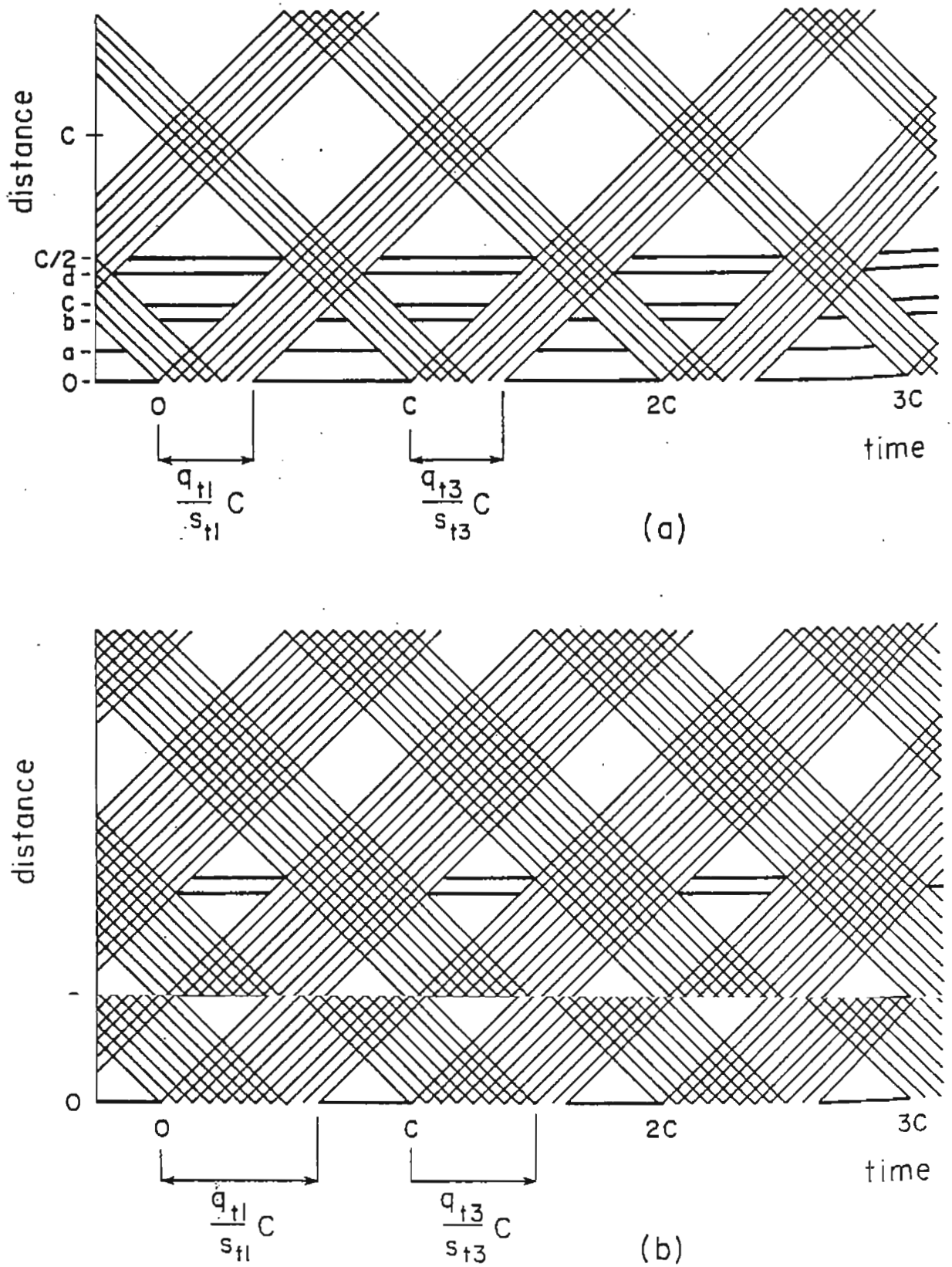


Fig. 4.2 - Trajectories of vehicles for a two-way progression for relatively light traffic (a) and heavy traffic (b).

whether or not an intersection can accommodate the cross traffic at any location does not depend on the time origin for the periodic behavior of the signal (it depends only on the fraction of available time).

From figure 4.2 one can see that the time available for any cross traffic is a periodic function of the "distance" with period C/2, i.e., with a period equal to the distance an uninterrupted vehicle can travel in a time C/2. A graph of the fraction of time available to the cross street vs "distance" (i.e., the uninterrupted trip time in units of C) would have a form as shown in figures 4.3a, b. There are two types of situations as illustrated in (a) and (b) depending on whether $1 - q_{t1}/s_{t1} - q_{t3}/s_{t3}$ is positive (a) or negative (b).

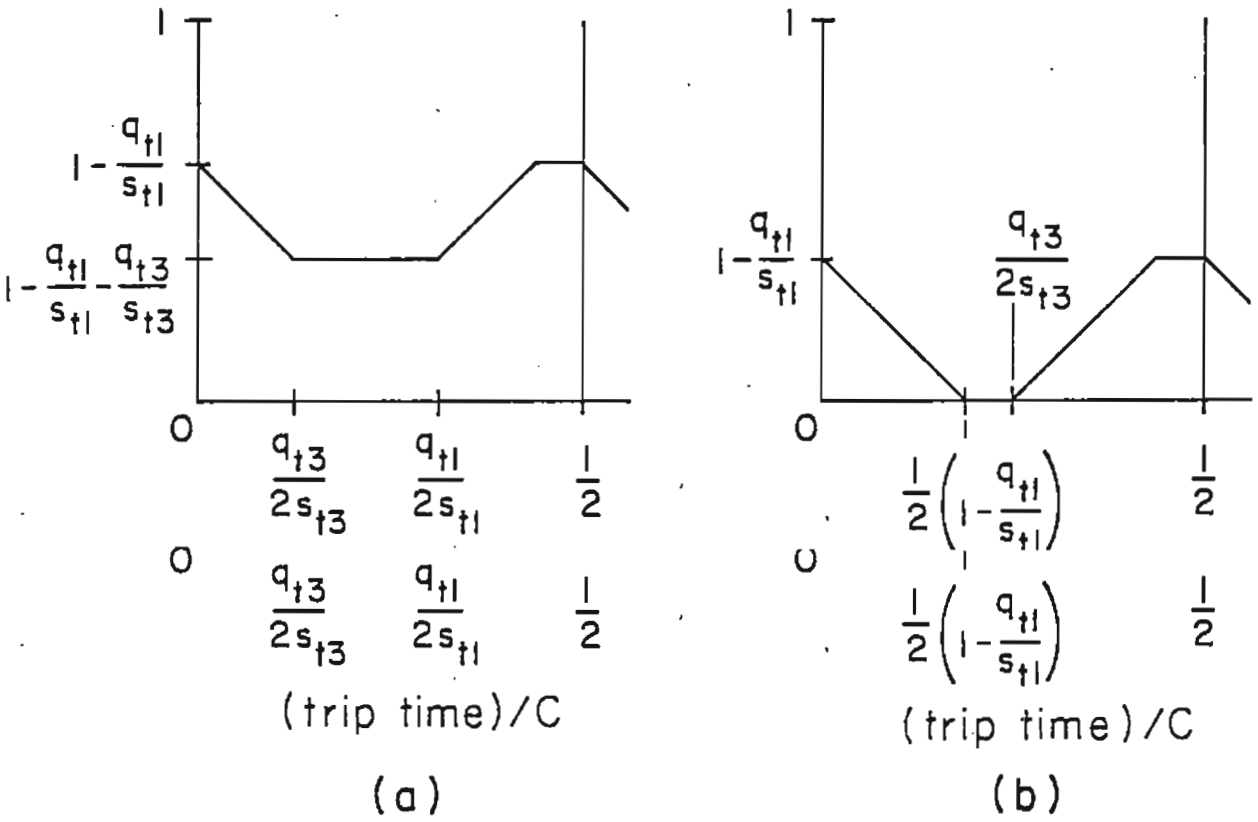


Fig. 4.3 - The fraction of a cycle time available to cross street traffic vs the trip time (location) for light arterial traffic (a) or heavy traffic (b).

As one can see from figure 4.2(a), the fraction of available time decreases with distance starting at the distance origin of figure 4.2a, because the overlap time of the two platoons in directions 1 and 3 decreases until the platoons do not overlap at all at the "location" $q_{t3}/2s_{t3}$ in figure 4.3a (also designated in figure 4.2 by location a). From this location until the location $q_{t1}/2s_{t1}$, the fraction of time available is the sum of two segments per cycle, a time interval between the end of a platoon in direction 1 until the start of a platoon in direction 3, and then from the end of a platoon in direction 3 until the start of the next platoon in direction 1 (as illustrated in figure 4.2a by location b).

From location $q_{t1}/2s_{t1}$ (designated in figure 4.2a by location c) to location $(q_{t3}/s_{t3} + q_{t1}/s_{t1})/2$ the platoons overlap again, increasingly so until the platoon in direction 1 completely covers that in direction 3 (location d in figure 4.2a). From this point until "location" C/2, the fraction of time available is constant as the position of the platoon in direction 3 shifts relative to (but within) that in direction 1. The pattern then repeated periodically with period C/2.

In figure 4.2b not only is $q_{t1}/s_{t1} > 1/2$ but also $q_{t1}/s_{t1} + q_{t3}/s_{t3} > 1$ so that the platoon in directions 1 and 3 overlap at all locations. Starting at location 0 where the platoons overlap completely, the available time to the cross street decreases with location until it vanishes at location $(1 - q_{t1}/s_{t1})/2$. It stays zero until location q_{t3}/q_{t3} but then increases again. The cross street decreases with location until it vanishes at location $(1 - q_{t1}/s_{t1})/2$. It stays zero until location q_{t3}/q_{t3} but then increases again. The fraction of available time stays constant over the final location interval width $(q_{t1}/s_{t1} - q_{t3}/s_{t3})/2$ as in figure 4.3a where the narrower platoon in direction 3 shifts relative to that in direction 1. Again this pattern is periodic with period C/2.

There is no guarantee; indeed it is not even likely, that one can accommodate the cross street traffic at a large number of intersections (signalized or not) while maintaining uninterrupted traffic in both directions 1 and 3. It

is not very difficult, however, to test whether or not it is possible (or if not, why not).

For any given geometry of intersections, one can firat draw a graph of the fraction of time needed by any cross street (i.e., the appropriate $q_2^{(m)}/s_2^{(m)}$) as a function of the uninterrupted trip time (in real time units) on the arterial from some reference point to the cross street. If there is no intersection at some location, the time needed at that location is zero. The graph is, therefore, positive only at discrete locations.

For any choice of C, one can rescale the horizontal coordinate of figure 4.3a or b to real time units and extend the graph periodically so as to obtain a graph of the available fraction of time vs the trip time from some arbitrary origin in real time units. One can then test to see if there is any horizontal translation of this graph of the fraction of time available which is everywhere above the graph of the fraction of time needed at each cross street. The former graph is certainly periodic in "distance" with period $C/2$, but the latter is probably not (unless intersections are equally spaced).

If the test fails for one trial value of C, usually because the peaks of one graph do not coincide with the peaks of the other, it should be obvious from inspection what new values of C one should try as likely candidates for success. It should usually also be obvious if no choice of C will work because the peaks of the curve for the time needed are aperiodic and incompatible success. It should usually also be obvious if no choice of C will work because the peaks of the curve for the time needed are aperiodic and incompatible with any acceptable choice of C.

Even if one does succeed in this first step, one must further verify that (for the proposed value of C) the available fraction of time for the cross streets exceed the fraction of time needed at each cross street by an acceptable amount, enough to accommodate the fraction of time lost in switching (L/C) plus some extra time to accommodate fluctuations in the cross street traffic. Note that if the available time for some cross street should occur in two intervals

per cycle, as in figure 4.2a at location b, then one must allow for two switching time losses per cycle. If one must also provide a minimum time for pedestrians, such a scheme will almost certainly fail.

One should first make the above test as if there were a traffic signal at every intersection, whether there is one or not. If the cross street traffic cannot be accommodated at some intersection with a signal, it certainly cannot be accommodated without a signal. On the other hand, if there is a workable scheme and one has signals at enough intersections to keep the platoons in shape and to control vehicles turning onto the arterial from cross streets, there may be some intersections with enough available time that the cross street traffic can be accommodated without a signal (particularly for intersections close to busy intersections where the platoons in the two directions overlap).

The extension of the above procedures to situations in which there is a small but nontrival amount of turning traffic at some intersections is, in principle, fairly straightforward. The existence of turning vehicles should have little effect on the (average) trip time of a vehicle near the rear of a platoon in either direction 1 or 3, and one should seek a two-way progression so that the vehicles at the rear of the platoon travel without interruption. As on a one-way arterial, the start of green for direction 1, for example, should be timed so that vehicles turning onto the arterial at intersection $m - 1$ com-

on a one-way arterial, the start of green for direction 1, for example, should be timed so that vehicles turning onto the arterial at intersection $m - 1$ com-

press any slack in the platoon at intersection m caused by vehicles turning off the arterial at intersection $m - 1$ or m. In a trajectory plot of through vehicles analogous to figure 4.2, the width of the platoon may vary slightly with locations as the $q_{t3}^{(m)}$ or $q_{\ell 3}^{(m)}$ vary with m. Thus, the time available for the cross traffic will not be exactly periodic in "distance."

One should also notice in figure 4.2 that anywhere there is a flow of traffic in direction 1 but not in direction 3, a vehicle which wishes to turn left from direction 1 should be able to do so with no interference from the through traffic

in direction 3. If there is a chance that some vehicle may be traveling in direction 3 outside the band, a straggler or a vehicle which turned onto the arterial at the upstream intersection, one could install a signal with a leading or lagging left for direction 1 so as to guarantee that the turning vehicles have a protected phase. A corresponding situation exists for vehicles traveling in direction 3. Thus, in the scheme illustrated in figure 4.1, one can insert a dotted line any place one has vehicles passing an intersection in only one direction.

If the most critical intersection requires a multiphase signal and one wishes to test whether or not one can have a two-way progression, one should probably first determine at which locations in the vertical coordinate of figure 4.2 one can place the critical intersection(s). For any location between locations 0 and a of figure 4.2(a) where there is traffic in direction 3 but not 1, for a sufficient time to accommodate the left-turning vehicles from direction 3 with a leading left for direction 3, and there is traffic in direction 1 but not 3 for a sufficient time to accommodate left-turning traffic from direction 1 with a lagging left for direction 1, there is a corresponding location between locations c and d of figure 4.2(a) where one could also accommodate the flows but with the leading and lagging phases reversed. One can also accommodate the left turning vehicles by placing the critical intersection date the flows but with the leading and lagging phases reversed. One can also accommodate the left turning vehicles by placing the critical intersection at location d but precede the arterial green with a simultaneous left turn phase as in sequence d of figure 2.12. These options provide several possible ways in which one might be able to establish a two-way progression.

We have neglected here the stochastic fluctuations in the arterial traffic. One certainly must provide green intervals at each intersection larger than $(q_{t1}/s_{t1})C$ and $(q_{t3}/s_{t3})C$ for the through traffic in directions 1 and 3, respectively, and it would be desirable to have bands wide enough to accommodate

typical fluctuations. At the intersections which constrain the bandwidths, however, there typically is excess time available to serve through traffic in directions 1 and 3 outside the bands. For example, at intersections where the bands do not overlap, a vehicle which cannot pass the intersection in the direction 1 band could pass during the green interval for direction 3 unless the signal is set so as to allow travel in only one direction (after turning vehicles have been served). A vehicle which cannot be accommodated in the band will be delayed somewhat but it will get through. The capacity of the arterial for the through traffic is typically appreciably larger than the capacity of the bands. The bands do, however, limit the capacity of the cross streets.

4.5. Special cases of two-way progression

(a) Alternating

In the textbook example of an ideal two-way progression, it is assumed that the spacings between intersections are all equal or, more generally, that the trip time $\tau^{(m)} + \tau'^{(m)}$ up and back between adjacent intersections can be expressed as an integer multiple of some time C (which is chosen as the cycle time) for all m. With no turning traffic, the signals are set so as to allow simultaneous flow in directions 1 and 3 for (at least) whatever time is needed to accommodate the flows in directions 1 and 3.

The resulting flow pattern is illustrated in figure 4.4a if $\tau^{(m)} + \tau'^{(m)} = C$ is needed to accommodate the flows in directions 1 and 3.

The resulting flow pattern is illustrated in figure 4.4a if $\tau^{(m)} + \tau'^{(m)} = C$ for all m. The platoons need not be compressed to the appropriate saturation rates, and the flows need not be the same in the two directions. There need not be intersections at all the places indicated as having signals; nor does one necessarily need signals at all intersections, but one should not have intersections at any locations other than those shown. The red and green intervals at an individual intersection need not be equal, but if they are, such a scheme is called an alternating system because one signal is green when its neighbors are red and vice versa.

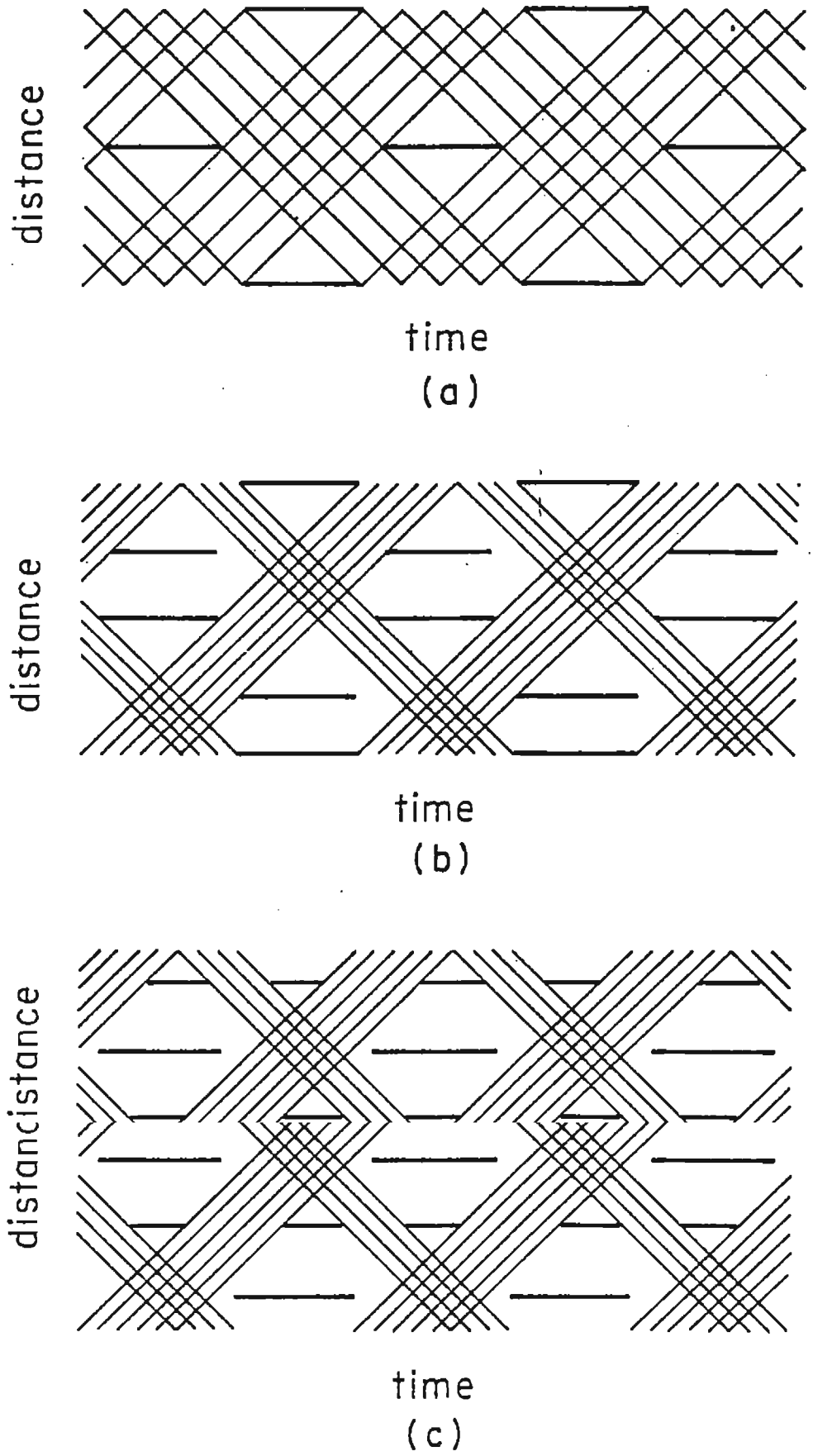


Fig. 4.4 - Some special cases of two-way progression.

The main limitation of this scheme, aside from the assumption that the intersections are regularly spaced, is that the spacing between adjacent intersections must be such that $\tau^{(m)} + \tau'^{(m)} = C$, or some integer multiple of C. On the other hand, C must be sufficiently large so as to accommodate the arterial and cross street traffic, i.e., a relation of the form

$$q_1^{(m)}/s_1^{(m)} + q_2^{(m)}/s_2^{(m)} < 1 - L^{(m)}/C$$

must hold at every intersection, and there may be constraints due to pedestrian crossing times. In any case, one would not likely choose a cycle time less than 30 seconds. But a value of $\tau^{(m)} + \tau^{(m')} = 30$ sec already represents a fairly large spacing (for a speed of 30 miles/hour, it would mean a spacing of at least 1/8 miles). For the more common choices of cycle times in the range of 45 seconds to 90 seconds, the typical spacing between urban intersections is about half what one would like for a two-way progression, which means that half the intersections would be located where the two platoons barely overlap, possibly where there is no time at all available for the cross street. Also, if the cycle time is chosen to establish a two-way progression, one cannot adjust the cycle time in response to changes in the flow.

It should be noted here that if the trip times between intersections are nearly (but not exactly) equal, no great harm is done. In particular, if the trip time between two intersections is slightly less than the design value (i.e., nearly (but not exactly) equal, no great harm is done. In particular, if the trip time between two intersections is slightly less than the design value (i.e., two intersections are a bit closer together than the rest), lead vehicles in platoons would be slowed down and, in effect, have a trip time equal to the design trip time between these intersections. If, however, some intersections are farther apart than the design value, vehicles passing an intersection near the rear of the green interval might be stopped at the next intersection and delayed into the next cycle. In either case, small irregularities give small delays, but one would probably prefer to delay an entire platoon by a small

amount than to displace a small number of vehicles into the next platoon. Thus, one might choose a cycle time C somewhat larger than the mean value of $\tau^{(m)} + \tau'^{(m)}$ if these values are slightly different.

The more general scheme described in the last section offers greater flexibility, particularly if there is a significant difference in the values of q_{t3}/s_{t3} and q_{t1}/s_{t1} (unbalanced flows). If $\tau^{(m)} + \tau'^{(m)}$ is nearly equal to C (or an integer multiple of it) for all m and the off-sets are chosen so as to maintain a (perfect) progression in direction 1, then a compact platoon in direction 3 might be able to pass many intersections without delay even though the $\tau^{(m)} + \tau'^{(m)}$ are not exactly equal. The platoon in direction 3 has some flexibility in its time to pass an intersection because it uses only some segment of the available green interval. There is also the possibility, even if the platoons passing some intersection do not overlay completely, that one can extend the arterial green interval there and still have sufficient time left in the cycle to serve the traffic on the (minor) cross street.

Even with this greater flexibility, one still has the basic problem that the $\tau^{(m)} + \tau'^{(m)}$ are typically considerably smaller than any acceptable value of C, and one cannot accommodate the cross street traffic at some intersections without displacing the arterial platoons.

(b) Double (or triple) alternate

In the standard transportation engineering literature, it is often sug-

(b) Double (or triple) alternate

In the standard transportation engineering literature, it is often suggested that if intersections are regularly spaced but too close together to permit an alternating scheme as described in (a) (except for unacceptably short cycle times) one should try a "double alternate" scheme as illustrated in figure 4.4b. Actually figure 4.4b is not quite what is usually called the double alternate. The usual scheme has equal width bands in the two directions and equal red and green intervals so that pairs of adjacent signals switch simultaneously but conservative pairs alternate phases as in figure 4.4a.

This scheme is also a special case of the general scheme described in section 4.4, but it will succeed only if the arterial green can accommodate both of the opposing platoons with no overlap, i.e., the fraction of arterial green must be at least $q_{t1}/s_{t1} + q_{t3}/s_{t3}$. Thus, the cross street flow must be accommodated in the remaining time,

$$q_{t2}^{(m)}/s_{t2}^{(m)} < 1 - q_{t1}/s_{t1} - q_{t3}/s_{t3} - L/C, \text{ for all } m. \qquad (4.5.1)$$

Note that, since the arterial platoons do not overlap at any intersection in this scheme, one could allow protected left turn movements in both arterial directions.

Condition (4.5.1) will be valid only for "light" traffic, typically only about half that which can be accommodated under saturated conditions. But if this condition is true, there are likely to be many other coordination schemes of the type described in section 4.4 which will also provide a two-way progression. Indeed, the intersections in figure 4.4b are located at points where the time needed by the arterial traffic is a maximum.

Under the somewhat more restrictive condition that the cross street traffic could be accommodated even with two lost times per cycle at every intersection, the scheme described in section 4.4 would allow one to have the intersections anywhere. Figure 4.4c, for example, shows the same arterial flow pattern as in figure 4.4b and the same spacing between intersections but with each inter-

anywhere. Figure 4.4c, for example, shows the same arterial flow pattern as in figure 4.4b and the same spacing between intersections but with each inter-

section midway between those shown in figure 4.4b. Half of the signals operate with the same split (and cycle time) as in figure 4.4b, but the platoons overlap at these intersections. There is an excess of time available for the arterial (which could be reassigned to the cross streets). The other signals operate with half the cycle time of the previous ones (and twice the fractional lost time per cycle due to switching).

These more general schemes of coordination are also very restrictive but not quite as restricted as the pattern shown in figure 4.4b. If there is just one major intersection, for example, which requires a relatively large fraction of time for the cross street, one should set the signals so that the arterial platoons overlap at this intersection and then see if one can still accommodate the cross street traffic at other intersections where the platoons may not overlap.

(c) Maximum bandwidth

As a generalization of goals achieved by the double alternate scheme of figure 4.4b, many people have devised schemes which maximize the sum of the "bandwidths" in the two arterial directions, i.e., the sum of the time intervals per cycle which can be covered by hypothetical uninterrupted vehicle trajectories in each of the two directions. Since a one-way progression would automatically give a bandwidth in the direction of the progression equal to the smallest of the arterial green intervals but possibly zero bandwidth in the opposing direction, it is usually further stipulated that bandwidths in both directions shall be positive (if possible).

In these schemes it is usually assumed that the red-green splits at each intersection are prespecified (maybe also the cycle time). Presumably these are dictated by the minimum fraction of the cycle time, including switching time, desired by each cross street, or by minimum pedestrian crossing times. are dictated by the minimum fraction of the cycle time, including switching time, desired by each cross street, or by minimum pedestrian crossing times. It is also assumed that all signals shall operate on the same cycle time. In contrast with the scheme shown in figure 4.4c, there shall also be only one cross street green interval per cycle.

This proposal that one should try to maximize the bandwidths has led to some interesting mathematical programming problems and many computer algorithms for determining off-sets and cycle times. The results, however, are not obviously consistent with what one wishes to achieve.

The most obvious deficiency of the maximum bandwidth scheme is that the maximum bandwidth (if positive in both directions) are typically only a small fraction of the minimum arterial green interval, for acceptable values of the cycle time. If the arterial traffic cannot be accommodated in the bands, the traffic may still be able to pass through the system, i.e., the system may be undersaturated. It is not obvious, however, that the off-sets which maximize the bandwidth will, under these conditions, also minimize the delays or other more realistic measures of performance. There are, in fact, situations in which the maximum bandwidth strategy will cause such long queues during the red intervals for nearly saturated flows that blocking of intersections will unnecessarily restrict the arterial flow ("grid lock"). There is a danger of this happening even for the double alternate scheme, and it is almost certain to happen for the triple alternate scheme.

If the arterial traffic can be accommodated in the bands, there would, of course, be no delays to arterial traffic except as the vehicles are delayed in entering the system either at the boundaries of the system or the cross streets. Under these conditions, however, there would be many possible choices of off-sets and splits which would achieve the same result of negligible delays on the arterial. Among these, the most desirable strategies would presumably be those which give the least delay to the cross streets. The general philosophy behind the maximum bandwidth scheme, however, seems to be that one will give those which give the least delay to the cross streets. The general philosophy behind the maximum bandwidth scheme, however, seems to be that one will give the cross street only some predetermined time, possibly the minimum acceptable time, and give any excess time to the arterial (whether the arterial can use it or not). In contrast with this, the scheme described in section 4.4 will give the arterial only the minimum time it needs to maintain the two-way progression and give any excess time to the cross street. It also admits the further possibility that one could assign the time to a cross street in two segments per cycle.

A further failure of the maximum bandwidth computer programs is that they faithfully do exactly what they are asked to do. If they are asked to find both a cycle time and off-sets which achieve a maximum bandwidth for a specified sequence of signals, with no restrictions on the cycle time, they may give a cycle time and an arterial green interval larger than the time it takes a vehicle to traverse the whole signal sequence forward and back. Certainly, as long as the signal stays green for the arterial, it can accommodate the traffic in both directions simultaneously. The cycle time which maximizes the sum of the bandwidths is undoubtedly infinite.

Our conclusion here is that maximizing the bandwidth is not likely to give the most desirable signal coordination under any conditions.

4.6. Closely spaced intersections

We have seen how one can test whether or not it is possible to have a two-way progression. Unfortunately, for signals in typical urban areas, the test often fails because the spacing between busy intersections (with $1 - q_1 s_1 - q_3/s_3$ too small to accommodate the cross street traffic) is too short, for reasonable values of the cycle time. To understand the problem further, we consider next an idealized situation in which we choose to have equal arterial green intervals at all intersections, but the intersections are so close together that we can treat them as a near continuum (analogous to sections 3.3b and d).

at all intersections, but the intersections are so close together that we can treat them as a near continuum (analogous to sections 3.3b and d).

In our dimensionless trajectory representation in which vehicles have "speed" of $+1$ or -1, we could choose the off-sets for any (dimensionless) progression speed u as illustrated in figure 4.5. Obviously we would not want to choose the progression speed between -1 and $+1$. If we choose $-1 < 1/u < +1$, the dimenionless form of (3.2.3) gives

$$\frac{\text{average travel time}}{\text{uninterrupted travel time}} = \begin{cases} 1 + \left(\frac{C}{G} - 1\right)\left(1 - \frac{1}{u}\right) & \text{for } u > 1 \\ & \text{in direction 1} \end{cases} \quad (4.6.1a)$$

$$\begin{cases} 1 + \left(\frac{C}{G} - 1\right)\left(1 + \frac{1}{u}\right) & \text{for } u > 1 \\ & \text{in direction 3} \end{cases} \quad (4.6.1b)$$

This is valid regardless of any possible interaction between vehicles, provided that the vehicles can travel between intersections at an average "speed" of ±1, and a typical vehicle travels a distance sufficiently long so as to be stopped at several intersections before leaving the arterial.

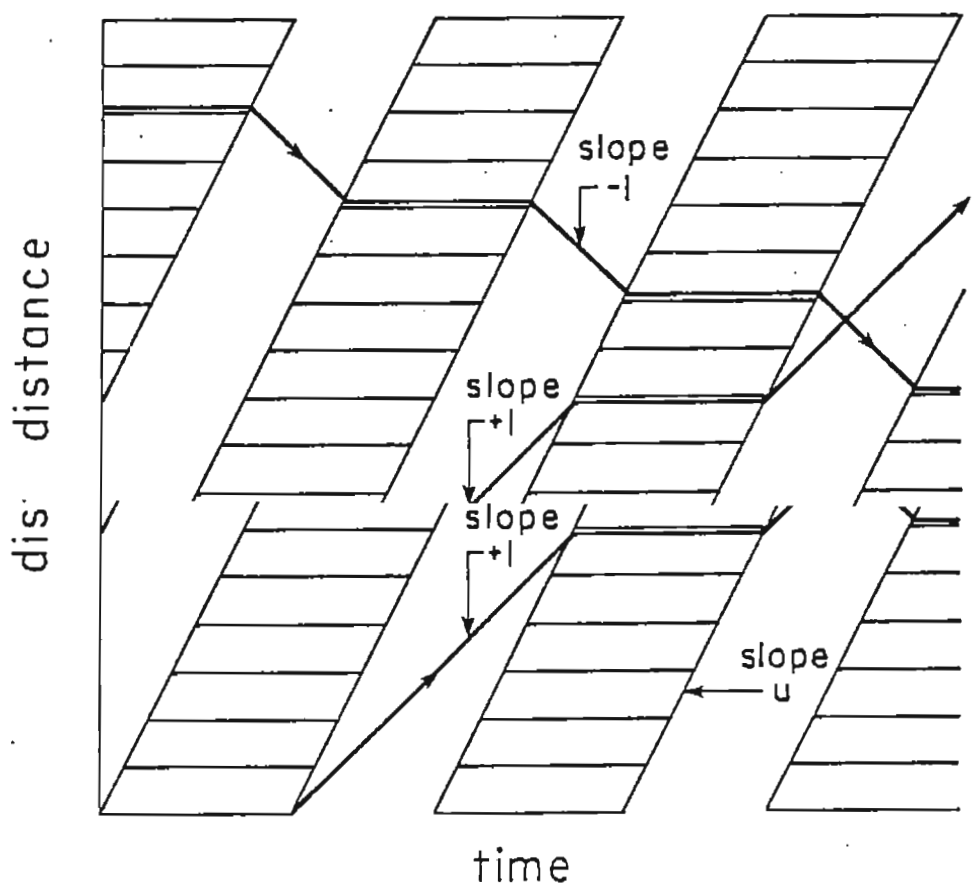


Fig. 4.5 – Some vehicle trajectories for closely spaced intersections.

If there is a flow q_1 in direction 1 and q_3 in direction 3, the total travel time for all arterial vehicles is the sum of (4.6.1a and b) weighted with the values of q_1 and q_3, i.e.,

$$\frac{\text{average travel time}}{\text{uninterrupted travel time}} = \frac{q_1}{(q_1 + q_3)} \left[1 + \left(\frac{C}{G} - 1\right)\left(1 - \frac{1}{u}\right)\right]$$

$$+ \frac{q_3}{(q_1 + q_3)} \left[1 + \left(\frac{C}{G} - 1\right)\left(1 + \frac{1}{u}\right)\right] \qquad (4.6.2)$$

$$= \frac{C}{G} - \frac{1}{u}\left(\frac{C}{G} - 1\right) \frac{(q_1 - q_3)}{(q_1 + q_3)} \quad \text{for } -1 < \frac{1}{u} < 1 \ .$$

One should first notice that if the flows are balanced, $q_1 = q_3$, (4.6.2) is independent of the progression speed u for $-1 < 1/u < 1$. In this case it would seem advisable to choose $u = \infty$, $1/u = 0$ (simultaneous coordination). There would be no reason to give excessive delays to direction 3, for example, in order to reduce the delay in direction 1. The travel time per vehicle, in this case would be C/G times the uninterrupted travel time (twice the latter for $C = 2G$).

If $q_1 > q_3$, the average delay as given by (4.6.2) is least for $u = 1$, coordination for the direction with the higher flow. This, however, gives a very heavy penalty to the traffic in direction 3. The travel time in direc-

coordination for the direction with the higher flow. This, however, gives a very heavy penalty to the traffic in direction 3. The travel time in direction 3 would be larger than the uninterrupted travel time by a factor of

$$1 + 2\ (C/G - 1) \ , \qquad (4.6.3)$$

a factor of 3 for $C = 2G$.

Actually for $u = 1$, a vehicle in direction 1 could theoretically travel for an infinite distance without being stopped by a signal, but his actual trip length is finite. If every vehicle in direction 1 enters the arterial (from a cross street) at the start of a green interval at some intersection,

one could choose a $1/u < 1$ such that most vehicles traveling in direction 1 would exit the arterial before the end of a green interval, without being stopped. The optimal u would then involve a trade-off between the delay to most vehicles in direction 3 and some vehicles in direction 1 (with exceptionally long trip length).

Disregarding the above dependence of delay on the trip length distribution, one should notice here, as in section 3.3d, that the delay on the arterial as given in (4.6.2) depends on the split C/G and the progression speed, but not on the cycle time itself. To obtain the total travel time of all vehicles in the system, one must add to the travel time (4.6.2) the delays to the cross street traffic and the delay to arterial vehicles as they enter the progression either from some "first" intersection or from cross streets, including stochastic queueing. All of these latter delays do depend on the cycle time and the splits, but they are essentially independent of the progression speed. Thus, the choice of an "optimal" cycle time C and an optimal progression speed u are completely independent of each other.

In this idealized signal system with equal arterial green times at all intersections, the choice of the cycle time will, as usual, involve a trade-off between deterministic and stochastic queueing, the former to the vehicles on <u>all</u> cross streets plus the vehicles entering the arterial, the latter to both the arterial and the cross streets but mostly at or caused by the critical inter-
cross streets plus the vehicles entering the arterial, the latter to both the arterial and the cross streets but mostly at or caused by the critical intersections. The choice of the split will involve a trade-off between the stochastic queueing on the arterial and the cross streets, particularly at the critical intersections, but also with the travel time on the arterial since the travel time in (4.6.2) would tend to favor giving more time to the arterial (particularly for direction 3 if the progression favors direction 1).

The choice of cycle time and split is somewhat arbitrary because one is not likely to give equal weight to delays for vehicles traveling in different

directions. It is particularly important, however, to recognize that the stochastic delay to vehicles on the arterial is certainly not the sum of delays that would exist at each intersection if each intersection were isolated (as proposed in most computer models). In the limit of a continuum of intersections (infinitely many) such a sum would be infinite. Actually the average stochastic delay to a single vehicle which enters the arterial at one intersection and exits at another is comparable with the delay it would experience at the most critical intersection between its entrance and exit if that intersection were an isolated intersection, and all other intersections were eliminated! The "optimal" choice of cycle time and (equal) splits should, therefore, be comparable with what one would have chosen at the most critical intersection if it were an isolated intersection.

In the present situation with equal arterial green intervals, the most critical intersections would be those with the largest values of $q_{t1}^{(m)}/s_{t1}^{(m)}$ for direction 1 or $q_{t3}^{(m)}/s_{t3}^{(m)}$ for direction 3. With traffic turning on and off the arterial, these will not necessarily be the same at all intersections. With unequal flows at different intersections, however, and perhaps some multiphase signals for turning vehicles, one may wish to use unequal splits. An analogous argument regarding the magnitude of the stochastic queueing still applies, but the critical intersections will be those which in isolation would analogous argument regarding the magnitude of the stochastic queueing still applies, but the critical intersections will be those which in isolation would have the largest stochastic queueing.

In comparing the theory described here with the theory of the two-way progression in section 4.4 and 4.5, it is obvious that, to reduce travel time in the system further, one must find some way for direction 3 vehicles to slide across the "red bands" of figure 4.5 with as little delay as possible (if the progression favors direction 1). In the two-way progression this was achieved either because the intersections were so far apart that vehicles could cross the red band in figure 4.5 between intersections, or because one could use two

green intervals per cycle for the arterial at some intersections making some slots through the red band to accommodate the arterial traffic in direction 3.

One could imagine placing traffic signals at every location along an arterial whether there were intersections there or not, and setting these signals so as to provide a progression as in figure 4.5. In our idealized model in which vehicles travel at speeds ±1 or 0, the introduction of extra traffic signals between intersections certainly would not decrease any vehicle's travel time. Thus, the simple formula (4.6.1), (4.6.2), and (4.6.3) represent upper bounds on the minimum travel times (delays) in the two directions for any system with signals at discrete locations.

The interesting feature of (4.6.1) is that it is independent of the location of the signals, the flows $q_{ti}^{(m)}$ (provided the system is undersaturated) and, as noted above, even the cycle time C (for given C/G). Indeed, (4.6.3) depends <u>only</u> on the value of C/G. A possible two-way progression, of course, gives a lower bound for the delays to through traffic in directions 1 and 3, namely zero, but the existence of a two-way progression depends on the spacings between all intersections, the cycle time, and the flows at all intersections.

If the cross streets are not part of any progression system in the cross directions, one could easily evaluate the delays to the cross street traffic for any strategy, since the delay to the cross street at the mth intersection depends on the signal setting at only the mth intersection. In the progression scheme of figure 4.5, one is essentially assigning all the time not needed for the through traffic (or turning traffic) to the cross street. For any choice of the cycle time, the scheme of figure 4.5 would, generally, give less delay to the cross street than a two-way progression which minimizes the delay in direction 3 at the expense of cross street delays.

4.7. Imbalanced Flows

If the flow in direction 1 is considerably larger than in direction 3, one would likely try first to set the signals so as to provide uninterrupted flow for the through traffic in direction 1, and then secondly to do the best one can for the traffic in direction 3, the cross traffic, and the turning traffic. This presupposes that the system is undersaturated. Certainly the primary issue is to choose the cycle time and splits at critical intersections so that they are undersaturated (if possible).

The minimum acceptable cycle time and minimum acceptable arterial green interval will be dictated by the capacity of the critical intersections (or by pedestrian crossing times). This cycle time is also likely to be such that one cannot establish a two-way progression over long distances. There may be some flexibility in the choice of the cycle time C but, whatever one chooses, it should be the same at all intersections. It may be possible to choose a cycle time so as to establish a two-way progression over certain segments of the arterial, but such a cycle time for one segment of the arterial may not be the same as that for another segment. It is not obvious that much can be gained in such circumstances by changing the cycle time. There is considerably more flexibility in how one partitions the cycle time at noncritical intersections particularly those with relatively light cross street traffic.

particularly those with relatively light cross street traffic.

Once one has specified the arterial green interval and the cycle time at the critical intersection and specified that through vehicles in direction 1 should not be interrupted, one has established a minimum through band for direction 1. At each intersection one has the option of partitioning the remaining time among various signal phases to accommodate the traffic in directions 3, 2, and 4, or turning traffic.

At the critical intersection(s), the cycle time and splits have presumably been chosen so that all signal phases are nearly fully utilized (otherwise one

should probably have chosen a shorter cycle time). If there are turning movements which require multiphase signals, one does have the option of using leading or lagging turn phases (or other multiphase strategies), but otherwise there is not much flexibility.

If there are turning movements at the noncritical intersections which require multiphase signals, some of the time assigned to the minimum through band for direction 1 may be assigned also to left turning movements for direction 1, but most of the time in the through band for direction 1 should be available also to the traffic in direction 3. Since, by assumption, the flow is predominantly in direction 1, the time available in this band for direction 3 vehicles is typically sufficient by itself to accommodate the flows $q_3^{(m)}$ in direction 3 at the saturation flow s_3 (but with some delay because the offsets have been chosen to favor direction 1).

At each noncritical intersection one also has a certain minimum acceptable time per cycle which must be allocated to the traffic in directions 2 and 4 (also possibly left turns from direction 3) to guarantee that these traffic directions are undersaturated and have acceptable delays. This must, of course, be allocated from the time outside the direction 1 through band (plus the time lost in switching). At the noncritical intersections, the minimum acceptable time needed by directions 2 and 4 is certainly less than the time available (otherwise the intersection would have been classified as "critical"). The time needed by directions 2 and 4 is certainly less than the time available (otherwise the intersection would have been classified as "critical"). The problem is to determine some strategy for partitioning the excess time between directions 2, 4, and 3 at each noncritical intersection. The objective is to reduce the delays in directions 2, 4, and 3 as much as possible without interrupting the through band for direction 1.

The delay in direction 3 is a very complicated function of the relative signal settings at all intersections. At certain intersections vehicles will arrive from direction 3 outside the green band for direction 1, but they will

not necessarily arrive in a single compact platoon. Particularly if some vehicles arrive just after the end of the minimum green band for direction 1 (or just before it is to start), one is tempted to extend this time to allow these vehicles to pass without delay (provided one has some extra time to allocate). This action will certainly not increase the total delay in direction 3, but it will not necessarily decrease the delay. It will, typically, increase the delay in directions 2 or 4.

It is very difficult to identify what one <u>should</u> do in order to minimize the total delay for some arbitrary locations of intersections and arbitrary values of the flows at all intersection, but one can identify certain things that one would <u>not</u> do, particularly just upstream or downstream of some critical intersections. One can also identify strategies which would give less delay than simply assigning all excess time to directions 2, 4 as suggested by figure 4.5.

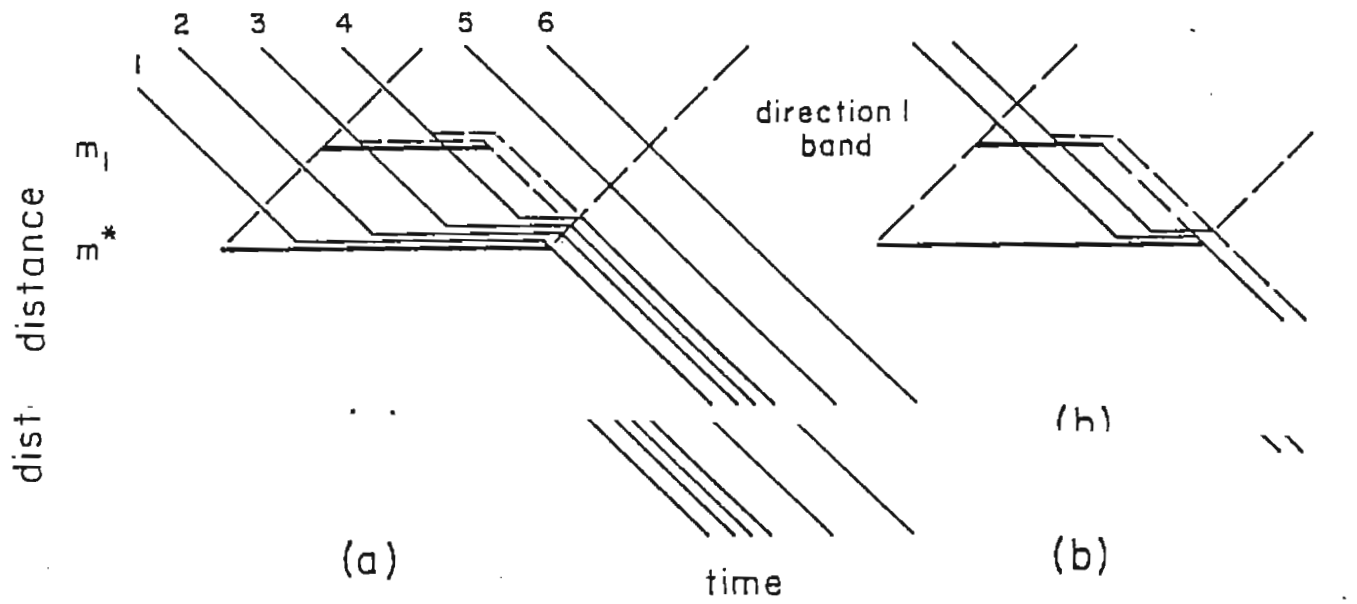


Fig. 4.6 - Trajectories of vehicles passing the critical intersection in direction 3 and the effect of a signal upstream.

Suppose, as illustrated in figure 4.6, for some critical intersection m*, we draw the effective red intervals for direction 3 (heavy line segments) and some hypothetical idealized trajectories of vehicles in direction 3 (solid lines

1-6) which would exist if there were no intersections immediately upstream (in direction 3) of m^*. Vehicles which are stopped by the signal at m^* will leave at headways $1/s_3$ after the signal turns green. Figure 4.6 also shows the extrapolation of the final trajectories (of slope -1) of these vehicles upstream (broken lines).

If now we were to insert a traffic signal at some intersection m_1 upstream of m^*, which would intercept vehicles 3 and 4, for example, this signal would cause no net increase in delay to any vehicle in direction 3, provided that the vehicles stopped at m_1 are released from m_1 no later than the time needed to follow the corresponding dashed lines so as to arrive at m^* before the time they were scheduled to leave m^*. The latest such termination of the red interval at m_1 will depend on the number of vehicles which are stopped at m^* but not at m_1, but it certainly will not be earlier than if there were no such vehicles, as illustrated in cycle (b) in which the first vehicle to leave m_1 when the signal turns green is also the first to leave m^*.

If we are to maintain the progression in direction 1, the signal at m_1 should not interfere with the through band for direction 1. The signals at m^* and/or m_1 might be multiphase signals but, if (as illustrated in figure 4.6) they are both two phase signals, then the effective red intervals for direction 3 will be (nearly) the same as those for direction 1. In this case there will

they are both two phase signals, then the effective red intervals for direction 3 will be (nearly) the same as those for direction 1. In this case there will be an earliest time that the red interval can start at m_1 without intercepting the direction 1 band of slope +1 ending with the start of red at m^*. The red interval at m_1 will, except for the lost time in switching be available to the cross street.

The point of the argument here is that if one has some intersection m^* which requires a relatively long time to serve the cross traffic and therefore a relatively long red interval for direction 3, then immediately upstream of m^* one can assign to the cross street traffic a certain minimum time, at no expense

to the traffic in direction 3 regardless of the arrival pattern in direction 3. One may need to assign more time than this to accommodate the traffic in directions 2 and 4, but nothing is gained by assigning less.

Note that the strategy illustrated in figure 4.6 does not necessarily minimize the delays to all vehicles at intersection m_1. To do so, one might, for example, have postponed the start of red at m_1 so as to pass vehicle 3, thus reducing the delay of this vehicle at m_1, but with no final benefit to this vehicle. If as a result of this, however, one must also displace the termination of the red at m_1 in order to serve the cross street vehicles, this may delay vehicle 4 not only at m_1 but also its departure from m^*.

The generalization of the above argument and figure 4.6 to cases in which the intersections m_1 and/or m^* may have multiphase signals is straightforward, but there are many possible modifications of figure 4.6. In all cases the effect of a signal at m_1 on the direction 3 vehicles will be determined by the effective red intervals for direction 3 as illustrated in figure 4.6. A direction 3 vehicle which is stopped at m_1 will suffer no increase in delay if it arrives at m^* prior to the time it would have left m^* in the absence of a signal at m_1. A multiphase signal, however, will influence the possible starting time of the effective red at m_1 and what part of this effective red is available to the cross street. The relative merits of various multiphase signal strategies may depend on the trajectory pattern of vehicles in direction 1 is available to the cross street. The relative merits of various multiphase signal strategies may depend on the trajectory pattern of vehicles in direction 3, which we will discuss later, but, for any strategy, there will be potentially some time available for directions 2 and 4 at no expense to direction 3 (regardless of the pattern of arrivals in direction 3).

When there is a left turn phase for direction 1, the signal will be red for direction 3 (and also the cross street). Thus, at m^*, the effective red for direction 3 would overlap the direction 1 band either at the start of the through band (leading left for direction 1) or the end (lagging left).

In figure 4.6, the direction 1 band would be wider than shown, with the same effective red for direction 3 at m^*. Since the cross street green at m_1 cannot start (at least) until the direction 1 band passes, a leading left for direction 1 at m^* would impose the least constraint on the signal at m_1. A lagging left at m^* would delay the start of the cross street green at m_1 relative to the red interval at m^*.

If one needed a direction 3 turn signal at m^*, then the direction 1 band could not pass during this turn phase. In figure 4.6, the direction 1 band would be narrower than shown (for a given effective red at m^*). A lagging turn phase for direction 3 would allow the cross street green to start earlier at m_1 (possibly at the expense of some delay to the turning vehicles in direction 3 which are stopped at m^*, but maybe to the benefit to the turning vehicles stopped at m_1). If one had both a leading left for direction 1 and a lagging left for direction 3 of equal duration at m^*, the direction 1 band in figure 4.6 would simply be shifted to the left giving more flexibility for the effective red at m_1.

If intersection m_1 is a minor intersection with light cross street traffic, chances are that there would not be much traffic turning left there from either direction 1 or 3, and one would not need turn signals at m_1. One can, however, insert a leading left phase for direction 3 at m_1 at no expense to the vehicles in direction 1 or 3, since the direction 3 vehicles are released from ever, insert a leading left phase for direction 3 at m_1 at no expense to the vehicles in direction 1 or 3, since the direction 3 vehicles are released from m_1 before the arrival of the direction 1 band.

If one needs a turn phase for direction 1 vehicles at m_1, it should obviously be a lagging left. The consequence of this would be that the effective red for direction 3 at m_1 would start earlier, overlapping the direction 1 band and potentially intercepting more direction 3 vehicles. This, however, would have little effect on the time when the cross street green starts because it cannot start until the direction 1 band passes. If the signal at m_1 is to

cause no increase in delay to any possible direction 3 vehicles, a lagging left for direction 1 at m_1 may indirectly decrease the time available for the cross street. It may decrease the number of direction 3 vehicles which pass m_1 but are stopped at m^* (such as vehicles 1 or 2 in figure 4.6a) and thus advance the time direction 3 vehicles must leave m_1 in order to pass m^* on schedule (as in figure 4.6b).

The above arguments may have rather limited applicability because the duration of the cross street green which one can insert at m_1 at no expense to the arterial traffic decreases rapidly as one moves upstream of m^*, typically at a rate -2, i.e., it would decrease to zero in a trip time from m^* of half the red interval at m^*. If there are some intersections m_1 to which this scheme applies with adequate cross street green to serve the cross traffic, there is a satisfactory way of setting the signals at m_1 independent of the pattern of arrivals in direction 3, i.e., independent of the setting of any other signals upstream of m^*. In choosing some overall scheme for the other intersections which do cause delay to direction 3 vehicles, one need not worry about the consequences of intersections like m_1.

If there is an intersection m_1 upstream of m^* for which one can provide some "free" time to the cross street; but it is not enough to serve the cross street traffic, one could assign this time to the cross street and then look for some other appropriate time to serve the rest of the cross street traffic. To take full advantage of the free time, however, one must release the direction 3 vehicles immediately after the termination of this free time in order for them to arrive at m^* at their appointed time. It is not obvious that, to serve more cross street traffic, one should simply extend the cross street green. It may be advantageous to serve the cross street in two intervals per cycle (with an extra lost time in switching). In any case, what one should do depends on the arrival pattern of vehicles in direction 3.

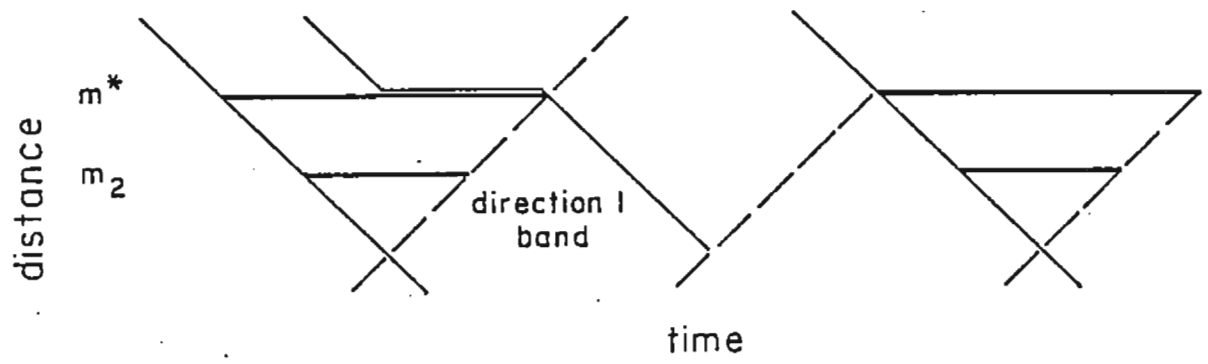


Fig. 4.7 - Trajectories of vehicles in direction 3 and the effect on a signal downstream.

An analogous (but simpler) argument to the above also applies to intersections downstream of m* in direction 3, as illustrated in figure 4.7. If there were no intersections downstream of m*, vehicles leaving m* anytime during the green interval for direction 3 would have trajectories of slope -1, but no such trajectories would leave m* during the red interval. If, at some intersection m_2 downstream of m*, one were to start a red interval for direction 3 after the passage of vehicle 1 in figure 4.7, it would cause no delay to vehicles in direction 3. To prevent interference with the vehicles in direction 1, however, the red interval for a two-phase signal must terminate with the start of the next through band for direction 1. With appropriate adjustments for switching times and turn phases, there is a certain time which can the start of the next through band for direction 1. With appropriate adjustments for switching times and turn phases, there is a certain time which can be assigned to the cross traffic at no penalty to the traffic in direction 3.

If there are turn signals at m* , intersection m_2 would, unfortunately, prefer the opposite sequence of turn phases to that of intersection m_1 in figure 4.6. For a given effective red at m* , one would like to shift the start of the direction 1 band to the right to increase the free time at m_2. If there are turn signals at m_2 , one obviously would prefer a lagging left for direction 3 (after the passage of the direction 1 band), and a leading

left for direction 1 (which would have a negligible effect on the cross street green). Analogous to figure 4.6, the free time available for the cross street decreases as one moves downstream from m^*. For two phase signals, this time vanishes at a trip time from m_2 of about half the red interval at m_2.

Aside from the fact that the intersections upstream or downstream of a critical intersection may prefer different sequences of turn phases at m^*, the two schemes in figures 4.6 and 4.7 are not necessarily independent of each other even for two phase signals. Suppose, for example, that one had two critical intersections m^* and m^{**} as illustrated in figure 4.8 with a trip time between them less than half the red interval. The two triangular regions upstream of m^* and downstream of m^{**} corresponding to figures 4.6 and 4.7 respectively, would now overlap. If there were no intersection between m^* and m^{**}, some typical trajectories would be as illustrated schematically by the curves 1 to 8, 1' to 5' of figure 4.8.

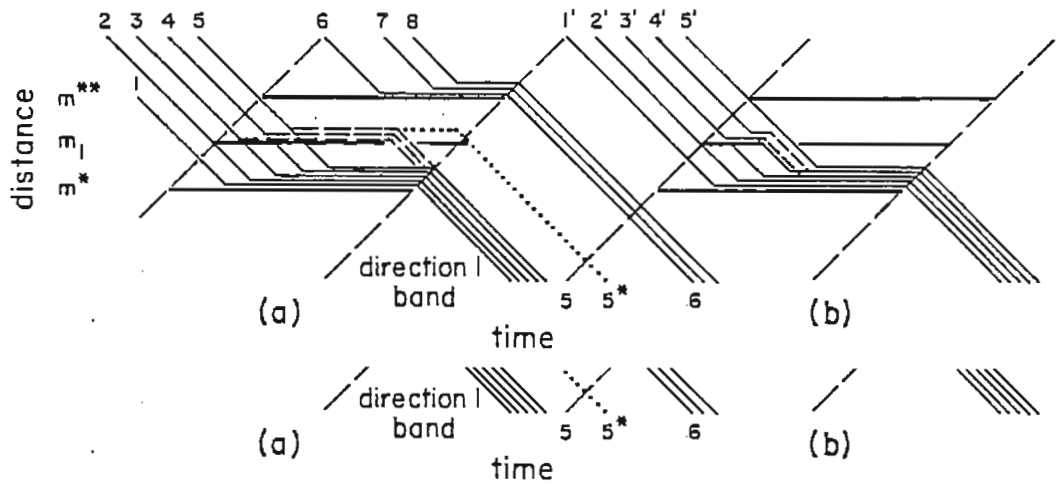


Fig. 4.8 - Trajectories of vehicles in direction 3 passing two critical intersections m^* and m^{**} with an intermediate signal at m_1.

If, now, one has an intersection m_1 between m^* and m^{**}, one would like to set the signal at m_1 so that it causes no additional delays to the vehicles in direction 3, if possible. Cycle (a) of figure 4.8 shows a scheme based upon applying the strategy of figure 4.6 first, i.e., vehicles 3, 4, and

5 are stopped at m_1 and the (first) red interval extends until vehicle 3 will leave m_1 just in time so as to pass intersection m* without delay. Similarly, vehicles 4 and 5 passing m_1 with headway $1/s_3$ will pass m* without delay. Note that this strategy displaces these trajectories into the triangular region corresponding to figure 4.7 downstream of m**. After vehicles 3, 4, and 5 have passed m_1, one can give any remaining time (if any) before the direction 1 band arrives to the cross street in a second red interval. If the flow in direction 3 is sufficiently large and m_1 is sufficiently close to m*, however, the flow in direction 3 after the first red interval could extend into the direction 1 band.

In cycle (b) of figure 4.8 with the same arrival pattern in direction 3 as in cycle (a), the second red interval of the cycle at m_1 starts as in figure 4.7 with the uninterrupted passage of vehicle 5' (analogous to vehicle 1 of figure 4.7). The first red interval starts at the end of the direction 1 band as in cycle (a) but terminates so as to give barely enough time to serve the vehicles 3', 4', and 5' before the start of the second red interval.

In either cycle (a) or (b), the time allocated to direction 3 is barely enough to serve vehicles 3, 4, and 5 or 3', 4', and 5'. If there is enough time between the direction 1 bands to serve the cross street traffic in two segments plus the direction 3 vehicles stopped in the first red interval, then time between the direction 1 bands to serve the cross street traffic in two segments plus the direction 3 vehicles stopped in the first red interval, then one can serve the latter at any time between the extremes shown in (a) and (b). The number of direction 3 vehicles which must be served between the two red intervals is the same in (a) or (b) or anything in between. If the time available is not sufficient to do this, one must displace some direction 3 vehicle to the start of the direction 1 band as illustrated by the dotted trajectory 5* in cycle (a). Even if one cannot serve all the direction 3 vehicles between the two red intervals, it is better to serve some than none.

In figure 4.8 we have chosen some rather arbitrary arrival pattern of direction 3 vehicles to intersection m**, but these vehicles typically leave m* in disjoint platoons (more than one per cycle) with flow s_3. If there were intersections upstream of m**, one would, of course, expect the arrivals at m** also to occur in disjoint platoons. The strategy shown in figure 4.8, if it works, is not very sensitive to the arrival pattern, but whether or not it will work does depend on the relative number of vehicles stopped at m_1 and m*.

If one must provide turn signals, particularly for left turning vehicles in direction 1, most signals are designed with either leading or lagging turn phases. If, however, the pattern of direction 3 vehicles passing m* is like that shown in figure 4.8, one obviously should consider the possibility of having the turn phase for direction 1 in the empty space between vehicles 5 and 6 (or 4 and 5* or 5* and 6) where there is no interference with direction 3 vehicles. Thus, in developing some signal strategy, it may be advantageous first to disregard the turning movements and then see if there are some times when one can introduce the turn phases with minimum interference with the through traffic.

4.8. Special Cases

As a practical matter, it probably is not worthwhile to seek an "optimal" strategy because such a strategy would likely be so sensitive to the flows that

As a practical matter, it probably is not worthwhile to seek an "optimal" strategy because such a strategy would likely be so sensitive to the flows that one would not find it worthwhile to keep changing the settings as the flow changes. Also, pedestrian crossings may prevent one from doing many of the things one would like to do. For any given constraints and geometry of intersections, it should not, however, be very difficult to devise some reasonably efficient strategy if one sketches some typical trajectories of vehicles in direction 3 and investigates the consequences of various actions.

One can attack the problem from two directions simultaneously. On the one hand, one can start by inserting traffic signals at all intersections set

according to a strategy as in figure 4.5 with all the excess time assigned to the cross streets (or turning phases). This will give the minimum delay to the cross street and an upper bound on the minimum travel time in direction 3. By drawing some typical trajectories in direction 3, one can see what benefits would derive by assigning some of the time from the cross traffic to direction 3 starting with those intersections which are least critical. On the other hand, one can start by inserting signals only at the most critical intersections for which one has little choice in strategy, particularly those intersections which clearly prevented one from creating a two-way progression. Since the insertion of additional signals cannot decrease the trip time in direction 3, this would, at each stage, give a lower bound on the minimum average trip time in direction 3.

In principle, the two schemes should converge to the "optimal" strategy, but it should suffice if one can identify some simple but efficient strategy which is not very sensitive to the flow.

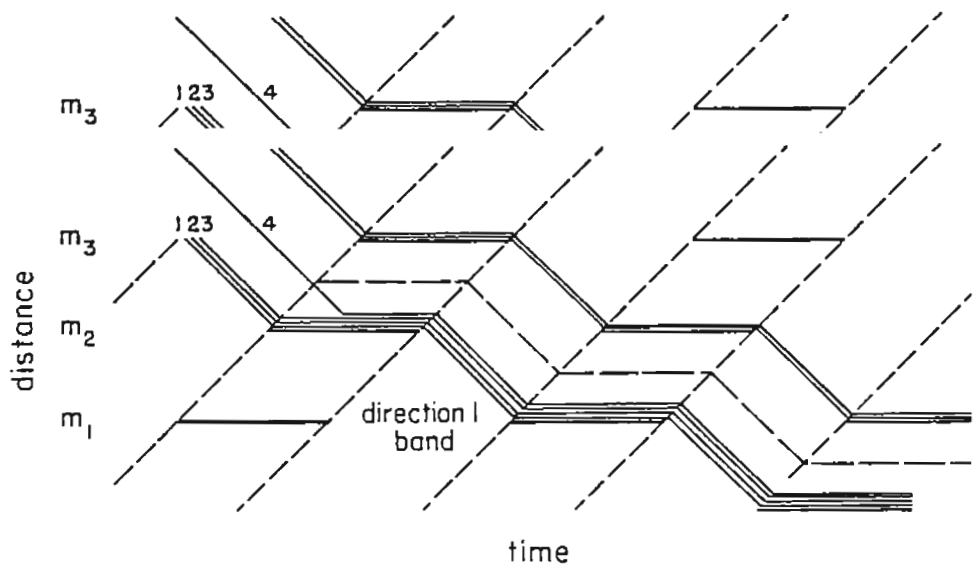


Fig. 4.9 - Signal coordination with maximum delay to direction 3 vehicles.

The "worst" situation is illustrated in figure 4.9. Suppose that the (critical) intersections all have a "spacing" (trip time between intersections) exactly equal to half the duration of the through band. Regardless of when a vehicle in direction 3 enters the system, once it has been stopped at some intersection, it will be stopped again at all subsequent intersections. Furthermore, except for variations in the time lost on entering the sytem, the trip time of this vehicle is independent of the flow and equal to the value described by (4.6.3) that would exist if there were signals everywhere as in figure 4.5. The solid line trajectory of vehicle 4 in figure 4.9 illustrates how this vehicle would behave it it entered the system at some arbritrary time and was stopped only at the critical intersections. The broken line shows the trajectory it would have it there were signals everywhere as in figure 4.5. The extra signals do increase the delay to this vehicle on entering the system, but do not cause any additional delay thereafter.

It is highly unlikely that one would have equally spaced and equally critical intersections. If one did, one would certainly try to avoid choosing a cycle time and splits so that the direction 1 bandwidth has twice the trip time between intersections. This illustration is helpful, however, as a basis for comparison of various strategies with the bounds described in section 4.6. Unbetweenintersections. In figure 4.5, in which travelers traverse the system at
comparison of various strategies with the bounds described in section 4.6. Un-like the pattern shown in figure 4.5, in which travelers traverse the system at more or less arbitrary headways, the vehicles in figure 4.9 are in compact platoons typical of any system with discrete intersections.

Fig. 4.10 shows the same through bands in direction 1 as figure 4.9 but a shorter spacing between intersections. The travel time in direction 3 for this geometry is appreciably less than in figure 4.9 because each stopped vehicle is delayed for a time less than the whole red interval. The travel time in direction 3 is, in fact, the same as if the red interval in direction

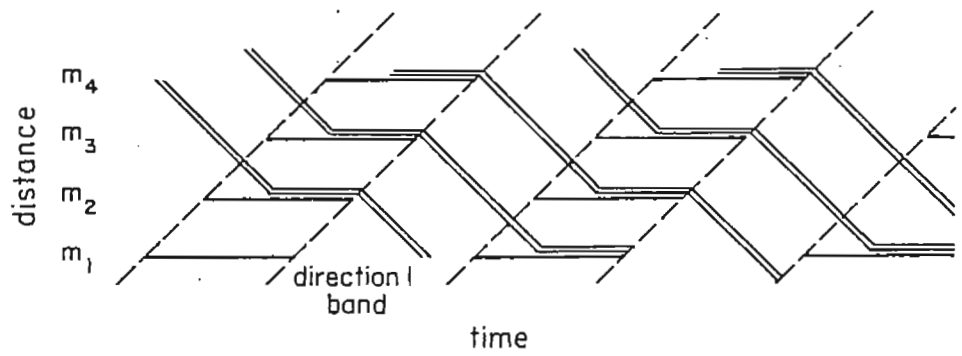


Fig. 4.10 – Signal coordination with an intersection spacing less than half the duration of the direction 1 band.

interval G in (4.6.3) was increased accordingly).

In figure 4.10 a vehicle is stopped either at the odd numbered intersections or the even numbered intersections. If we eliminated all the odd (or even) numbered intersections, the trip time in direction 3 would be nearly the same, but with all vehicles stopped only at the odd (or even) numbered intersections. This, however, would yield a spacing between intersections larger than in figure 4.9. Thus, the travel time for direction 3 in figure 4.9 is larger than if the intersection spacing were either increased or decreased.

In figure 4.10 the flow in direction 3 is light (it uses only a small

creased.

In figure 4.10 the flow in direction 3 is light (it uses only a small fraction of the available time). One can readily see that the trajectory picture would change appreciably if the platoon leaving intersection m_3, for example, at the start of green were so long that the vehicles in this platoon could not clear intersections m_2 before the signal at m_2 turns red. In this case, some vehicles would pass m_2 without delay, but vehicles which are stopped at m_2 will be delayed for the entire red interval. Furthermore, the stopped vehicles will delay the vehicles in the next platoon which left m_4

at the start of green but cleared intersection m_3. Thus, any vehicles which cannot pass a signal at the end of the green interval are shifted to a platoon at the start of the next green.

In the limit of saturated flow in direction 3, one can see from figure 4.10 that the travel time in direction 3 would be the same as (4.6.3) or figure 4.9 as if there were signals at all locations. This, in turn, implies that, for closely spaced intersections and heavy flow in direction 3, nothing much is to be gained by giving additional green time to direction 3 at only a few intersections (obviously one would reduce the delay if one gave more green time at all intersections). Indeed, one might not reduce the delay in direction 3 even if one gave all the available time to direction 3 at some intersection, i.e., in effect, eliminate some signals.

The intersection spacing shown in figure 4.9 (trip time between intersections equal to half the bandwidth for direction 1) is critical also if the flow is heavy in direction 3. The above conclusion that the travel time in direction 3 for saturated flow is independent of the signal spacing applies only for signal spacings closer than in figure 4.9. If, in this case, one draws additional vehicle trajectories in figure 4.10, one sees that the band for direction 1 is completely covered by trajectories in direction 3. If, however, one removes the even (or odd) numbered intersections in figure 4.10 so as to create an intersection spacing larger than in figure 4.9, the travel time in direction

removes the even (or odd) numbered intersections in figure 4.10 so as to create an intersection spacing larger than in figure 4.9, the travel time in direction

3 is less than in figure 4.9. Indeed, the travel time in direction 3 is independent of the flow in direction 3. The delay is now a linearly decreasing function of the intersection spacing, reaching a value of zero when the intersection spacing is sufficiently large as to give a two-way progression.

The details of the above examples are not important because we do not expect to have a large number of equally spaced intersections or to have the same splits at all intersections. They are important only to the extent that

some of the conclusions may be insensitive to geometry. We are particularly interested in identifying situations in which the travel time in direction 3 may be appreciably less than in (4.6.3) and what strategies of giving extra green time to direction 3 at certain intersections may significantly reduce the delays in direction 3. These examples, however, clearly illustrate again that any strategy designed to reduce the delays at any single intersection will not necessarily reduce the total delay for all intersections.

From figures 4.9 or 4.10 or variations thereof, one can see that any reduction in total time relative to (4.6.3) derives primarily from avoiding having vehicles arrive at some intersection just after the signal turns red and being delayed for the entire red interval (if we assume that the red interval is not partitioned into two segments as discussed in section 4.7).

If one had a pair of adjacent intersections with a spacing as in figure 4.9, one could reduce the delay at this pair of intersections either by changing the cycle time or the splits. To change the cycle time may only lead to the creation of similar problems at other intersections. There is typically much more flexibility in the choice of the splits.

It is unlikely that two adjacent intersections would both be critical intersections and equally so with a spacing as in figure 4.9. Suppose that intersection m_2 in figure 4.9 were the most critical intersection so that one could not afford to give any extra arterial green time at m_2. If the intersection m_2 in figure 4.9 were the most critical intersection so that one could not afford to give any extra arterial green time at m_2. If the spacing between m_2 and m_3 were as shown in figure 4.9 and one could afford to give some extra time to the arterial at m_3, then one obviously can reduce the delays in direction 3 if one could release the platoon of vehicles in direction 3 at m_3 a bit earlier so that some of these vehicles could pass m_2 before the signal turns red. Any such vehicles which pass m_2 without delay will likely be stopped at m_1 but only for part of the red interval at m_1. They are not likely ever to be forced back into the rest of the platoon which was stopped at m_1.

Suppose, on the other hand, that intersection m_2 was again the most critical intersection, but the spacing between m_2 and m_1 were as shown in figure 4.9. If one can afford to give some extra time to the arterial traffic at m_1, then it would obviously be beneficial to delay the start of red at m_1 so as to allow part of the platoon from m_2 to pass m_1 without delay.

Of course, if neither m_1 nor m_2 or m_3 is a critical intersection, one has even more flexibility to avoid the situation illustrated in figure 4.9. To compare various strategies, one must look further upstream or downstream of these intersections.

Although it is typically easy to make improvements on the worst situation as illustrated in figure 4.9, it is not as easy to make improvements in the pattern shown in figure 4.10. If, analogously, intersection m_2 is the most critical intersection in figure 4.10, releasing the direction 3 platoon early from intersection m_3 will not decrease the travel time in direction 3 unless this platoon is so long that some vehicles in this platoon are stopped at m_2. Otherwise the platoon passes m_2 without delay and leaves intersection m_1 at a time independent of any (small) variations in the time of release from m_3. Similarly, if the spacing between m_2 and m_1 is as shown in figure 4.10 but the entire platoon leaving m_2 passes m_1 without delay, then nothing is gained by delaying the start of red at m_1.

Generally, a reduction in delay to vehicles at one intersection resulting from a change in splits, will not yield a significant reduction in the total delay in direction 3 unless the action taken will guarantee that some vehicle will clear some intersection that it would not have cleared otherwise (either at the intersection where the action is taken or at some downstream intersection).

4.9. Traffic Responsive Strategies, Two-way Progression

The above discussion has dealt only with pretimed strategies. Stochastic effects were not discussed explicitly although it was implied that the cycle time would be chosen large enough to serve the flows at the most critical intersection. Indeed, one would probably choose a cycle time and splits (including turn phases) at the critical intersection similar to what one would use if this were an isolated signal. Also, the minimum time assigned to any cross street at noncritical intersections would take into consideration stochastic delays on the cross street. One cannot assign to the cross streets barely the minimum time needed to keep them undersaturated.

The number of vehicles arriving in any signal phase will vary from cycle to cycle and for any efficient pretimed signal plan there will be stochastic queueing. The question we wish to consider now is if and how one might use information from vehicle detectors to reduce the stochastic queueing. This question was considered in some detail for a one-way arterial. Many of the qualitative aspects of a one-way arterial carry over to the two-way arterial, particularly if there is a dominant direction of flow and one wishes to maintain a progression for (at least) one direction.

For a one-way arterial we saw that it was very difficult to respond to any stochastic variations in the cross street traffic. By the time one had observed some excess or deficiency of arrivals on a cross street, it was too any stochastic variations in the cross street traffic. By the time one had observed some excess or deficiency of arrivals on a cross street, it was too late to make any worthwhile modifications in the timing for the arterial platoons which were released earlier from various upstream signals. If, for a two-way arterial, one wishes to maintain a progression in (at least) direction 1, it is almost certain that one will need to have the signal phases for the through traffic in direction 1 start according to some pretimed strategy at each intersection. Even if traffic turning onto the arterial causes a (small) queue to form ahead of an approaching platoon, one should not typically modify the

signal timing to respond to this. One should be concerned mostly with the timing for the vehicles at the rear of the platoon rather than the front.

For a one-way arterial, about the only stochastic properties of the traffic flow to which one could respond was that the number of vehicles in an arterial platoon would vary from cycle to cycle. A pretimed plan would need to allocate more than the average green interval needed for the platoon in order to prevent excessive stochastic queueing of the arterial traffic. A traffic responsive system, however, can terminate the arterial signal phase as soon as it observes that the arterial platoon has passed that signal, so that the average time allocated to the arterial platoon is (nearly) equal to the average time needed. The excess time that would have been allocated to the arterial flow in a pretimed plan can be assigned to the cross street. This will reduce the stochastic queueing on the cross street which, in turn, may also allow one to use a shorter (preset) cycle time for the whole system. Stochastic queueing on the arterial is also (nearly) eliminated.

For a two-way arterial, the situation is much more complicated because one not only must consider the traffic in direction 3, but also possible turning movements, particularly from directions 1 and 3. Most signal phases will be serving two movements simultaneously.

If, per chance, one could create a two-way pretimed progression over some section of the arterial as in figure 4.2 with bandwidths in the two directions

If, per chance, one could create a two-way pretimed progression over some section of the arterial as in figure 4.2 with bandwidths in the two directions sufficiently wide to accommodate the flows in directions 1 and 3 with a bit to spare and there was still enough time left to accommodate the flows on each cross street (without excessive stochastic queueing), then one can certainly devise a traffic responsive strategy which will give less delay due to stochastic queueing than the pretimed strategy. As with the one-way arterial, the goal is to reduce queueing on the arterial and also give as much time as possible to the cross street.

In the traffic responsive strategy the signal phases serving the through traffic movements in directions 1 and 3 should each start according to the pretimed plan. There are several cases to consider, however, depending on if or how the through bands overlap and the sequence of turning phases (if any) at each signal.

(a) If at some intersection the pretimed bands illustrated in figure 4.2 for directions 1 and 3 do not overlap and the cross street traffic is served in two disjoint intervals per cycle, the left turning vehicles in direction 1 can be served at the same time as the through vehicles in direction 1; similarly for direction 3. Typically, one need not worry about the turning traffic.

In a traffic responsive modification of the pretimed plan, the green interval for direction 1 would start at the preset time but terminate as soon as the platoon clears the intersection (or the green interval has reached some preset maximum at least as large as for the original pretimed strategy); similarly for direction 3. The preset maximum times may be chosen so as to provide at least some minimum acceptable time intervals for the cross streets.

As for the one-way arterial, this strategy reduces or eliminates any stochastic queues for both arterial directions (depending on how large one can set the maximum arterial green intervals) at the intersection in question. It also reduces the stochastic queueing on the cross street because the average time available for the cross street is larger than for the pretimed plan.
reduces the stochastic queueing on the cross street because the average time available for the cross street is larger than for the pretimed plan.

(b) If at some intersection the pretimed bands for directions 1 and 3 overlap with the direction 3 band preceding the direction 1 band, then the pretimed strategy probably gave a leading left phase for direction 3 and a lagging left for direction 1. If this is the case and the pretimed strategy provides adequate time for the left turns from direction 3, the traffic responsive strategy would use the pretimed strategy for the leading left phase. When the direction 3 platoon has passed (presumably after the direction 1 green has

already started) or some preset maximum time is reached, the signal may switch to a lagging left for direction 3 if the pretimed plan provided such a phase. In any case, there should typically be ample time to serve the turning traffic from direction 1, at least an ample average time. If so, one can now terminate this phase as soon as the direction 1 platoon has passed (or some preset maximum time is reached).

(c) If the through bands overlap but the direction 1 band precedes the direction 3 band, the pretimed plan may or may not have given a leading left to direction 1. In either case, the traffic responsive strategy would follow the pretimed strategy until the signal was green for both directions 1 and 3. It would probably be advantageous next for the traffic responsive system to follow the same sequence of phases as in the pretimed plan but to terminate each phase as soon as it was no longer needed (or the phase lasted some preset maximum time). Since the platoons in direction 3 are presumably shorter than in direction 1, the direction 3 platoon may terminate before the direction 1 platoon. In this case there would be additional time available for turning vehicles from direction 1, with or without a turn signal. If one provides a turn signal, however, one must make some special provision for the left turn vehicles from direction 3, because one does not wish to serve them following a protected turn phase for direction 1. In any case one can terminate the phase for the direction 1 through traffic as soon as the platoon passes (or the time a protected turn phase for direction 1. In any case one can terminate the phase for the direction 1 through traffic as soon as the platoon passes (or the time has reached some preset maximum). If this phase is followed by some turning phase, the latter can also be made traffic responsive and terminate when the flow drops. Whatever the sequence, the traffic responsive plan would terminate on the average earlier than the pretimed strategy which needed to provide some slack time to accommodate fluctuations.

If in the pretimed plan the green interval for direction 3 continues after that for direction 1, one probably also had a leading left for direction 1 of

sufficient duration to serve all the turning traffic in direction 1. After the direction 1 platoon passes there will also be some time available for direction 3 turning traffic, with or without a turn signal. As soon as the platoon in direction 3 passes; and, if necessary, one has provided some extra time for the left turn vehicles from direction 3, one can switch the signal to the cross traffic.

The above schemes, of course, presuppose that one can establish a two-way progression. Since the stochastic queueing for the traffic responsive strategy is considerably less than for a corresponding pretimed strategy, there may be situations in which one can satisfactorily operate a traffic responsive two-way progression with a shorter cycle time than one would tolerate for a completely pretimed strategy. Unfortunately, the more common situation is that one cannot establish a two-way progression at all.

4.10. Traffic Responsive Strategies, One-way Progression

If one cannot establish a two-way progression, one would probably like to have a progression in at least direction 1. In developing an efficient pretimed strategy, one would have considered the potential benefits to the traffic in direction 3 resulting from giving extra time to direction 3 at various intersections at various times in the cycle, including the possibility of interrupting the cross street flow and serving the cross street in two segments per cycle. Whereas the penalty resulting from giving extra time to direction 3 at intering the cross street flow and serving the cross street in two segments per cycle. Whereas the penalty resulting from giving extra time to direction 3 at intersection i is only to the cross street traffic at intersection i, the benefit to the traffic in direction 3 depends on whether or not the local benefits to these vehicles translates into net benefits downstream.

Any traffic responsive strategy at intersection i which balances only the local benefits to various traffic movements at intersection i without considering the consequences at other intersections is not likely to be as efficient as a pretimed plan which does consider the interactions. It is

reasonable to assume, therefore, that an efficient traffic responsive system would merely attempt to make appropriate adjustments to some pretimed plan at intersection i based upon information obtained from detectors at intersection i.

There would, of course, be a hierarchy of more complicated strategies in which a signal at intersection i would respond to information obtained not only at intersection i but also at neighboring intersections. It is not obvious, however, that information obtained from neighboring intersections would be useful and, if so, could be obtained soon enough to influence one's response at intersection i. As was true for the one-way arterial, information obtained about the current status of the cross street traffic at intersection i cannot typically be used to modify the timing of the arterial platoon which was released upstream. It is reasonable to assume, therefore, that the start of the green for the through traffic in direction 1 at various intersections should follow a pretimed plan.

In designing the pretimed plan, one tries to avoid, wherever possible, having the arterial green terminate so as to cut off part of a direction 3 platoon. One might have even designed the plan for a somewhat higher flow in direction 3 than actually exists so that there will be some intersections at which the direction 3 flow vanishes near the end of the arterial green for direction 1. The plan also allowed more green time for direction 1 than the which the direction 3 flow vanishes near the end of the arterial green for direction 1. The plan also allowed more green time for direction 1 than the average duration of the platoon. Thus, in a significant fraction of the cycles, the direction 1 platoon will terminate before the pretimed end of the green. If now the flow in <u>both</u> directions 1 and 3 should vanish before the end of an arterial green at some intersection, there is no benefit to allowing the arterial green to continue. Any unused time can be given to some other signal phase and eventually to the cross street.

If the cross street traffic arrives at a uniform rate, any excess time given to the cross street will certainly result in a net benefit even if the queue vanishes on the cross street. It will give a larger benefit if there is stochastic queueing on the cross street.

At some other intersections the pretimed plan may have assigned some extra time to the direction 3 traffic outside the direction 1 band, either at the end of the band or possibly interrupting the cross street flow. In either case, the direction 3 flow might terminate before the alloted time expires, at a time when there is no flow in direction 1. One can, therefore, terminate this phase when the flow vanishes and give any excess time to the cross street. (Note that there is no analogous strategy if one gives extra green time to direction 3 ahead of the direction 1 band). If there are special phases for turning traffic in directions 1 or 3 in the pretimed plan, it would also be advantageous usually to terminate these phases early if the detectors indicate that the queue for turning vehicles has been served, at least in those situations for which the subsequent phase serves direction 2.

All the above modifications of the pretimed plan are guaranteed to give a reduction in the delay to the cross street traffic at no expense to any other traffic, thus a net gain for the system.

The original choice of a pretimed strategy involved various compromises between the average delays in directions 2 and 1 or 3. If, as a result of the

The original choice of a pretimed strategy involved various compromises between the average delays in directions 2 and 1 or 3. If, as a result of the above traffic responsive modifications at interseciton i, one reduces the average delay in direction 2 at intersection i, one could revise the original pretimed plan giving less scheduled time to direction 2 at intersection i and more time to direction 1 and/or 3. One would, of course, need to evaluate the effect of any such change at intersection i on the delays at other intersections. If, for example, the saving in time at intersection i resulted from terminating the direction 1 green ahead of schedule in some cycles, one could

give the excess time back to direction 1 by allowing the direction 1 green to run to a maximum larger than in the original plan so as to reduce stochastic queueing in direction 1 when there is an excess of traffic in some cycle. If, however, one gives extra time to direction 1 at only a few intersections, there is no guarantee that extra vehicles which are allowed to pass intersection i will also be able to pass intersections downstream with no delay.

The above traffic responsive improvements apply only to some of the intersections and to certain signal phases. In particular, no modification of the pretimed plan was proposed at an intersection where the pretimed plan would terminate the direction 1 and 3 phases simultaneously at a time when this was likely to cut off a direction 3 platoon. Particularly at the critical intersections with relative large cross street traffic, the pretimed plan may provide no extra time for direction 3. Furthermore, if the flow in direction 3 covers a significant fraction of the arterial through band, it may be impossible to avoid cutting off some direction 3 platoon, regardless of how one adjusts the timing of signals upstream. In this situation any traffic response strategy which reduces the delay for one traffic direction is likely to increase the delay for some other direction.

Since the pretimed strategy has already been designed to give favorable treatment to direction 1 and then to do the best it can for directions 2 and 3, perhaps one would like to go one step further and give the best possible ser-

treatment to direction 1 and then to do the best it can for directions 2 and 3, perhaps one would like to go one step further and give the best possible ser-

vice to direction 1 even with a stochastic demand in direction 1. Then consider what one can do for the other movements. Suppose that the direction 1 green intervals start according to a pretimed plan and, at <u>every</u> intersection, the green extends at least until the direction 1 platoon has passed (or some generous preset maximum time has expired). This would virtually eliminate stochastic queueing in direction 1 and overcome the problem noted above that, if one allows some extra direction 1 vehicles to pass one intersection, they

At each intersection, the time not used by direction 1 can now be allocated between directions 2 and 3. This time will vary from cycle to cycle (and even from one intersection to the next), but it will have an average value at each intersection larger than it would be for any pretimed plan which had to provide some slack time in direction 1 to control stochastic queueing. If, at the critical intersection, one allocated all of this time to direction 2, the stochastic delays for direction 2 would typically be even less than for a pretimed plan, as was explained in section 3.5 for a one-way arterial. This strategy, however, is expected to be less advantageous for direction 3 than a pretimed strategy, depending on how one can set the noncritical intersections for direction 3, because, by reducing the average time for direction 1, one has also reduced the average time available for the direction 3 vehicles to travel in the direction 1 band. Because of the reduced delay in direction 2 at the critical intersection, however, one may now be able to use a shorter cycle time for the whole system, which is likely to reduce the average stopped time for direction 3.

At the noncritical intersections the average time available for direction 2 is more than it needs, so one can afford to reassign some of the time to direction 3. The situation now is similar to that described in section 4.7 and 4.8 for a pretimed plan except that the direction 1 bands are variable from cycle to cycle (and not quite of uniform width at all intersections) with and 4.8 for a pretimed plan except that the direction 1 bands are variable from cycle to cycle (and not quite of uniform width at all intersections) with a mean value less than one would have used for a pretimed plan. Also at the time when one must make a decision at some ith intersection of how much time to assign to direction 3 one does not know the actual width of the subsequent direction 1 band (unless one can relay information from upstream intersections after the direction 1 platoon has passed an upstream intersection). One only knows its probability distribution (a mean and variance). Neither does one know precisely the future arrivals in direction 3, or any past traffic responsive actions taken at other intersections.

There will be some intersections with such light cross street traffic that they cause no problems. One can assign time to direction 3 in such a way as to cause no extra delay to direction 3 and still have ample time to serve direction 2. The signal may, in effect, be part of a two-way progression in which one may have assigned time to the cross street in either one or two segments per cycle, or one may have a situation analogous to figure 4.6 or 4.7. We are interested here primarily in situations in which there is a genuine conflict between directions 2 and 3.

Suppose one is considering if and how much to extend the arterial green for direction 3 traffic at some intersection after the direction 1 platoon has passed; for example, at an intersection downstream in direction 3 from a critical intersection. At the time the arterial platoon in direction 1 has passed, one knows how much time is available until the start of the next (pretimed) green for direction 1. From observations in previous signal cycles or estimations from the strategy at upstream intersections, one can evaluate the average number of direction 3 vehicles expected in the arriving platoon but one cannot know it precisely without updated information from upstream intersections. With special equipment one could estimate the queue length on the cross street but with most equipment, one could only estimate the number of vehicles between the detectors (if any) and the intersection. One does know cross street but with most equipment, one could only estimate the number of vehicles between the detectors (if any) and the intersection. One does know the (average) cross street flow per cycle.

We expect that there will be some stochastic queueing on the cross street because the cross street green will terminate at a preset time whether there is a queue then or not. If in some signal cycle we know that the available time is insufficient to serve both the direction 3 platoon and the cross street queue, we must have some rule for partitioning the available time. The rule could provide a fixed time for direction 2 or a variable time depending on the time

available but it must give to direction 2 an average time adequate to keep the stochastic queue at an acceptable level. Having adopted some such rule, one can then make some traffic responsive modifications. If during some cycle the platoon in direction 3 should terminate before the specified time expires, one can end the direction 3 phase early and give the excess time to direction 2. Also, if, when the specified time for direction 3 expires, one has a count of the queue length in direction 2 (because the queue does not extend beyond the detectors) and the time assigned to direction 2 exceeds that needed, one might extend the time for direction 3.

Clearly, it would be rather difficult to evaluate an "optimal" strategy theoretically, but it should not be too difficult to recognize deficiencies in an existing strategy and to make appropriate corrections.

Suppose now that one wishes to provide a leading green for direction 3 at some intersection; for example, at an intersection upstream from the critical intersection, so that some direction 3 vehicles will be able to clear the critical intersection (or some other downstream intersection) without delay. At the intersection in question, the cross street green presumably begins after an arterial platoon has passed in direction 1 (and one has provided turn phases, if necessary), at which time one knows how much time is available until the preset time for the start of the next green for direction 1. This should typically be more than enough time to serve the average cross street queue.
time for the start of the next green for direction 1. This should typically be more than enough time to serve the average cross street queue.

The penalty for terminating the cross street green before the queue vanishes is quite high. The residual queue must wait a whole cycle time and one still must find some (better) time to serve it in a later cycle. The benefit for giving some extra time to direction 3, however, is that some vehicles might then be able to clear a downstream intersection and save some fraction of a red time. But one does not know for sure if this will happen, because one does not

know precisely the width of the subsequent arterial band. If one has detectors on the cross street, it would seem, therefore, that one should allow the cross street green to extend at least until the queue vanishes (or all the available time has expired).

If in some cycle one has an excess of available time, one might consider continuing the cross street green after the queue has vanished (or interrupting the cross street green to serve the cross street in two segments). This would at least save a whole cycle time of delay for any vehicles which might arrive in direction 2 after the queue vanishes, and eliminate the need to serve it in the next cycle which might have less available time. Even if the benefit from giving such time to direction 2 should exceed the benefit from giving it to direction 3, maybe one would give it to direction 3 anyway because that is the simplest thing to do, and it is likely to give some benefit to direction 3.

It should be emphasized again that one should not apply the same strategy at all intersections. The most important aspect of the above scheme is that one starts with some tentative pretimed plan with a band in direction 1. With the available time for direction 3 at the critical intersections limited by the cross street demands, one then tries to find ways to avoid having direction 3 vehicles delayed for a whole red time at any intersection or experience any extra delay at all due to minor intersections.

vehicles delayed for a whole red time at any intersection or experience any extra delay at all due to minor intersections.

This determines the global strategy of what one is trying to do at each intersection. The final step is to modify this pretimed plan by a traffic responsive strategy in which each signal responds to the local conditions making sure that any time not being used by one traffic movement is reassigned to some other movements. One will reassign time to direction 3, however, only if this action is likely to allow some direction 3 vehicle to pass an intersection downstream in an earlier cycle than it would otherwise.

4.11. A Typical Plan

In Chapter 5 we will propose that if one has a collection of parallel roads in some urban network and there is an imbalance of flow with the flow predominantly in direction 1, then one should set the signals on most (but not all) roads to give a progression in direction 1. Some (possibly one-way) roads should have a progression in direction 3. The argument is that if each driver chooses the fastest route between his origin and destination and he has the choice of using any of several possible roads in directions 1 or 3, he will choose a road with a progression in his direction of travel, if there is one. If one provides some roads with a progression in direction 3, they should carry most of the traffic in direction 3. Any two-way road with a progression in direction 1 is likely, therefore, to have relatively light traffic in direction 3 probably also with relatively short trip lengths along that road.

The purpose of the argument here is to point out that it is usually not realistic to minimize the total travel time of all vehicles for fixed flows along a single arterial, if this arterial is part of a network in which drivers have alternative routes. In particular, there is some question as to how much weight one should give to delays in direction 3 if the flow is predominantly in direction 1. Obviously, if one can reduce the delay in direction 3 with negligible effect on any delays for other directions, one should do so. Even if one could establish a two-way progression, but with a restriction on the

negligible effect on any delays for other directions, one should do so. Even if one could establish a two-way progression, but with a restriction on the combined widths of the two through bands, one would likely give a preference to one direction and hope that any demand in the other directions which cannot fit in the band will find other routes (which have been provided for it).

Under these conditions, the first stage in the development of a signal plan is to establish a progression for one direction (direction 1) similar to that described in Chapter 3 for a one-way arterial. One may, however, need to make some allowance for turning traffic (particularly left turns from direc-

tion 3). We assume that one measured or estimated all the relevant q/s values for all movements at all intersections (and suitable effective lost times per cycle), and that it is possible to operate all signals so as to be undersaturated. The proposed sequence of steps starts as in section 3.6.

1. Identify the most critical intersections. These are the intersections which would require the longest cycle time if they were isolated signals. One must consider here which intersections my require multiphase signals as described in section 2.11. These issues are actually independent of any possible coordination scheme, and the critical intersections are likely to be those which require turn signals.

2. For the most critical intersection, determine a cycle time (or a range of possible cycle times) C and splits for all signal phases that one would use if this intersection were an isolated intersection with stochastic arrivals in all directions. For a pretimed plan this C should be comparable with twice the minimum value which could accommodate all flows. If a two-phase signal gave an average of more than 3/2 turning vehicles in any direction per cycle, one may need separate turn phases.

If one needs a multiphase signal (particularly for directions 1 and 3), the t_1 and ℓ_3 movements (also the t_3 and ℓ_1) must be served in separate phases. We expect, and will assume here, that the average time needed per the t_1 and ℓ_3 movements (also the t_3 and ℓ_1) must be served in separate phases. We expect, and will assume here, that the average time needed per cycle by the t_1 plus ℓ_3 flows exceeds that needed by the t_3 plus ℓ_1 flows (although we could easily deal with the opposite case). If the "optimal" cycle time C for an isolated F-C signal is larger than acceptable, one might discount some of the stochastic queueing in anticipation that one can reduce the stochastic queueing by some traffic responsive strategy.

If the signal at the critical intersection is to be coordinated with other signals, a traffic responsive strategy will have a specified C and a pretimed start for the through traffic in direction 1, but a traffic responsive split.

This would virtually require that the critical intersection have a lagging left for direction 3 if one wishes to salvage as much of the time as possible not needed by directions 1 and 3 to give to direction 2 (and 4). The through phase for direction 1 would end promptly when the platoon passes (or the interval exceeds some preset maximum). If, at that time, there are two or less $\ell 3$ vehicles waiting, one can skip the $\ell 3$ turn phase and expect these vehicles to pass during the yellow. Otherwise one will provide a lagging left (preferably without a turn arrow) for direction 3 terminating when the $\ell 3$ queue vanishes, independent of any possible flow of $t3$ vehicles. This will guarantee that direction 2 receives any time not needed by the $t1$ and $\ell 3$ movements.

It is assumed here that the <u>average</u> time needed to serve the $t1$ and $\ell 3$ movements is adequate to serve the average $t3$ plus $\ell 1$ flows. If, as in figure 4.8, the timing of the neighboring signals is such that there will be some idle time between platoons arriving in direction 3 (during the $t1$ phase), one might be able to insert a protected $\ell 1$ phase in the gap between direction 3 platoons. Otherwise, at the expense of some additional delay to the direction 3 platoons, one would probably just use a leading left for direction 1. If this phase were traffic responsive, it would terminate when the $\ell 1$ queue vanishes, possibly continuing, however, at least until there has been a call from the $t3$ direction. We expect that there will be delay to platoons in direction 3 and possibly some (but not much) stochastic queueing for the $t3$ direction. If the direction 2 phases, however, are also traffic responsive and the queue vanishes in direction 2 before the scheduled start of the $t1$ phase, one may give some extra time to the $t3$ direction (particularly if there is a queue).

If the cross street at the critical intersection is itself part of a signal progression in the cross direction, one would also like the cross street green for the through traffic to start no later than some scheduled arrival time of a platoon in the direction of the progression for the cross street. If one

needs a lagging left for direction 3, this may cause some loss in efficiency for a traffic responsive signal because one would now need to have a prescheduled minimum time for the $\ell 3$ phase and thus a constraining maximum time for the $t1$ phase to guarantee that the $\ell 3$ traffic is adequately served before the scheduled switch of the signal to direction 2. There would likely be stochastic queueing on both the arterial (directon 1) and the cross street, but one could at least make sure that any time not used by one direction is used to reduce the queue in another direction.

3. As in the case of a one-way arterial, one should next consider which intersections other than the most critical intersection could operate on a cycle time $C/2$. If the critical intersection requires a multiphase signal, this is a particularly important consideration because the multiphase signal typically requires a much longer cycle time than a two phase signal.

Since the average number of turning vehicles per cycle is proportional to the cycle time, a cycle time of C at some intersection other than the critical intersection might necessitate a multiphase signal there also, even though this intersection could operate as a two phase signal at a cycle time of $C/2$. Using a cycle time of $C/2$ will not only reduce the delay to the cross street and turning traffic, it may also increase flexibility in timing signals for direction 3.

cross street and turning traffic, it may also increase flexibility in timing signals for direction 3.

4. As for the one-way arterial, one should measure the trip times between intersections for typical vehicles at the rear of platoons in direction 1. This will define the progression speed for direction 1. For a completely pretimed plan it will define the times for termination of $t1$ signal phases at each intersection relative to the critical intersection, thus the ending time of the direction 1 band. For a traffic responsive signal it will serve as a reference time to determine the pretimed starts of the $t1$ phases at each intersection.

5. If vehicles turn on and off the arterial, the $q_{t1}^{(m)}$ will vary with m, and so, therefore, will the average time needed to accommodate the direction 1 band. As for the one-way arterial, adjustments in the width of the band should be made at the start of the band relative to the ending times determined in step 4.

The procedure for doing this for a one-way arterial as described in section 3.2 i and j was somewhat flexible. It provided some extra time for the arterial direction to allow for the uncertainty in the number of turning vehicles upstream or downstream of critical intersections, but otherwise restricted the direction 1 band to accommodate only those vehicles which could pass the critical intersection. Any time not needed by the arterial traffic was given to the cross street, even at intersections with light cross traffic.

For a two-way arterial we could start with the same direction 1 band we would use for a one-way arterial, but with a width restricted by the time allotted at the critical intersection(s) to the through traffic in direction 1. At a later stage some of the time which would be generously allocated to direction 2 at noncritical intersections for a one-way arterial may be reassigned to direction 3 at appropriate times. Indeed at some intersections we may even squeeze the direction 1 band a little tighter or displace the assigned to direction 3 at appropriate times. Indeed at some intersections we may even squeeze the direction 1 band a little tighter or displace the off-sets a little if this would give a significant benefit to direction 3.

Here, as for the one-way arterial, is where the present theory departs significantly from the theory incorporated in most computer programs. The latter (incorrectly) insert a stochastic delay at <u>every</u> intersection as if each intersection were an isolated intersection, despite the fact that the stochastic delays at noncritical intersections in the arterial direction are constrained by what can pass more critical intersections. The result of this is that these programs will assign extra time to the arterial band at non-

critical intersection to accommodate the exaggerated fluctuations (and record this as a benefit). If the program includes platoon spreading (which is likely also to be exaggerated), it is further inclined to assign more time to the arterial to allow the platoon to spread despite the fact that the platoon will later be compressed when it reaches a more critical intersection. The program may then record a benefit for this from reduced "stops."

The consequences of assigning unnecessary time to the direction 1 band may be considerably larger for a two-way arterial than a one-way arterial. For a one-way arterial, the benefit from assigning extra time to the cross street may be marginal at noncritical intersections where there is negligible stochastic queueing on the cross street. For a two-way arterial, however, the extra time might be reassigned to direction 3, possibly splitting the cross street green into two intervals.

6. With a tentative specification of the direction 1 band, it is now time to consider the partition of the remaining time at each intersection among directions 2, 3 and turning phases (if any). For this, it is advantageous to construct a "dimensionless" time-space diagram as described in section 4.2 based upon the typical trip times between intersections in directions 1 and 3 (for vehicles at the rear of platoons). One must also know the q/s values for the cross streets, the time needed for turn phases, and the effective lost (for vehicles at the rear of platoons). One must also know the q/s values for the cross streets, the time needed for turn phases, and the effective lost times in switching so that one can assess how much excess time is available for possible reallocation to direction 3 at each intersection.

As explained in sections 4.7 and 4.8, one can, on the one hand, start by drawing some platoon patterns for direction 3 that would exist if there were signals at all intersections set with all time outside the direction 1 band assigned to direction 2 or required turn phases. One can then identify where giving some extra time to direction 3 can be beneficial. On the other hand, one could also draw some platoon patterns that would exist if there were

signals only at critical intersections or at intersections which obviously cause problems in constructing through bands for direction 3, particularly omitting minor intersections which are close to critical intersections. One can then try to insert signals at the other intersections with a minimum description of traffic in direction 3.

It is not possible to give a general recipe of what one should do in all circumstances, but in any particular situation, one should be able to recognize from the time-space diagram where the problems are and what options are available to remedy them.

7. We have already anticipated in step 2 that, in order to reduce the cycle time, one might use a traffic responsive strategy at the most critical intersection, giving any time not used by direction 1 to direction 2 (and vice versa if the cross street at the critical intersection is part of a coordinated signal system in the cross direction). It would also be advantageous to do this at any other (less) critical intersection where there may be stochastic queueing in direction 2 despite having all the time not used by direction 1 (and $\ell 3$) assigned to direction 2. To implement this would require vehicle detectors only for direction 1 (and $\ell 3$) at the critical intersections.

If one had vehicle detectors in direction 2 at critical intersections, one could also switch the signal back to directions 1 and/or 3 if the queue in

If one had vehicle detectors in direction 2 at critical intersections, one could also switch the signal back to directions 1 and/or 3 if the queue in direction 2 vanished before the scheduled time required to serve the direction 1 band. Such a strategy, however, is not guaranteed to give a net benefit to directions 1 and 3 unless there is stochastic queueing in direction 1 or 3. One would need to check the time-space diagram for the pretimed plan to see if this is worthwhile (at the expense of a likely increased queue in direction 2 in the next cycle).

It is typically advantageous at critical intersections to have traffic responsive turn phases (if required), particularly if any time saving can

directly or indirectly be assigned to direction 2. The reasons for this are that (a) turning movements typically have low saturation flows and, therefore, require a disproportionally large fraction of the cycle time per vehicle, (b) the number of turning vehicles per cycle is typically small, but the fractional fluctuations are therefore relatively large as compared with the through movements, and (c) if the turn bays have restricted lengths, one must virtually eliminate stochastic queueing for the turning vehicles so they will not overflow the turnbays and block the through lanes. Because of all these effects collectively, a pretimed turn phase operates rather "inefficiently."

Whether or not it is advantageous to have traffic responsive signal phases at other intersections depends on the pretimed plan. If at some (noncritical) intersection one finds it advantageous to give extra time to direction 3 (possibly splitting the cross street green into two intervals), at the expense of causing some stochastic queueing in direction 2, then it would be advantageous to reduce this queueing by giving back to direction 2 any time not used by direction 1 or 3. The benefit from this, however, may not justify the additional cost of the detectors and controllers needed to implement such a strategy.

5. COORDINATION OF SIGNALS IN A NETWORK

5.1. Introduction

If one wishes to coordinate signals in a network of roads, one cannot generally choose a coordination of the signals on one arterial independent of the coordination on other arterials. Suppose, for example in figure 5.1 the lines a-b, b-c, c-d, and d-a represents sections of four arterials, each with a coordinated signal system with a common cycle time. There may also be other roads and signals between those shown in the figure. If one chooses some "optimal" off-set between signals at a and b, and between b and c, this will also specify the timing of the signal at c relative to a. On the other hand, if one chose the optimal off-set of d relative to a and c relative to d, this would also specify the timing of c relative to a, but this timing of the signal at c may not be compatible with that determined via the path a to b to c.

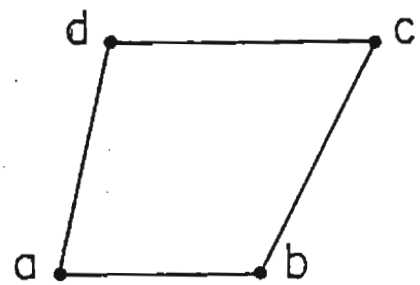


Fig. 5.1 - Geometry for

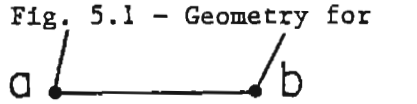


Fig. 5.1 - Geometry for loop conditions.

More generally if there is more than one path along a network by which one can travel from one signal to another, the relative timing of the two signals determined from the cumulative off-sets along a path must be uniquely specific independent of the path. Equivalently, one can also say that if one

traces any "loop" as in figure 5.1 from a to b to c to d to a, the relative timing of successive signals must be such that one comes back to the same timing of a from which one started, i.e., the sum of the off-sets around the loop must add to an integer multiple of the cycle time. This, of course, must be true for every loop in the network.

It is possible to have traffic responsive signals anywhere in the network but, as was true even for a single arterial, it is not generally possible to devise an efficient plan in which the cycle time C is allowed to vary from cycle to cycle in response to fluctuations in traffic. At each intersection at least one signal phase should start at a periodic scheduled time, usually the start of green for one (or both) of the through traffic movements. The off-sets along any road can then be interpreted as the off-set between the scheduled phases (not necessarily "corresponding" phases at different intersections). The loop conditions would then require that the sum of the off-sets between the scheduled times around the loop back to the same scheduled phase at the starting point should be an integer multiple of C. This does not exclude the possibility that some signals might serve some or all traffic movements twice per cycle. They could, for example, operate on a cycle time of C/2. The loop conditions would, however, apply only to scheduled events which are periodic with period C. It is not necessary that all signals have the same splits or the same sequences of turn phases (if any).

are periodic with period C. It is not necessary that all signals have the same splits or the same sequences of turn phases (if any).

In most of the literature dealing with coordination of signals on a network, it is postulated that the flows are specified on all links of the network (including turning movements), and that these flows are independent of the signal strategy. The total travel time (delay) in the network is expressed as the sum of the travel times on all links or arterials of the network, and the travel time on each link or arterial is represented as a function of the signal plan on the individual links or arterials. For simplicity, it is usually postulated

that the travel time (including delays) associated with any link between adjacent signals can be approximated as a function of the signal settings of only these two signals, independent of the off-sets between nonadjacent signals. The mathematical complication, however, comes mainly from the fact that the setting for each signal along one arterial affects the setting of the signal also for the cross street which is part of another arterial coordination in the cross direction. In addition, one has all the "loop constraints" mentioned above.

The problem is usually formulated as a very large mathematical program with a very large number of constraints. Various iterative schemes are used in an attempt to determine the (feasible) signal settings which minimize the total travel time in the system. For example, one might try to minimize the sum of the delays on all approaches to a single intersection for given settings of signals at adjacent intersections in all directions. Of course, a change in the setting of one signal to obtain a "local optimal" at that signal will affect the delays at the adjacent signals. The latter will be readjusted in subsequent iterations.

There are three major deficiencies in this approach:

1. The formulas used for the delays are so idealized as to be highly questionable. In particular, they typically include an exaggerated

1. The formulas used for the delays are so idealized as to be highly questionable. In particular, they typically include an exaggerated stochastic delay at each intersection, as described in previous chapters.

2. The numerical procedures do not give much insight into the issues involved, in particular, how the signal coordination and travel times depend on the geometry of the network.
3. The assumption that the flow pattern is independent of the signal coordination is completely unrealistic.

As regards the third point, one could (and people have) proposed a more general formulation in which one specifies the origins and destinations of trips within the network (or at boundaries of the network), and then combine the signal coordination with a traffic assignment scheme. This, however, leads to a (nonconvex) mathematical program of enormous complexity. None of these computational schemes appears to be capable of identifying, even qualitatively, a reasonable "global" strategy (with reassignment of flows). At best they can only "fine tune" some proposed strategy.

We will be concerned here mostly with some coordination schemes for certain idealized geometries and with some qualitative issues.

5.2. Special Geometries, Rectangular Grids

a. No loops

Suppose one wished to coordinate the signals along just one arterial running E-W, for example, (either one-way or two-way) and several arterials running N-S as illustrated in figure 5.2 by the solid lines. There might be other roads indicated by the broken lines but, if so, we are not concerned with the travel times on these roads. They may be simply access roads to residential areas and carry negligible through traffic.

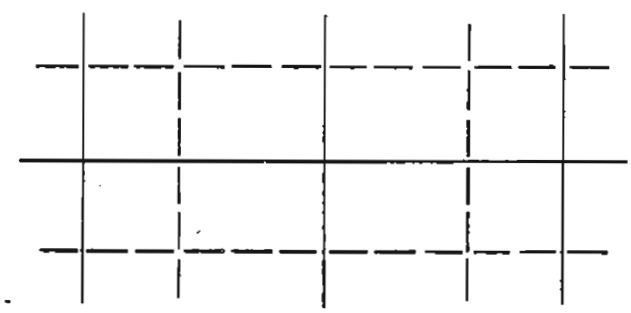


Fig. 5.2 - A single arterial with several cross streets.

For this geometry there are no loops involving the solid lines, and we can coordinate the signals along all arterials (nearly) independently. We start by coordinating the signals along the E-W arterial according to any appropriate scheme discussed in Chapters 3 or 4. In these schemes, the cycle time and the splits (including turn phases) were chosen with suitable allowance for the cross street flows.

The delays along each N-S arterial depend on the off-sets along that arterial but are independent of any choice of a "time origin" or the off-sets of other N-S arterials. Thus, on each N-S arterial, we can measure time relative to the setting of the signal at the intersection with the E-W arterial.

In discussing the coordination of signals on a single arterial in Chapters 3 and 4, we admitted the possibility that a cross street might itself be part of a coordinated system (particularly at the most critical intersection). Whether it was or not had little effect on a pretimed strategy for the arterial, because the total stochastic delay at or caused by the critical intersection is comparable with what would exist only at the critical intersection if it were an isolated intersection, for both the arterial and the cross street. Thus the choice of a cycle time (for the whole system) and the splits at the critical intersection are essentially independent of whether or not signals on the cross street are coordinated.

intersection are essentially independent of whether or not signals on the cross street are coordinated.

A coordination on the cross street did restrict somewhat the flexibility of a traffic responsive strategy because the through traffic phases in both directions should have a scheduled starting time. One does, however, still have the option of reassigning any time not used by one direction to reduce stochastic queueing in the other direction.

For any cross street other than at the critical intersection, the proposed coordination scheme along the E-W arterial would automatically assign any time not needed by the arterial (or most of it) to the (N-S) cross street,

if most of the E-W vehicles passing the critical intersection also pass the intersection in question. If, however, the intersections are so far apart that few E-W vehicles pass both intersections, the splits at the two intersections are determined independently. In either case, the stochastic queueing should be (considerably) less at the less critical intersection.

In the former (more common) case, one should indeed give most of the excess time not needed by the E-W traffic to the N-S direction at the less critical intersection because the stochastic queue in the N-S direction at this intersection is independent of that in the E-W direction and in the N-S direction at the critical intersection. Even if there is stochastic queueing in the E-W direction upstream of the critical intersection (in the direction of the heavier flow for a two-way arterial) giving more time to the E-W arterial will not change the total stochastic queueing in the E-W direction; it will only transfer some of the queue at this intersection to the critical intersection. Downstream of the critical intersection there should be little queueing on the E-W arterial because the E-W traffic is constrained by what can pass the critical intersection, so we obviously would not need to give the E-W arterial more time (contrary to what would result from most computer programs).

If, for some N-S arterial other than the one passing through the most programs).

If, for some N-S arterial other than the one passing through the most critical intersection in the E-W arterial, the intersection with the E-W arterial is the most critical intersection along the N-S arterial, the strategy described above would assign more time to the N-S flow on this arterial than one would have given it if the signals operated on the same cycle time C but the intersection with the E-W arterial is also the critical intersection on the E-W arterial. Although the coordination of the signals along the various N-S arterials are "independent of each other" in the sense that the off-sets are determined by the trip times along each arterial, independent of

the trip times on other N-S or E-W arterials, the "optimal" splits at the various intersections along a N-S arterial are highly correlated with the split at the most critical intersection along that arterial. But if this most critical intersection in the N-S direction is at a coordinated E-W arterial, the split at this intersection is also highly correlated with the split at the most critical intersection on the E-W arterial. Thus, the most critical intersection in the network (at least for this geometry), not only determines the common cycle time C for all signals, but it also affects the choice of splits at all intersections within a distance (in both directions) from the most critical intersection comparable with the mean trip length on the network. i.e., within an area around the critical intersection such that a significant fraction of the travelers in this area would have entered or left the coordinated network in this area or made at least on turning movement.

The actual delay in the network is not expected to be very sensitive to the choice of the cycle time or the splits, as long as the choices are reasonably close to the optimal. Indeed we have not even seriously proposed that one add the "deterministic" components of the delays at all intersections, or that one even try to calculate the total delay. The general procedure, however, suggested by this special geometry, would be to start from the most critical intersection in the whole network (or the most critical intersection in some suggested by this special geometry, would be to start from the most critical intersection in the whole network (or the most critical intersection in some area of the network). Choose the cycle time and splits at this intersection. Coordinate the signals along these two intersecting arterials giving any time not needed by these arterials to their cross streets, and "fan out" from there.

b. Highly imbalanced flows

Suppose one had a network which was (approximately) a rectangular grid with arbitrary spacings between parallel roads. All of the roads are two-way, but the flows on all parallel roads are imbalanced in the same direction. For example, if the roads run N-S and E-W, the northbound flow is larger than the

southbound flow on all N-S roads, and the eastbound flow is larger than the westbound flow on all E-W roads. Suppose, also, that the uninterrupted trip time between intersections is the same on all parallel roads.

If, in this (hypothetical) situation, one wished to minimize the total travel time for fixed flows, one would want to coordinate the signals on all E-W roads to give a progression in the eastbound direction and all N-S roads to give a progression in the northbound direction. There should be no conflict between the coordination on different roads. The loop conditions will be satisfied everywhere because the cumulative off-sets between any pair of intersections would be independent of the path.

Although this scheme of coordination will give the minimum total travel time for fixed values of the flow and is also "stable" in the sense that no driver can find a better route between his origin and destination than that which produced the given flows, this may not be the "optimal" design. If one could somehow convert a few of the N-S roads to a progression in the southbound direction and a few of the E-W roads to a progression in the westbound direction without disrupting too much the northbound or eastbound progression on other roads, the whole flow pattern would shift so that most of the trips are made along roads with a progression in the direction of the trip. In particular, most westbound or southbound trip components would focus on the roads are made along roads with a progression in the direction of the trip. In particular, most westbound or southbound trip components would focus on the roads provided for them, with considerable benefit to them.

We will not pursue further the details of how one would determine the cycle times, splits, etc., for a network with all parallel roads coordinated in the same direction. The purpose of this example was to illustrate the fact that any program designed to set traffic signals so as to minimize travel time for given flows on all links does not necessarily give the same pattern as one designed to minimize the travel time for given origins and destinations of trips. The latter problem is difficult to solve even for idealized geometries

(a rectangular grid) and idealized trip patterns. Even if one could evaluate the "optimal" plan, it would be quite impractical to implement because it would be very sensitive to the trip pattern. A slight change in the pattern can result in a whole new geometry of routing (typical of nonconvex optimization problems).

The goal in designing a signal system is not to minimize some hypothetical objective function. The purpose is to find some geometries and coordination schemes which are reasonably "efficient," can be easily implemented on an existing network, are soqially acceptable (equitable), can function both in the morning and evening peaks, etc.

c. Idealized two-way progression

If one had a square grid of roads and the trip time per block for the N-S and E-W directions were equal to each other and the same in all blocks, one could have a two-way progression on all roads with a cycle time C equal to twice the trip time per block. All roads would have an alternating coordintion as described in section 4.3(a). The signal scheme is shown schematically in figure 5.3. The solid lines represent the roads with the coordinated signals. The O's and X's represent intersections with opposite signal phases.

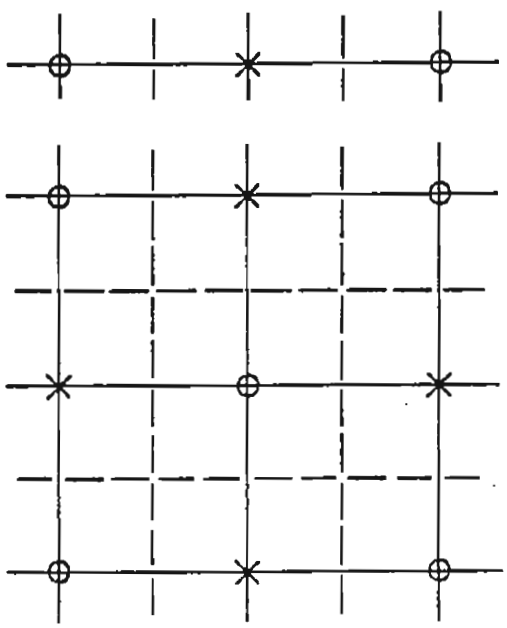


Fig. 5.3 - A grid of two-way streets.

The problem here, as for any two-way progression on even a single arterial, is that the cycle time C for such a scheme is unacceptably short for typical urban block lengths. There are, of course, other analogous schemes corresponding to the removal from figure 5.3 of any subset of the roads shown in figure 5.3. In particular, if one removed all odd numbered streets in the N-S and/or E-W directions, one would have a simultaneous coordination of the signals on the even numbered streets with a trip time between intersections of C (an even less likely situation).

One might imagine also that the network of figure 5.3 is only part of the complete network. For example, there might be other roads represented by the broken lines. Perhaps now, with coordinated signals only on the roads represented by the solid lines, it would be possible to choose a cycle time C equal to twice the trip time between intersections of the solid line network. The problem here is that at the intermediate intersections of the solid lines and broken lines, the platoons traveling in the east and west directions (or the north and south directions) would likely not overlap. The time available (if any) for the through traffic on the secondary network at these intersections (signalized or not) would be limited to what was not used by either of the two primary directions. (There would be ample time for right turning traffic, however.)

the two primary directions. (There would be ample time for right turning traffic, however.)

Actually the design of a network (or part of a network) for a two-way progression on all or some roads is not nearly as restrictive as suggested by the above idealized example. Indeed it is probably fairly common practice for traffic engineers to try to create a two-way progression on some streets, at least for light traffic, i.e., to create two directional through bands on some or all streets of a network.

If, as in example b, one were first to create a progression system for the two directions of heavier flow, for example, northbound and the eastbound

directions, with equal travel times on parallel roads of the coordinated network, this will automatically satisfy the loop conditions on all loops of the coordinated network. The splits between northbound and eastbound directions need not be equal to each other. They need not necessarily even be the same at all intersections. It is only necessary that each signal have a uniquely defined timing relative to some common reference.

For any given flows, turning movements, etc., at various intersections, it would be advantageous first to locate the most critical intersection and specify the relative splits there between the northbound and eastbound directions, but leave the cycle time as flexible as possible. From this intersection one can construct minimum width bands needed to accommodate the northbound and eastbound flows at all intersections. If the flows at the critical intersection (and all others) are well below saturation (degree of saturation less than 1/2, for example), there will be considerable flexibility in the relative timing of the northbound and eastbound through bands at the critical intersection and elsewhere.

Having established a family of possible progression schemes for the northbound and eastbound directions, the problem now is to see if any of them will also accommodate some through bands for the southbound and westbound directions. Of course, this was difficult to do even for a single two-way arterial as dis-

also accommodate some through bands for the southbound and westbound directions. Of course, this was difficult to do even for a single two-way arterial as dis-

cussed in section 4.4. One is not likely to succeed (except possibly for very light flows) unless the intersections are nearly equally spaced and the cycle time can be chosen as two or four times the trip time between intersections (but we presume here that the former is unacceptable). For this to be true in both the N-S and E-W direction, it is (almost) necessary that the network be a square grid of roads or possibly a rectangular grid with block lengths in one direction some integer multiple of those in the other direction.

For the square grid illustrated in figure 5.3, it would be possible on the combined network of the solid lines and broken lines to have a "double alternate" type scheme as described in section 4.5b in both the N-S and E-W directions, but not necessarily with equal bandwidths in any of the four directions. In such a scheme, however, the bands (as illustrated in figure 4.4b) would not overlap in the north and south directions, nor in the east and west directions at intersections. For the traffic to be accommodated within the band, the traffic must be sufficiently light that it can be served one through direction at a time with no overlap (for the specified value of C), The turning traffic would typically cause no problem, but it is not obvious that this would be a desirable scheme for higher flows than can be accommodated in the bands.

Alternatively, one could construct a two-way progression only on the solid lines of figure 5.3 with potentially a full utilization of the capacity at the intersection of the solid lines for the two cross directions with the heavier flows. The through bands, analogous to those illustrated in figure 4.4c, however, would not overlap as they pass the intersection of the solid line with the broken lines of figure 5.3. Through traffic on the E-W broken lines could pass only in the time (if any) between the passage of the N and S bands in two segments per cycle; similarly for the traffic in the N-S broken lines could pass only in the time (if any) between the passage of the N and S bands in two segments per cycle; similarly for the traffic in the N-S broken lines (One might need a median strip so that pedestrians can cross these intersections in two stages.)

The problem with this scheme is that if the signals at the intersection of the solid and broken lines are set so as to pass the nonoverlapping bands on the solid lines network, any increase in the flow on the solid lines will be made at the expense of a loss in capacity on the broken line network for through traffic. In some cases, the capacity would be zero. This might be appropriate if the road network were designed so that the solid lines carried

the through traffic, but the broken lines were used only for access to parking lots (with no through traffic and no left turns during rush hours).

It is not necessary that one use the same band coordination on all N-S or E-W roads. At the most critical intersection where the N-S street with the heaviest flow (actually the value of q/s) meets the E-W street with the heaviest flow, one might need to have the northbound and southbound bands overlap if they are to accommodate the flows; also the eastbound and westbound bands. On these two streets, there may be some problems, as described above, at the intersection with the broken line network where the opposing bands do not overlap. For other N-S or E-W streets with lighter flows, however, one has more options for shifting the relative positions of the (narrower) N-S and E-W bands. The positions of the bands are actually determined by the setting of the signals at the intersections of the broken and solid lines of figure 5.3.

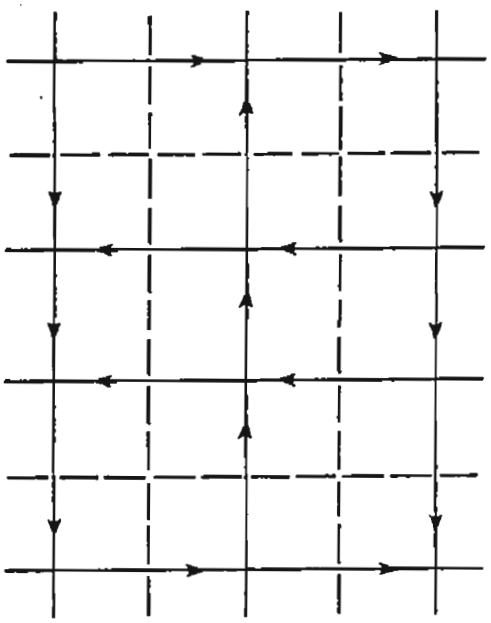


Fig. 5.4 - A grid of two-way streets.

d. Idealized one-way progression

Suppose, as illustrated in figure 5.4, the road network is a square grid, and one wishes to coordinate the signals on the roads represented by the solid lines for a progression only in the directions indicated by the arrows. Successive E-W roads on the solid lines alternate between progression for the eastbound and westbound directions, and successive N-S roads alternate between progression for the northbound and southbound directions. These roads could be either one-way or two-way roads, but, if the latter, we are not concerned with the behavior of traffic traveling against the progression. There may also be other roads in the network as indicated by the broken lines (not necessarily regularly spaced) which are probably two-way roads. The signals (if any) at the intersections of the broken and solid lines are coordinated for the progression on the solid lines, which typically will not provide a progression in either direction along the broken lines.

Except for the arrows, the geometry of figure 5.4 is the same as figure 5.3. One can readily verify, however, from the geometry of figure 5.4, that the loop conditions will be satisfied if the progression speeds are equal on all parallel roads (but in opposite directions on adjacent roads) and the total trip time around any loop on the solid lines is an integer multiple of the cycle time, typically equal to C for the smallest loop. These conditions will also be satisfied for any modification of the network shown in figure 5.4 the cycle time, typically equal to C for the smallest loop. These conditions will also be satisfied for any modification of the network shown in figure 5.4 in which any of the roads shown are deleted.

With the broken line network deleted, this scheme would give a trip time per block of $C/4$ ("quarter cycle off-sets"). In contrast with the two-way progression scheme in (c) with half-cycle off-sets, the resulting value of C is now likely to be in an acceptable range for typical urban networks (perhaps 40 to 60 seconds for the one-way grid as compared with 20 or 30 seconds for the two-way grid). For cities with very short blocks or for networks which

may require a long cycle time (because of pedestrian or whatever), one still has the option of coordinating the signals on only part of the network as illustrated in figure 5.4 including the broken lines.

In contrast with the two-way progression scheme in (c), the existence of the broken line network in figure 5.4 does not cause any complication in the one-way progression system. Any signal at the intersection of the solid and broken lines is constrained to maintain the progression on the solid lines only in one direction. The time needed per cycle to accommodate the traffic in the progression direction at this intersection should not typically be any larger than is needed at the adjacent intersections of the solid lines. We would also expect that, if the solid lines are two-way streets, the traffic going against the progression would be relatively light and would easily be accommodated in the time allotted to the progression direction. The through (or turning) traffic on the broken lines is also expected to be lighter than on the parallel solid lines. Indeed it might be so light that there would be some time not needed by either the progression direction or the cross street, which could be reallocated back to the solid lines at such a time as to assist some vehicles traveling against the progression.

If a traffic engineer or city planner has some freedom to select which streets of a network shall be one-way streets, or even two-way streets designed to carry through traffic, it would clearly be desirable to create a geometry streets of a network shall be one-way streets, or even two-way streets designed to carry through traffic, it would clearly be desirable to create a geometry such that these streets are part of a solid line network of the type shown in figure 5.4 with a cycle time chosen to satisfy the loop conditions. If he could relocate parking lots or other entrances or exits to the network, he might also restrict these to the broken lines (if any). He could also use the signals at the intersections with the solid lines to "meter" the flow leaving the parking lots. He could also ban curb parking on the solid line network.

There is some flexibility in the design of such a system. The cycle time for the whole system is, of course, determined by the loop conditions and the spacing between the coordinated signals, but one can vary the splits.

It is advantageous to start from the most critical intersection where the E-W street requiring the most time meets the N-S street requiring the most time. If this intersection is undersaturated, one could choose the split here between the N or S and E or W traffic as one would if this were an isolated intersection serving the prevailing flows. This would determine the through bands for the two cross directions at this intersection which are to have a progression, say the northbound and eastbound directions.

If there is negligible turning traffic, one could propose a preliminary plan in which every intersection has the same split between the N-S and E-W directions as at the critical intersection. All northbound and southbound bands would have the same width as would all eastbound and westbound bands. Such a plan would satisfy all the loop conditions even though the bandwidths are different in the N-S and E-W directions. If the critical intersection can accommodate the flows, then so can any other intersection. Indeed some intersections may be highly underutilized. If the critical intersection is at the junction of northbound and eastbound bands, and the traffic is predominantly in the northbound and eastbound directions throughout the network, the southbound and westbound bands particularly would be underutilized and the signals in the northbound and eastbound directions throughout the network, the southbound and westbound bands particularly would be underutilized and the signals at the junctions of the latter bands would have considerable excess time.

In the absence of turning vehicles, there would be no reason to modify this pretimed plan despite its apparent underutilization. Stochastic delays have already been taken into account in the choice of the split at the critical intersection, because, it is reasonable to assume that the total stochastic delay along the two streets crossing at the critical intersection is determined by the split at the critical intersection. There would be no reason to make

the bands through the critical intersection wider either upstream or downstream of this intersection in either the northbound or eastbound directions, even at the (underutilized) intersections with the southbound or westbound roads. Any time not needed by the through bands is given to the cross street, thus minimizing the stochastic delay on the cross street. The stochastic delays at any other intersection should be relatively unimportant.

If there is some turning traffic, but only a small fraction of turning vehicles at any intersection, it may be advantageous to vary the width of a through band along its length as described for a single arterial in Chapters 3 and 4. One would do this first along the two streets which meet at the critical intersection. Along the eastbound street, for example, if one wishes to widen the eastbound band at some intersection, one would advance the start of the band but keep the end of the band in time with the progression. This would be done at the expense of the northbound (or southbound) band at this intersection, displacing the rear of this band. The signals at all other intersections along this northbound band should now be displaced by the same amount so as to maintain the progression for the rear of the northbound platoon.

If one does the same thing along the northbound street passing through the critical intersection, this will displace the rear of some eastbound (or westbound) bands crossing this northbound street. This in turn may change the starting time of some northbound (or southbound) bands at streets crossing

bound) bands crossing this northbound street. This in turn may change the starting time of some northbound (or southbound) bands at streets crossing

the eastbound streets, including some intersections where one may have already displaced the rear of the northbound bands in the previous stage.

To see how one can manipulate the signals while maintaining a progression in four directions, it may be advantageous to construct a three-dimensional time-space diagram. From some slotted sheets of clear plastic, one can make an assembly like the corrugated cardboard in a carton used for shipping bottles. On each sheet one can draw the space time trajectories

of vehicles or through bands on a single street with the time measured along the axis of the slots and the intersections at the slots. In the assembled array, the signals displayed along the slots of the sheets must be compatible with each other at the intersection of the sheets.

To make an efficient traffic responsive system for this geometry, one should start from a pretimed plan such as described above, preferably one in which the two through bands which meet at the critical intersection(s) are a little narrower at the critical intersections than elsewhere. It would typically suffice to make only the signals at the critical intersections traffic responsive.

A signal phase at a critical intersection should terminate if a platoon in either direction passes the intersection before the scheduled termination time, so as to guarantee that the intersection is kept busy whenever there is a queue in either direction. Since the cycle time is prespecified, however, one may need to make some modification of this strategy during the off-peak when there might be no queue in either direction. One could have a preset earliest time to terminate a phase and/or a provision that a phase will not terminate unless there has been a call for another phase.

e. Mixed strategies

We saw in part (b) that it was possible to have a progression in the same

e. Mixed strategies

We saw in part (b) that it was possible to have a progression in the same direction for all parallel streets on a rectangular grid of arbitrary spacing and in part (d) that it might be possible to have a progression for half of the N-S (E-W) streets in one direction and half in the opposite direction on a square grid. It would be nice if one could provide a progression for some arbitrary fraction of the N-S streets for the northbound direction and some arbitrary fraction of the E-W streets for the eastbound direction, reversible

for the morning and evening peaks, but, unfortunately, this is impossible. One can do it for either the N-S directions or the E-W directions but not both.

Some cities have roads with reversible lanes and perhaps even reversible one-way streets, but this is quite rare because it is very awkward to implement. In most cases, the total amounts of pavement available for opposing traffic directions are equal and constant throughout the day, and one-way streets retain the same direction at all times. Street networks are typically a mixture of one-way and two-way streets, so about the only flexibility one has in adjusting to the morning and evening peaks is to reverse the direction of progression (if possible) on some of the two-way streets. One can change the splits at intersections between the N-S and E-W directions, but if the morning split is primarily between the northbound and eastbound directions, this is probably not much different from the desired evening split between the southbound and westbound directions.

Presumably the "capacity" of a network for northbound traffic, for example, would be independent of the signal off-sets. The capacity of any single N-S road for northbound traffic would be determined by the signal split (and cycle time) at the critical intersection. Chances are that the critical intersection is the one carrying the heaviest E-W traffic, and that the same E-W street is the critical intersection for many N-S streets. Also the same N-S street is is the one carrying the heaviest E-W traffic, and that the same E-W street is the critical intersection for many N-S streets. Also the same N-S street is likely to be the critical intersection for many E-W streets.

If the trip time (including delays) along all northbound streets were equal, then, ideally, the traffic should distribute itself so as to utilize the available capacity on all streets and to maintain the equality of the trip times. If, however, some streets have a good progression for the northbound direction and some do not, the streets with the good progression will attract more traffic and reach capacity first as the flow increases. Theoretically,

if the flow increases further, queues will form on the "good roads" and travelers will not divert to the other roads until the good roads are so congested that the trip time on these roads is the same as on some other roads. In this situation, the travel time in the network depends on the trip time on the latter roads and is nearly independent of the progression scheme on the "good roads."

The conclusion from this is that it would be desirable to have a network of roads with nearly equal trip time on many parallel roads so that the traffic would be distributed. It would obviously also be desirable that the capacity of these roads can accommodate the demand.

There is another potential complication. In an attempt to provide a progression on a few roads for vehicles traveling against the predominant direction of flow, one might choose to interrupt the progression on some roads at just a few locations. This may result in giving some signal almost the opposite phase from what is needed to maintain the progression in a direction of heavy flow. The danger here is that if a platoon is timed at arrive at some intersection just as the signal turns red, a queue will form during the red interval. If the red interval is long enough, the queue could back up past the upstream intersection. In the subsequent green interval, the signal may be able to serve only part of the queue because the upstream signal may turn red at such a time as to cut off the flow. This "blocking" (grid lock) will cause a loss in capacity at this intersection, which, in turn, may become a bottleneck restricting the capacity of the whole road.

Not all roads are intended primarily to carry through traffic. Some roads may be access roads to residential areas, parking lots, etc. Not only are they not intended for through traffic, one does not even want through traffic to use them, even if the roads intended for the through traffic are congested.

If the primary purpose of a street network were to carry through traffic, then, for the typical spacing of urban roads on a square grid, the network with

the highest capacity and least travel time would be a network of all one-way streets in a pattern of the solid lines of figure 5.4. This would be true even if the traffic is highly imbalanced, but in opposite directions for the morning and evening peak. In such a case, the logical competitor for such a system would be that described in part (b) where all signals on N-S roads were set for progression in the northbound direction in the morning, for example, and in the southbound direction in the evening; similarly for the E-W roads. These roads, however, would be two-way roads. If it were not possible to have reversible lanes, one would need just as many southbound lanes as northbound lanes. Half of these lanes would be underutilized to the same extent as the one-way roads of figure 5.4 in the direction of the lighter traffic. The one-way roads, however,would give a progression for both directions and fewer conflicting traffic movements at intersections.

Obviously, the capacity for through traffic is not a primary issue for most road networks. Two-way roads provide better access to facilities along the road, and the utilization of any road involves conflicting objectives. People involved in activities along a road would prefer that through traffic go elsewhere. Unfortunately, the traffic engineer cannot usually decide the use of the roads, and which roads are used for through traffic is often linked with political and economic issues. If only part of the road network is to be designed for through traffic, and one has some control over which roads they with political and economic issues. If only part of the road network is to be designed for through traffic, and one has some control over which roads they are, there are some schemes which will provide a progression on some of the roads without limiting the capacity of the others.

If the traffic is much heavier in the N-S direction than the E-W direction, possibly also imbalanced in the N-S direction reversing directions for the morning and evening peaks, the pattern described in part (a) permits an arbitrary coordination on the N-S roads and on a single E-W road for any choice of cycle time and split at the critical intersection. In particular

one can have an arbitrary fraction of the N-S roads have a progression for the northbound direction.

If there were other E-W roads at a trip time of C (or an integer multiple of C) north or south of the one shown in figure 5.2, they could also be coordinated in the same pattern as the original E-W road. One even has some flexibility in the choice of C to make this happen. Unfortunately any E-W roads between the coordinated E-W streets would likely have a rather chaotic signal pattern. If the signals on roads with a progression are set for the eastbound traffic, there would not likely be any roads with a progression for the westbound direction unless one can create a geometry similar to figure 5.4.

In part (c) we considered a possible two-way progression as illustrated in figure 5.3. If the trip time between intersections on the solid lines is C/2, but there are also roads (broken lines) at a spacing corresponding to a trip time of C/4 , one could set the signals at the intersection of the solid and broken lines so as to give a progression in only one direction on the solid lines, in any combination, and reversible for the morning and evening peaks. In the off-peak, one might be able to set these signals to provide a two-way progression, but, when the traffic exceeds the capacity of the through bands, switch to a pattern with just a one-way progression on each road. The signals on the broken line network would, of course, typically have a rather chaotic pattern at all times unless the progression direction alternates so as to give

on the broken line network would, of course, typically have a rather chaotic pattern at all times unless the progression direction alternates so as to give the pattern of figure 5.4

5.3. Comments

Most road networks are not square grids and, because of the loop conditions, one cannot coordinate the signals on all roads to provide a good progression on every road where one might like it. There is no simple procedure for choosing an "optimal" coordination of signals on a network and most computer programs which pretend to do so may actually do the "wrong" thing.

Generally, one is not trying to design a new network, but simply to make improvements on an existing system. Because of land use, political considerations, etc., one does not have many options for changing the function of the existing roads. One should recognize, however, that a change in the signal coordination may induce a change in the flow pattern over the network.

We do not have very accurate models of traffic assignment, so any strategy of signal coordination must be, in part, a multistage "trial and error" procedure. If one reduces the travel time on some road, it will likely attract more trips to that road and reduce the flow on neighboring roads. This should typically yield a net benefit to any drivers who change routes (otherwise they would not have shifted routes) and probably also to the drivers who did not shift routes, because of the reduced congestion. It is difficult, however, to quantify these benefits because not all drivers perceive the benefits equally and any change in the pattern has disbenefits of some sort. In the final analysis, any change must be consistent with what is acceptable to society.

Certainly the first step in any analysis of a signal system is to identify the most critical intersections in the existing systems. If some intersection is oversaturated, changing the coordination of the signals upstream (in the oversaturated flow directions for two-way streets) of the intersection on the two intersecting streets is not likely to have much effect on the trip time along these two streets, because the trip time is determined mostly by the de-

the two intersecting streets is not likely to have much effect on the trip time along these two streets, because the trip time is determined mostly by the de-

parture process from the critical intersection. One must either provide some additional capacity and more attractive routes somewhere so as to divert the traffic to other roads, or just accept the situation as it is. One does have the option of changing the splits at the critical intersection and thereby shifting the delays between different classes of drivers, but there is no clear choice as to who should suffer the most. The issue here is the same as for an isolated intersection as discussed in Chapter 2.

When the critical intersection is undersaturated, chances are that one provided a progression for the two heavier flow directions on the two streets which intersect there (or at least gave them favorable consideration if there were some loop constraints). But if this intersection becomes oversaturated, the progression is ineffective upstream of this intersection for the two directions of heavy flow. One still should set the signals on these streets so as to prevent blocking in these directions, but otherwise one has the freedom to reset the signals to provide some benefit to any other traffic movements.

If these are two-way streets, one might, for example, set the coordination to favor the traffic in the opposing directions (upstream of the critical intersection). But if there were loop constraints which affected the coordination of signals on cross streets and on other adjacent streets, relaxing the progression scheme on part of the two streets which meet at the critical intersection may allow one to provide a better progression on some other streets (particularly streets which could provide an alternative route for drivers who were going through the critical intersection).

Even if the critical intersections are always undersaturated, the stochastic queueing in the network is dominated by that at or caused by the critical intersections. If some traffic going through the critical intersection could be diverted so as to go through less critical intersections, this would typically give a net benefit. In the absence of loop constraints, however, it would generally be desirable to provide the best possible coordination on each street individually. Even though it might be advantageous to society if some drivers would use less critical intersections, it would be difficult to induce drivers to choose routes which are less desirable to them individually. If, however, because of the loop conditions, one must make some compromise between the coordination on one street or another, then one should try to do it in such a way that the traffic will distribute itself so as to utilize all

Although there is no clear recipe for implementing these principles, a bit of common sense will go a long way. In contrast with this, most computer schemes are not likely to do what should be done. Aside from the fact that the computer programs contain a grossly exaggerated estimate of the stochastic delay and an even worse estimate of delay for oversaturated conditions, the scheme of minimizing delay for fixed flows tends to give preference to traffic movements with the heavier flows. This, in turn, is likely to make streets with heavy flow even more advantageous so as to attract more flow, whereas one would like to divert flow to streets which are underutilized.

INDEX OF NOTATION

Symbol	Meaning	Page
G_{mi}	minimum green interval for direction i	43
G_M	maximum green interval	97
G_{Mi}	maximum green interval for direcion i	118
G^*_{Mi}	maximum green intervals for a modified strategy	130
G^*_{M3}	a maximum green for a six-phase signal	202
h	headway	141
h_0	minimum acceptable headway	142
h_1	headways between vehicles crossing in a gap	142
H_j	a headway for the jth vehicle	141
$H(\mu)$	a tabulated function	54
i	index numbering flow directions	39
I	$I_A + I_D$	52
I_A	variance to mean ratio for arrivals	52
I_D	variance to mean ratio for departures	19
I_1, I_2	the value of I for directions 1, 2	57
$I_i(t)$	time-dependent value of I_i	65
j	index numbering vehicles	2
j	integer index	141
k	spatial density of vehicles	29
k	index numbering cycles	16
k	spatial density of vehicles	29
k	index numbering cycles	16
k_j	jam density of stopped vehicles	31
ℓ	index numbering channels	4
ℓ	integer index	231
$\ell_j(t)$	channel of the jth vehicle at time t	4
L	effective lost time per cycle for a F-C signal	39
L_i	lost time in switching from direction i to the cross direction	111, 162

Symbol	Meaning	Page
L^*	effective lost time per cycle for a V-A signal	113
L_3	lost time per cycle for a 3-phase signal	184
L_4	lost time per cycle for a 4-phase signal	181
L_5	lost time per cycle for a 5-phase signal	196
L_6	lost time per cycle for a 6-phase signal	187
$L_{1\ell}$, $L_{2\ell}$	lost time from switching for a turning vehicle	162
L'_t	lost times per cycle with turn bays in only directions 1, 3	170
$L'_{t\ell}$	lost time per cycle if signal switches for left turn vehicles in directions 2 or 4	170
$L^{(m)}$	lost time per cycle at intersection m	226
L_6^*	lost time per cycle with short turn bays	210
m	integer	50
m	intersection number	225
m*	critical intersection	234
m'	an intersection upstream of m	234
m''	an intersection upstream of m'	235
m''	an intersection downstream of m*	279
m_1, m_2, m_3	three intersections	253
m_1	an intersection upstream in direction 3 of the critical intersection	377
$\bar{n}_1$, $\bar{n}_2$, $\bar{n}_3$	----------	
m_1	an intersection upstream in direction 3 of the critical intersection	377
m(h)	the number of vehicles served in a headway h	141
M	number of vehicles to arrive in a cycle	116
n	number of vehicles	19
n(t, x)	cumulative number of vehicles to pass x by time t	9
$\bar{n}(t, x)$	average of n(t, x)	16
$\bar{n}_i(t)$	average cumulative number of vehicles to arrive in direction i	68

Symbol	Meaning	Page
$n_k(t, x)$	cumulative number of arrials at x in the kth cycle by time t	16
$n_{\ell i}$	storage capacity of left turn way for direction i	176
n_1^*	a critical number of departures	218
$\bar{n}_i^{(m^*)}(t)$	average cumulative arrivals by time t at intersection m* in direction i	281
N	number of cycles observed	16
p	$p_i = p$ for symmetric intersection	155
p_i	fraction of left turning vehicles in direction i	151
p_i'	fraction of right turning vehicles in direction i	151
p*	an average of p_1, p_2	156
$p^{(m^*)}$	fraction of vehicles which pass m* which entered at m* - 1	275
p(w)	price of a unit of delay for someone who has already waited a time w	75
P	total cost of an interruption of traffic	134
q	flow	27
q(t, x)	flow at time t, location x	17
q_i	approach flow for direction i	39
q_i	flow for direction i averaged over several cycles	225
$q_i(t)$	flow in direction i at time t	65
$q_{\ell i}$	flow for left turns from direction i	185
$q_i(t)$	flow in direction i at time t	65
$q_{\ell i}$	flow for left turns from direction i	185
q_{ti}	flow for through traffic in direction i	186
$q_i^{(m)}$	flow in direction i at intersection m	225
$q_{\ell i}^{(m)}$	flow for left turns from direction i at intersection m	226
$q_{ti}^{(m)}$	flow for through traffic in direction i at intersection m	226
q_2^*	a flux of cross street traffic	298

Symbol	Meaning	Page
Q	average residual queue at the start of red	49
$Q(t)$	queue length at time t	30
Q_i	residual queue for direction i	54
$Q^{(m)}$	residual queue caused by the mth intersection	275
R	effective red interval	134
s	saturation flow	17
s_i	saturation flow for direction i	39
$s_{\ell i}$	saturation flow for left turns from direction i	186
s_{ti}	saturation flow for through traffic in direction i	186
$s_i(t)$	a time-dependent saturation flow in direction i	214
$s_i^{(m)}$	saturation flow for direction i at intersection m	225
s^*	flow per direction at a stop sign	138
s^*	an average of s_1, s_2	156
s_i^*	a decreased saturation flow in direction i	218
t	time	2
t_j	time jth vehicle passes intersection	13
$t_j(x)$	time jth vehicle passes location x	9
$t(x)$	time measured relative to an uninterrupted vehicle	231
t_y	time a signal switches from green to yellow	82
t_r	time a signal switches from yellow to red	85
t_y	time a signal switches from green to yellow	82
t_r	time a signal switches from yellow to red	85
t_r'	an early time to switch from yellow to red	92
t_0	time last vehicle passes a detector	91
t_0	time to start a red interval	237
t_0	time a vehicle leaves intersection 1	311
t_0'	time a vehicle reaches intersection m*	311
t_1, $t_2 \ldots t_6$	times in figure 3.15	312
t_1', $t_2' \ldots$	times in figure 3.15	312

Symbol	Meaning	Page
t_1^*	time in figure 3.15	312
t^*	time a last platoon vehicle passes an intersection	86
t^*	time of a drop in flow	239
$\bar{t}_i(n_i)$	time to serve n_i vehicles in direction i	214
T	a travel time	289
u	average number of turning vehicles served per cycle	185
u	a dimensionless progression speed	368
v	velocity	82
v_j	time average velocity of jth vehicle	29
$\bar{v}$	average velocity of all vehicles	29
$v_j(t)$	velocity of jth vehicle at time t	6
v_p	velocity of a platoon vehicle	82
v^*	design speed for progression	285
v_m	speed which gives maximum flow	285
v'	a speed for which the green interval is fully utilized	290
V-A	vehicle-actuated	
Var(C)	Variance of C	108
w	waiting time	75
w_1, w_2	waiting time of vehicles at the head of the queue for directions 1, 2,	75
w_1, w_2	waiting time of vehicles at the head of the queue for directions 1, 2,	75
w_0	a waiting time for a semi-actuated signal	135
W(t)	waiting time at time t	29
$\bar{W}$	average wait per vehicle	34
$\bar{W}_i$	average wait per vehicles in direction i	44
$W_j^{(m)}$	wait for a jth vehicle passing the mth intersection	264
$\bar{W}^{(m)}$	average of the $W_j^{(m)}$ over all vehicles	265
x	a coordinate along a channel	2

Symbol	Meaning	Page
$\|x\|$	absolute value of x	29
x'	location of an observer	11
$x_j(t)$	position of a jth vehicle at time t	2
y	a coordinate in two-dimensions	2
y	a location along a channel	25
$y_j(t)$	y-coordinate of a jth vehicle at time t	2
y_1, y_2	locations of signals	7
α	the delay equivalent of one stop	46
α'	the delay equivalent of a stop and go in queue	62
β	time needed to detect the end of a platoon	88
β_i	time to detect the end of a platoon in direction 1	111
β'	minimum time for a loop detector to be unoccupied	100
δ	deviation of the off-set per unit distance	289
$\delta^{(m)}$	off-set of signal at m relative to $m-1$	229
$\Delta_j^{(m)}$	delay to a jth vehicle at intersection m	265
$\Delta^{(m)}$	the average delay to all vehicles at intersection m	265
ε	a small time interval	70, 74, 238
η	a cycle split	154
μ	a parameter	54
ρ	degree of saturation	50
μ	a parameter	54
ρ	degree of saturation	50
$\rho(t)$	degree of saturation at time t	51
ρ_i	degree of saturation for direction i	126
$\rho_{\ell i}$	degree of saturation for left turning vehicles in direction i	175
τ	time a queue vanishes	33, 237
τ	trip time between intersections	227
$\tau^{(m)}$	trip time from intersection $m-1$ to m	229
$\tau'^{(m)}$	trip time from intersection m to $m-1$	348