

CHAPTER 2

STATE-OF-THE-ART IN

DELAY, QUEUE AND STOP ESTIMATION

2.1 OVERVIEW

The delays experienced on the arterial signalized streets are mainly associated with the intersection where conflicting movements are separated and controlled by traffic signals. These traffic signals can operate under an isolated control strategy, with the signal settings of each signal set independently of the settings of adjacent signals. The delay is defined as the difference in travel time when a vehicle is unaffected by the controlled intersection and when a vehicle is affected by the controlled intersection. This delay includes lost time due to deceleration and acceleration as well as stopped delay. Thus, intersection delay estimates are directed toward estimating total delay or simply stopped delay.

The research presented in this dissertation deals with the concept of vehicle delay at signalized intersections. More specifically, the main objectives of the research are to review the methods currently available for estimating the delay incurred by motorists at signalized intersections, to identify the main limitations of each method, and to propose methods for improving the accuracy of delay estimates.

To fulfill these objectives, the chapter provides some background material related to the concept of delay at signalized intersections. In the first section, the chapter describes the various types of traffic flows commonly encountered in a transportation network. The chapter then presents a detailed description of the various models that are currently available for estimating delay at signalized intersections based on vertical queuing models such as the Highway Capacity Manual (1994 and 1998 versions) and the Canadian Capacity Guide. This presentation is then followed by a discussion of a series of fundamental traffic flow characteristics and that are used in the

development of delay estimation models that consider the spatial extent of queues. Next, an overview of shock wave analysis for horizontal queuing models is presented. Finally, a detailed description of the INTEGRATION microscopic traffic simulation model that will be used in the analysis of the delay estimation model is provided.

2.2 CATEGORIES OF TRAFFIC FLOWS

Vehicle flow on transportation facilities may generally be classified into two categories:

- Uninterrupted flow, and
- Interrupted flow.

Uninterrupted flow occurs on facilities on which there are no external factors causing periodic interruptions to the traffic stream. As indicated in Table 2.1, such flows exist primarily on freeway and other limited-access facilities, where there are no traffic signals, stop or yield signs, or at-grade intersections to interrupt the continuous movement of vehicles. Such flows can also occur on long sections of surface highways between signalized intersections where the geometric and driving characteristics approach those usually found on a limited-access facility. On uninterrupted flow facilities, traffic flow conditions are thus primarily the result of the interactions among the vehicles within the stream and of the interactions between the vehicles and the geometric characteristics of the roadway. If congestion occurs, the breakdown of traffic flow then is strictly the results of frictions internal to the traffic stream and not the result of external causes.

Interrupted flow occurs on transportation facilities that have fixed elements causing period interruptions to the traffic stream irrespective of existing traffic conditions. As indicated in Table 2.1, these flows occur on facilities on which traffic signals, stop signs, yield signs, and other types of control devices force motorists to interrupt their progression at specific locations. On these facilities, traffic flow characteristics thus not only depend on the interactions between

the vehicles within the stream and with the roadway geometry, but also on the external factors causing the interruptions.

Table 2.1: Types of Transportation Facilities

Uninterrupted Flow	Freeways
	Multilane highways
	Two-lane highways
Interrupted Flow	Signalized Streets
	Unsignalized Streets with Stop Signs
	Arterials
	Transits
	Pedestrian Walkways
	Bicycle Paths

The principal device creating periodical interruptions to traffic flow in transportation networks is the traffic signal. Because these signals allow designed movements to occur only part of the time, they cause traffic to flow in platoons. A platoon is usually defined as a group of vehicles traveling closely together along a facility. These platoons are formed during the red interval of a traffic signal, when vehicles are being queued upstream of the stop line, and are released at the beginning of a green interval. As platoons depart a signalized intersection, they tend to disperse. This dispersion is a function of the spacing between signalized intersections, driver behavior and traffic conditions along the roadway. When successive intersections are far enough apart, the extent of the dispersion may even become sufficient to assume that uninterrupted flow exists on some part of the roadway between them. As a general guide, a spacing of 3.2 kilometers between intersections is often thought as being sufficient for considering uninterrupted flow to exist at some points between signalized intersections (McShane and Roess, 1990).

2.3 DELAY ESTIMATES USING VERTICAL QUEUING ANALYSIS

In transportation systems, queues are formed whenever the number of arrivals at a given location exceeds the maximum rate at which vehicles can go through the location. When such a situation occurs, the excess vehicles are stored upstream of the bottleneck or service area and their departure is delayed to a later time period. Depending on the type of service provided, the queues that are formed may be either moving or completely stopped. Typically, moving queues are formed at locations where the flow of vehicles across the bottleneck or service area is never completely stopped. Stopped queues occur on the other hand when there are completely interruption of service for a significant amount of time.

Examples of queuing processes in highway systems include stop-controlled and signalized intersections, toll plazas, parking facilities, freeway bottlenecks, incident sites, merge areas near freeway on-ramps, and traffic disruptions caused by slow moving vehicles. In each case, queuing theory can be used to analyze the formation and dissipation of queues and determine the amount of delay incurred by motorists while waiting in or moving through a queue. Typically, two types of analyses can be done: a deterministic analysis and a stochastic analysis. In a deterministic analysis, the vehicle arrivals are assumed to follow a uniform pattern. In a stochastic analysis, it is instead assumed that the arrivals are distributed according to some probability distribution. To simplify the analysis, it is also assumed in both approaches that the vehicles queue vertically, i.e., occupy no space while queued, and that the vehicles accelerate and decelerate instantaneously. Both of these two approaches are discussed in more detail in the following sections.

2.3.1 DETERMINISTIC QUEUING ANALYSIS

Deterministic queuing analysis can be undertaken at two different levels. First, the analysis can be carried out at the macroscopic level, where the arrival and service patterns of vehicles are considered to be continuous. The analysis can also be carried out at the microscopic level, where both the arrival and service patterns are considered to be discrete. Typically, a macroscopic analysis is conducted when the arrival and service rates are high, while a microscopic analysis is often conducted when the arrival and service rates are low.

To illustrate how deterministic queuing analysis can be employed to determine the delay incurred by motorists at a bottleneck, the diagram of Figure 2.1 illustrates the evolution of a queue of vehicles at an undersaturated signalized intersection where vehicles arrive at a uniform rate. The diagram first indicates that vehicles join the stop line queue at a rate corresponding to the vehicle arrival rate during the red interval and that the maximum queue size occurs just before the signal turns to green. In such case, the maximum queue size would therefore correspond to the number of vehicles that have reached the intersection during the red interval. Following the display of the green signal, the diagram indicates that the queue dissipates at a rate that corresponds to the difference between the rate at which vehicle leave the queue at the front and the rate at which vehicles join the queue at the back. If the vehicles are assumed to enter the intersection at the saturation flow rate, the time required to dissipate a queue in Figure 2.1 would then be given by Equation 2.1.

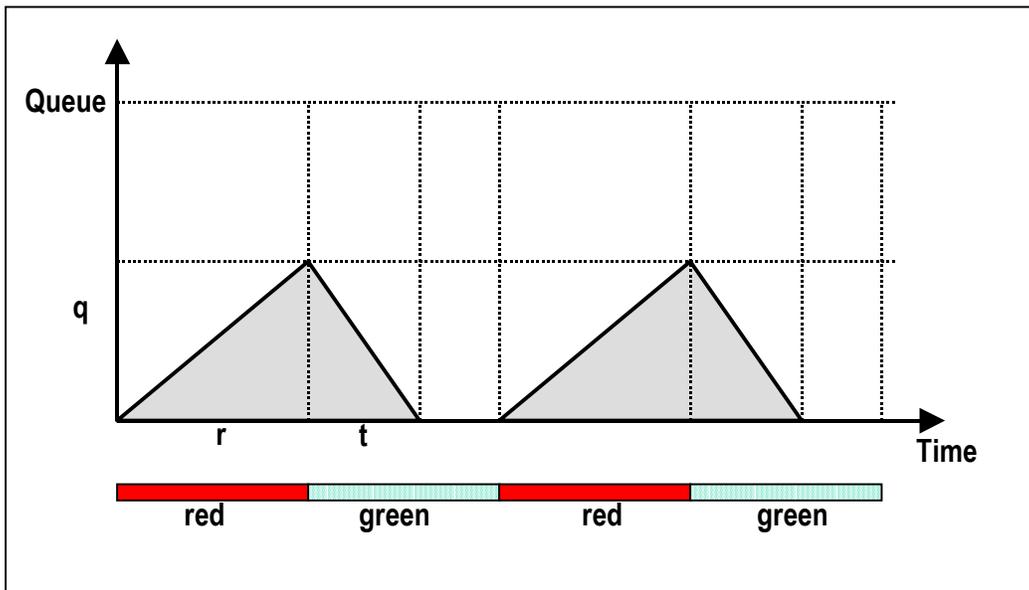


Figure 2.1: Queue Formation at a Signalized Intersection with Uniform Arrivals and Undersaturated Conditions

$$t = \frac{q}{s_i - v_i} \quad (2.1)$$

where:

- t = time to dissipate queue (seconds),
- q = queue size at end of red interval (vehicles),
- s_i = saturation flow rate for lane group i (vehicles/hour/lane),
- v_i = vehicle arrival rate for lane group i (vehicles/hour).

At an undersaturated intersection, the time required to dissipate the queue that forms during the previous red interval is always less than the duration of the green interval. In such case, the total delay generated by the traffic signal operation in each signal cycle corresponds to the shaded area below each triangle in Figure 2.1. If the intersection is oversaturated, the green time during each cycle is insufficient to dissipate the queue that forms during the previous red interval. Consequently, there is in such case a growing residual queue that remains at the end of each cycle and that is propagated to the subsequent cycles, as illustrated in Figure 2.2. In the case of oversaturation, the uniform delay generated by the traffic signal operation thus comprises the area within the triangles T1 and T2, while the delay due to oversaturation is given by the area of the triangle T3. It can also be observed in Figure 2.2 that while the uniform delay remains constant, the oversaturation delay increases as the analysis time horizon increases.

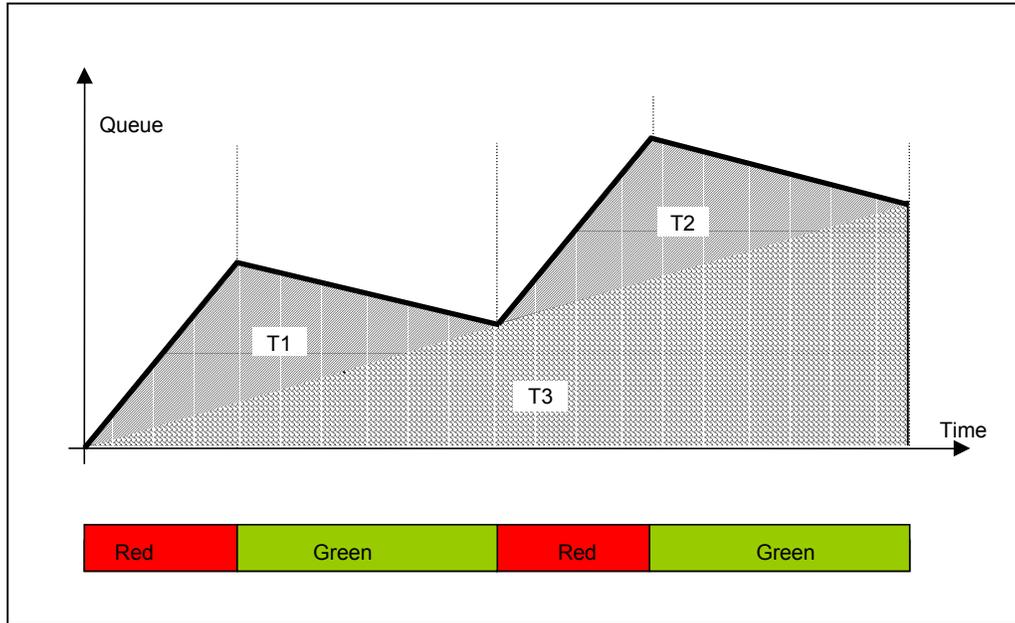


Figure 2.2: Queue Formation Upstream at a Signalized Intersection under Uniform Arrivals in Oversaturated Condition

Using deterministic queuing analysis, Webster (1958) developed a model for estimating the delay incurred by motorists at undersaturated signalized intersections that became the basis for all subsequent delay models. The model developed is presented in Equation 2.2.

$$d = \frac{C(1-\lambda)^2}{2(1-\lambda X)} + \frac{X^2}{2v(1-X)} - 0.65 \left(\frac{c}{v^2} \right)^{1/3} [X^{2+5\lambda}] \quad (2.2)$$

where:

- d = average overall delay per vehicle (seconds),
- λ = proportion of the cycle that is effective green (g/C),
- C = cycle length (seconds),
- v = arrival rate (vehicles/hour),
- c = capacity for lane group (vehicles/hour),
- g = effective green time (seconds).

In Equation 2.2, the first term represents the average delay to the vehicles assuming uniform arrivals. The second term estimates the additional delay due to the randomness of vehicle arrivals. This additional delay is attributed to the probability that sudden surges in vehicle

arrivals may cause the temporary oversaturation of the signal operation. The third term, finally, is an adjustment factor that was introduced in the model to correct the delay estimates and that was developed semi-empirically.

Following Webster's work, numerous studies were conducted on the subject of how to estimate delays at signalized intersections. As a result of these studies, a number of delay models based on deterministic queuing theory were proposed to suite different field conditions. Among these, the most noticeable are the models developed by Miller (1963) and Akcelik (1981), and the models developed for use in the Highway Capacity Manual (TRB, 1985, 1994, 1998) and the Canadian Capacity Guide for Signalized Intersections (ITE, 1984, 1995).

Cronje (1983) analyzed existing formulas, namely, Webster's (1958) and Miller's (1963) equations for average delay, overflow, and average number of stops for under-saturated conditions. These formulas were examined over a large variation of flows and cycle lengths. He concluded that the Miller formula (1963) gave the most accurate results. Akcelik (1988) compared the 1985 HCM delay formula with the 1981 Australian and 1984 Canadian formula. A generalized formula that embraces them all was presented. Akcelik concluded, only the incremental delay (overflow delay) would differ between alternative models. Once total delay is converted to stopped delay, the uniform term - in both the 1985 HCM and Akcelik's equation - will be identical to the first term in Webster's equation. Burrow (1989) compared Akcelik's general formula with an expression of similar form reported by Kimber and Hollis (1978). He found that the uniform term of Kimber and Hollis's formula takes the same value as given by Akcelik with a difference in the incremental terms. Hagen et al. (1989) compared the HCM (1985) delay model with the models used in the Australian Signal Operations Analysis Package (SOAP 1985) and the TRANSYT-7F Release 5 computer packages. They focussed on the effect of the degree of saturation, the peak-hour factor, the length of period of flow observation on delay computations, and the effect of straight-ahead traffic on right-turning vehicles. The results of all these models agreed closely at volume levels below saturation point. When conditions became over-saturated, the models diverged. Thus, an alternative model based on a deterministic queuing process was proposed and evaluated by him.

Teply (1989) studied two approaches to measure delay in the field and explained various problems related to each. He concluded that delay cannot be precisely measured and that a perfect match between the results of an analytical delay formula and delay values measured in the field cannot be expected.

Akcelik and Roupail (1993) proposed a delay model for signalized intersections, which is suitable for variable demand conditions. This model is applicable to the entire range of expected operations, including highly over-saturated conditions with initial queues at the beginning of the analysis period. The model clarified several issues related to the determination of the peak flow period, as well as the periods immediately preceding and following the peak.

2.3.2 STOCHASTIC QUEUING ANALYSIS

Stochastic queuing analysis methods are commonly applied to highway systems to estimate performance characteristics such as delay and queue length. A few example applications include the analysis of signalized (Little, 1971; Newell, 1988) and stop-controlled (Weiss and Maradudin, 1962; Madanat et al., 1993) intersections. However, accurate estimation of vehicle delay using stochastic analysis is difficult because of the randomness of the traffic flow process and the uncertainty associated with the various factors affecting intersection capacity. As a result, the mathematical models currently used to predict delay within stochastic processes usually include several simplifications. The most typical assumptions are that vehicle arrivals follow the Poisson process and that the mean arrival flow rate is constant throughout the period of analysis. In addition, deterministic and stochastic queuing analyses assume that vehicles queue vertically and that they can decelerate and accelerate instantaneously.

Unlike deterministic queuing analysis, the stochastic queuing analysis attempts to estimate the delay incurred by vehicles at a bottleneck or a service area by determining the statistical distributions of delays and queue lengths that result from the assumed arrival and service distributions. Similarly to deterministic queuing analysis, stochastic queuing analysis can also be used for undersaturated traffic conditions. In standard queuing theory, the degree of saturation of

a traffic stream is represented by its intensity (Equation 2.3), which is analogous to the volume to capacity ratio.

$$\rho = \frac{\lambda}{\mu} \quad (2.3)$$

where:

- ρ = traffic intensity,
- λ = mean vehicle arrival rate,
- μ = mean vehicle service rate.

In a stochastic process, undersaturated traffic conditions only occur when the traffic intensity is less than 1. If the intensity is greater than 1, there is no mathematical solution to the problem. To estimate the delay incurred by motorists, one can thus convert the queuing process into a deterministic queuing problem using multiple time slices with varying mean arrival and service rates. Other approaches include the use of microscopic simulation techniques.

There are many types of probability distributions that can be used to model the arrival and discharge processes of vehicles at a transportation facility. A classification scheme based on the more commonly used distributions is shown in Table 2.2. In the table, the letter *D* denotes the use of a constant mean value, while the letters *M*, *E*, and *G* respectively represent the use of a random, Erlang and generalized probability distributions. In the table, it is interesting to note that the assumption of constant arrival and service distributions generally lead to the use of deterministic queuing theory.

Table 2.2: Classification of Probability Distributions Used in Stochastic Queuing Problem

Arrival Distribution	Service Distribution			
	Constant	Random	Erlang	Generalized
Constant	<i>Deterministic queuing</i>	<i>D/M</i>	<i>D/E</i>	<i>D/G</i>
Random	<i>M/D</i>	<i>M/M</i>	<i>M/E</i>	<i>M/G</i>
Erlang	<i>E/D</i>	<i>E/M</i>	<i>E/E</i>	<i>E/G</i>
Generalized	<i>G/D</i>	<i>G/M</i>	<i>G/E</i>	<i>G/G</i>

The limitation of existing deterministic delay formulas seem to justify the need for the use of stochastic delay models that would allow for the calculation of queue size and delay probabilities that evolve over time. In that objective, an approach to study overflow queue evolution based on a Markov chain model was developed to compute delay (Brilon and Wu, 1990). The computations performed by their model enable the evaluation of average delays incurred by vehicles arriving at a fixed-time traffic signal under time-dependent input volumes and Poisson or non-Poisson arrival conditions. They developed a new approximate set of formulas which proved to be in good agreement with empirically based data. Their approach also provided a sound evaluation of the distribution of queue lengths, e.g., the 95-percentile, and their profile over time. In another research effort, Heidemann (1994) derived the probability generating function of the queue length distribution and the Laplace-transform of the delay distribution for both signalized and unsignalized intersections. Olszewski (1994) further investigated the sensitivity of delay models to different degrees of saturation, arrival types and control conditions using delay probability distribution, while Akcelik and Roupail (1994) extended the traditional delay models for signalized intersections to the case of platooned arrivals.

Stochastic queuing methods typically assume that the system in question operate under steady-state conditions over the entire time period of interest. However, the validity of this assumption with respect to traffic-flow facilities is questionable. Therefore, it is somewhat surprising that the vast body of literature concerned with the application of stochastic queuing methods for the assessment of highway operation includes very little discussion on the relevance of the steady-state assumptions. Son *et al.* (1995) are among the only few to have investigated the potential violations of the steady-state assumption. They conducted their analysis based on a simulation

methodology. They found that the assumption of steady-state operation may not always be reasonable and suggested that further investigations regarding the steady-state assumption is needed before using stochastic queuing techniques.

2.4 CAPACITY GUIDE DELAY MODELS

Numerous efforts have been devoted to the development of delay estimation models that would take into account both the deterministic and random aspects of traffic behavior. Following the work of Webster and Cobbe (1966), a number of stochastic delay models that take both these components into account have been developed using queuing analysis principles (Hurdle, 1984; Akcelik, 1988; Teply, 1989; Akcelik and Roupail, 1993 among numerous publications). These models all share the same basic assumptions. First, they still implicitly consider that a vehicle can decelerate and accelerate instantaneously. Second, they all assume that vehicle arrives at an average constant rate within each analysis period. Finally, it is usually assumed that the relation of delay to the arrival pattern is deterministic and that a Poisson distribution can describe the process by which vehicles arrive at an intersection. This section of the dissertation describes the delay models that have been developed for the HCM 1994, the CCG 1995, and the HCM 2000, as these models are among the most widely used in North America.

2.4.1 HIGHWAY CAPACITY MANUAL 1994

In the HCM 1994, the average stopped delay per vehicle for a given lane group, with stopped delay defined as completely immobilized vehicle, is computed using Equations 2.4, 2.5 and 2.6 (TRB, 1994).

$$d = d_1 \times (CF \text{ or } DF) + d_2 \quad (2.4)$$

$$d_1 = 0.38C \frac{\left(1 - \frac{g}{C}\right)^2}{\left(1 - \frac{g}{C} X\right)} \quad (2.5)$$

$$d_2 = 173X^2 \left[(X - 1) + \sqrt{(X - 1)^2 + \frac{m}{c} X} \right] \quad (2.6)$$

where:

- d = stopped delay per vehicle (seconds/vehicle),
- d_1 = uniform stopped delay (seconds/vehicle),
- d_2 = incremental, or random, stopped delay (seconds/vehicle),
- DF = delay adjustment factor for quality of progression and control type,
- CF = Adjustment factor for control type,
- X = volume to capacity ratio of lane group,
- C = traffic signal cycle length (seconds),
- c = capacity of lane group (vehicles/hour),
- g = effective green time for lane group (seconds),
- m = an incremental delay calibration term representing the effect of arrival type and degree and platooning (assumed to be 16 for random arrival conditions).

In Equation 2.4, the parameter d_1 estimates the uniform delay, or delay associated with perfectly uniform arrivals and stable flow in undersaturated flow regimes. This parameter is based on the first term of Webster's delay formulation (Equation 2.2) is widely accepted as an accurate depiction of delay for the idealized case of uniform arrivals in undersaturated traffic conditions (HCM, 1994). As such, this parameter is only valid for the analysis of delays on intersection approaches on which the parameter X , which represents the volume-to-capacity ratio, has a value of less than 1.0.

The second delay parameter, d_2 , estimates the incremental delay caused by the randomness of vehicle arrivals. Similarly to Equation 2.5, Equation 2.6 is only valid for oversaturated traffic conditions. However, the equation may be utilized with some caution for values of X that do not exceed the minimum of 1.2 or 1/PHF, where PHF is the peak hour factor. In cases in which the parameter X exceeds a value of 1.0, the delay estimate then only applies to the vehicles arriving during the first 15-minute of the analysis period as Equation 2.6 does not account for the cumulative effect of residual queues from previous 15-minute periods.

The final element of Equation 2.4 is the delay adjustment factor DF . This factor accounts for the impact of the type of traffic signal control and of the quality of traffic progression between successive signalized intersections on the estimation of uniform delay. Since these two effects are considered mutually exclusive, the parameter DF reflects either the type of traffic signal control or the quality of progression between intersections. If isolated semi-actuated or fully actuated control is used, the adjustment factor takes a value of 0.85. For all other types of control modes, a value of 1.0 is assumed. The adjustment factor for the quality of progression (PF) adjusts for the delay estimates depending on the arrival pattern. This factor is calculated using Equation 2.7. Since poor signal progression will result in a low percentage of vehicles arriving during the green interval, values greater than 1.0 are usually used to reflect such conditions. Similarly, values of less than 1.0 are used to reflect good progression.

$$PF = \frac{(1 - P)f_p}{1 - \frac{g}{C}} \quad (2.7)$$

where:

- P = proportion of vehicles arriving during a green interval,
- g/C = proportion of green time available,
- f_p = supplemental adjustment factor for when the platoon arriving during the green.

Similar to the parameter DF in Equation 2.4, the parameter m in Equation 2.6 accounts for the effect of the arrival type and the degree of platooning of arriving vehicles on the incremental delay. In this case, a value of 16 represents typical random arrival conditions, while lower values are used in coordinated systems to reflect the fact that signal coordination usually reduces demand variations from one cycle to the other by creating repetitive traffic patterns.

2.4.2 CANADIAN CAPACITY GUIDE 1995

The 1995 Canadian Capacity Guide computes the total delay at an intersection using Equations 2.8 through 2.11 (ITE, 1995). It can be demonstrated that Equation 2.9 and the first two components of Equation 2.10 are derived from standard queuing theory.

$$d = d_1 \cdot k_f + d_2 \quad (2.8)$$

$$d_1 = 0.50 \cdot C \cdot \frac{\left(1 - \frac{g}{C}\right)^2}{\left(1 - \frac{g}{C} \cdot \min(X, 1.0)\right)} \quad (2.9)$$

$$d_2 = 15t_e \left[(X - 1) + \sqrt{(x - 1)^2 + \frac{240X}{ct_e}} \right] \quad (2.10)$$

$$k_f = \frac{\left(1 - \frac{q_{gr}}{q}\right) f_p}{1 - \frac{g}{C}} \quad (2.11)$$

where:

- d = average overall delay per vehicle (seconds/passenger car units),
- d_1 = uniform delay (seconds/passenger car units),
- d_2 = incremental, or random, delay (seconds/passenger car units),
- k_f = adjustment factor for the effect of the quality of progression in coordinated systems,
- C = traffic signal cycle time (seconds),
- g = effective green time for lane group (seconds),
- X = volume to capacity ratio of lane group,
- c = capacity of lane group (passenger car units/hour),
- t_e = evaluation time (minutes),
- q_{gr} = arrival flow during the green interval (passenger car units/hour),
- q = total arrival flow (passenger car units /hour),
- f_p = supplemental adjustment factor platoon arriving during the green.

The CCG model computes the overall delay and is identical to the deterministic queuing model. In this model, with reports the overall delay rather than stopped delay, there is no fixed relation between stopped and overall delays. While the HCM 1994 model assumed that 76 percent of all incurred delay could be considered as stopped delay, both Teply (1989) and Olszewski (1993) agree that this factor is incorrect for very low and very high signal delays. To correct this problem Teply proposed using multiplicative adjustment factors that varies with the value of signal delay, while Olszewski recommended a subtractive adjustment factor that varies with the approach speed. Based on their research, the factors of Table 2.3 were developed for use in the 1995 CCG to convert overall delay estimates into stopped delay estimated.

Table 2.3: 1995 CCG Conversion Factors from Overall Delay to Stopped Delay

Red interval (seconds)	20	25	30	40	50	60	70	80	≥ 90
k_1	0.36	0.46	0.56	0.71	0.76	0.76	0.76	0.76	0.83

The 1995 CCG delay model is generally similar to the HCM 1994 model but overcomes some of its limitation. First, Equation 2.10 is not restricted to the analysis of a 15-minute peak period. Second, the model produces delay estimates at intersections with volume to capacity ratio between 1.0 and 1.2 that are more consistent with deterministic queuing theory.

However, caution must be applied when comparing the two models, as they use slightly different approaches to estimate delays. First, Equations 2.9 and 2.10 estimate overall delay while Equations 2.5 and 2.6 compute stopped delay only. Second, the HCM 1994 model uses vehicles to quantify traffic flows while the 1995 CCG uses passenger car units, a measuring unit that takes into account the relative size of vehicles. For instance, while an ordinary car is assumed to correspond to one passenger car unit, 1.5 to 3.5 passenger car units may be associated with a particular truck, depending on its size. Finally, saturation flows are estimated in both models using slightly different methods. As a general rule, it is indicated that the saturation flows estimated by the HCM 1994 are usually 5 percent higher than those estimated by following the

1995 CCG procedure. Therefore, appropriate adjustments must be applied when comparing both models.

2.4.3 HIGHWAY CAPACITY MANUAL 2000

After the release of the Highway Capacity Manual 1994, numerous research has been undertaken to assess the changes that were made in the delay estimated model with respect to the 1985 version of the manual. First, Daniel *et al.* (1996) examined the effects of non-random arrivals on random delay estimates and calibrated the parameters k and I of Equation X to account for nonrandom arrivals, signal controller types, and the quality of traffic progression between successive intersections. Braun and Ivan (1996) and Prevedouros and Koga (1996) further compared the 1985 and 1994 delay models using field data. In another research project, Akcelik (1996) extended the 1994 HCM delay progression factor to account for the prediction of queue length, queue clearance time, proportion of stopped vehicles in a queue, and queue move-up rate. Fambro and Roupail (1997) finally proposed a generalized delay model that corrected some of the problems found in the 1994 HCM model and that is now the delay model found in the HCM 2000.

In the HCM 2000, the average delay per vehicle for a lane group is given by Equations 2.12 to 2.15 (TRB, 1998).

$$d = d_1 \times PF + d_2 + d_3 \quad (2.12)$$

$$d_1 = 0.5C \frac{\left(1 - \frac{g}{C}\right)^2}{\left(1 - \text{Min}\left(1, X\right) \frac{g}{C}\right)} \quad (2.13)$$

$$d_2 = 900T \left[(X - 1) + \sqrt{(X - 1)^2 + \frac{8kIX}{cT}} \right] \quad (2.14)$$

$$PF = \frac{(1 - P)f_p}{1 - \frac{g}{C}} \quad (2.15)$$

where:

- d = average overall delay per vehicle (seconds/vehicles),
- d_1 = uniform delay (seconds/vehicles),
- d_2 = incremental, or random, delay (seconds/vehicles),
- d_3 = residual demand delay to account for over-saturation queues that may have existed before the analysis period (seconds/vehicles),
- PF = adjustment factor for the effect of the quality of progression in coordinated systems,
- C = traffic signal cycle time (seconds),
- g = effective green time for lane group (seconds),
- X = volume to capacity ratio of lane group,
- c = capacity of lane group (vehicles/hour),
- k = incremental delay factor dependent on signal controller setting (0.50 for pretimed signals; vary between 0.04 to 0.50 for actuated controllers),
- I = upstream filtering/metering adjustment factor (1.0 for an isolated intersection),
- T = evaluation time (hours),
- P = proportion of vehicles arriving during the green interval,
- f_p = progression adjustment factor.

This delay model is identical to the 1995 CCG model, except for the addition of the residual delay components d_3 and the use of vehicles instead of passenger car units to quantify traffic flows. The period analysis T is reported in hours instead of minutes, but this change is reflected in the use of a different multiplication factor in each term involving the variable T . In Equation 2.14, parameters k and I are also introduced in the last term of the equation, but this term reduces to the same one of Equation 2.10 when the values associated with pre-timed traffic signal control at an isolated intersection are used.

2.5 TRAFFIC FLOW MODELS

Traffic flow is a comprehensive stochastic process of interactions between drivers, vehicles and the geometric conditions of the roadway. To be able to design and operate transportation systems in the most efficient way, one must therefore be able to model such flow. In particular, one must be able to model the fundamental interactions between the various components of the traffic stream. To do so usually implies determining the relationship between the traffic flow, traffic density and traffic speed. More specifically, the development of traffic flow models usually require the determination of the following elements (Ross, 1988):

- The general equation of traffic stream, where flow is the product of speed and density;
- The equation of conservation of vehicles, where the difference between the number of vehicles entering a link and those leaving it during a given time interval corresponds to the change in the number of vehicles traveling on the link (Lighthill and Whitham, 1955);
- The relationship between speed and density or between the flow-density gradient.

At the present time, there is no unified theory regarding the modeling of traffic flow. Much of the knowledge currently available in this field is largely empirical. Research on this subject was initiated in the 1930's and mainly focused then on the development of speed-flow and speed-density relations based on traffic observations. The importance of the principle of the conservation of vehicles within a traffic stream was not emphasized until the development of high-speed computers allowed for a more extensive use of computer simulation to test the proposed models.

Depending upon the level of flow description used, existing traffic flow models can be divided into two main categories. The first category comprises macroscopic traffic stream models that are only concerned with the average behavior of a stream of vehicles. These models do not attempt to capture the specific interactions between individual vehicles, but simply the fundamental relations between traffic flow, traffic density and traffic speed. The second category comprises microscopic traffic stream models that are more concerned with the modeling of

interactions between individual vehicles. As such, models that belong to this category are often referred to as car-following models.

2.5.1 TRAFFIC STREAM DESCRIPTIVE PARAMETERS

Numerous variables are used to characterize traffic streams. In this section, the following descriptive parameters are introduced and defined:

- average travel speed;
- volume or rate of flow;
- traffic density or concentration;
- lane occupancy;
- spacing and headway;
- gap and clearance.

2.5.1.1 SPEED

Speed expresses a rate of motion and is usually defined in terms of traveled distance per unit time. Because there is generally a wide range of individual speeds that may be observed in a traffic stream, an average travel speed reflecting the general characteristics of the traffic stream is often used in traffic studies in replacement of individual vehicle travel speeds.

To obtain the average speed of a traffic stream, two different calculations can be performed. In the first case, the average travel speed along a given roadway segment can be obtained by weighting the travel time of individual vehicles along the segment. This computation, which is expressed by Equation 2.16, yields the *space mean speed* and the weighting of travel times effectively computes the average amount of time that each vehicles spends over a particular point in space.

$$u_s = \frac{L}{\sum_{i=1}^n \frac{t_i}{n}} = \frac{nL}{\sum_{i=1}^n t_i} \quad (2.16)$$

where:

$$u_s = \text{space mean speed (kilometers/hour),}$$

- L = length of the highway segment (kilometers),
- t_i = travel time of the i th vehicle to traverse the section (hours),
- n = number of vehicles observed.

In the second case, the average travel time can be obtained by taking the arithmetic mean of the measured speeds of all vehicles passing a fixed roadside point during a given interval time. This use of spot speeds, as indicated in Equation 2.17, yields the *time mean speed*.

$$u_t = \frac{\sum_{i=1}^n u_i}{n} \quad (2.17)$$

where:

- u_t = time mean speed (kilometers/hour),
- u_i = spot speed of vehicle I (kilometers/hour),
- n = number of vehicles observed.

2.5.1.2 VOLUME AND RATE OF FLOW

Two measures are often used to quantify the amount of traffic passing a point on a lane or roadway during a designated time interval. The first measure, *volume*, expresses the total number of vehicles that are observed or predicted to pass in front of an observation point during a given time interval. The second measure, *rate of flow*, expresses the same number of observed vehicles in terms of vehicle arrivals per unit time. The distinction between the two quantities is best described by an example. As such, consider an observation point where 200 vehicle passages are observed within a 10-minute period. In this case, the total volume of 200 vehicles would yield a 1200-vehicle/hour arrival ($200 \times 60/10$). In this case, the arrival rate does not indicate that 1200 vehicles were observed over a one-hour period, but simply expresses that vehicles were observed to arrive at that average rate over a 10-minute period. The distinction between these two parameters is important because the flow rate is used to characterize the flow of traffic using the standard traffic flow relationships.

2.5.1.3 DENSITY OR CONCENTRATION

Traffic *density*, or *concentration*, is defined as the number of vehicles occupying a unit length of a lane or roadway at a particular instant. Direct measurement of this parameter can be obtained through aerial photography; however, when the average travel speed and rate of flow of a traffic stream is known, density is more commonly calculated using Equation 2.18.

$$k = q / u \quad (2.18)$$

where:

- k = density (vehicles/kilometer),
- q = rate of flow (vehicles/hour),
- u = average travel speed (kilometers/hour).

2.5.1.4 LANE OCCUPANCY

Lane occupancy is a parameter related to the concept of traffic density. As indicated in Equation 2.19, this parameter expresses the usage of a given traffic lane by computing the ratio of the total length of all vehicles present on the lane to the total length of the lane.

$$R_j = \frac{\sum_{i=1}^{n_j} Lv_i}{L_j} \quad (2.19)$$

where:

- R = Occupancy of lane j ,
- n_j = number of vehicles on lane j ,
- L_j = total length of lane j (meters),
- Lv_i = length of vehicle i (meters).

If the lane occupancy is divided by the average length of a vehicle, one could then obtain that average density of the traffic stream.

2.5.1.5 SPACING AND HEADWAY

Spacing and headway are two additional parameters often used to characterize traffic streams. The first parameter, *spacing*, is defined as the distance between successive vehicles in a traffic stream, and is usually measured from the front bumper of a vehicle to the front bumper of the successive vehicle. Similarly, the second parameter, *headway*, is defined as the time interval between successive vehicles as they pass a point on a lane or roadway. Identical to the spacing between two vehicles, the headway of a pair of vehicles is measured from the moment the front bumper of the first vehicle passes over the observation point to the moment the front bumper of the next vehicles passes over the same point.

Both the spacing and headway are considered to be *microscopic* descriptive traffic stream parameters as they characterize individual pairs of vehicles within the traffic stream. Within any traffic stream, both the spacing and headway of individual pairs of vehicles are usually distributed over a range of values. These distributions are in turn generally related to the speed of the traffic stream and to the prevailing traffic conditions. From a more aggregate, or *macroscopic*, point of view, these two microscopic parameters are also related to the traffic density and the rate of flow. As expressed by Equations 2.20 and 2.21, respectively, vehicle spacing is the mathematical inverse of the traffic flow density, while the headway is the inverse of the traffic flow rate. Equation 2.21 also indicates that the headway between two vehicles is also linked to their spacing through the vehicles' average space-mean speed.

$$s = \frac{l}{k} \quad (2.20)$$

$$h = \frac{3600}{q} = \frac{3600 s}{u} \quad (2.21)$$

where:

- s = spacing (kilometers/vehicle),
- k = density (vehicles/kilometer),
- h = headway (seconds/vehicle),
- q = rate of flow (vehicles/hour),

u = vehicles' average space-mean speed (kilometers/hour).

2.5.1.6 CLEARANCE AND GAP

Similarly to concepts of spacing and headway, the *gap* and *clearance* parameters express the distance between two consecutive vehicles. However, contrary to spacing and headway, which express the distance from the front bumper of a vehicle to front bumper of the next vehicle, the gap and clearance parameter refer only to the distance between vehicles. As indicated in Equation 2.22, the difference between spacing and clearance is the average length of a vehicle. Similarly, as shown in Equation 2.23, the only difference between headway and gap is the time required by a vehicle to travel the average length of a vehicle at the traffic stream average speed.

$$g = h - \left(\frac{L_v}{u} \right) \quad (2.22)$$

$$c = s - L_v = g \cdot u \quad (2.23)$$

where:

- g = gap between two vehicles (seconds),
- h = headway between two vehicles (seconds),
- L_v = average length of a vehicle (meters),
- u = traffic stream average space-mean speed (meter/seconds),
- c = clearance between two vehicles (meters).

2.5.2 MACROSCOPIC TRAFFIC STREAM MODELS

Early research on traffic stream models focused on the development of simple flow-density-speed relationships based on traffic observations. The basis structure of the macroscopic models that were developed in these research projects is defined by the first-order speed-density relationship of Equation 2.24. It should be noted that the space-mean-speed is used in all subsequent relationships and will not be explicitly termed as space-mean-speed.

$$u = u_e(k) \quad (2.24)$$

where:

$u_e(k)$ = equilibrium relationship between speed and density.

The first model to be developed based on Equation 2.24 was produced by Greenshields (1934), who proposed a simple linear relationship between traffic speed and traffic density. The relationship developed by Greenshields is given in Equation 2.25 and is further illustrated in the top left diagram of Figure 2.3. The figure also illustrates the familiar parabolic shape of the speed-flow and flow-density relationships that can be developed from the Greenshields model using the fundamental relationship between speed, flow and density as given in Equation 2.18.

$$u = u_f \left(1 - \frac{k}{k_j} \right) \quad (2.25)$$

where:

u = traffic average speed (kilometers/hour),

u_f = speed at which traffic is moving freely, also known as the *free-flow speed* (kilometers/hour),

k_j = density at which the concentration of vehicles is maximum, also known as the *jam density* as it usually occurs when traffic is completely stopped (vehicles/kilometers).

Greenberg (1959) later developed a logarithmic speed-density model that assumes that the traffic flow along a roadway can be considered as a one-dimensional fluid. The model he developed is given by Equation 2.26. In particular, it was later found that this model could be related to one of the microscopic car-following models that are discussed later.

$$u = u_m \ln \left(\frac{k_j}{k} \right) \quad (2.26)$$

where:

u_m = speed at which the traffic flow reaches the roadway capacity.

Underwood (1961) proposed another macroscopic traffic stream model, given by Equation 2.27.

$$u = u_f e^{-\left(\frac{k}{k_m}\right)} \quad (2.27)$$

where:

k_m = density at which the traffic flow reaches the roadway capacity.

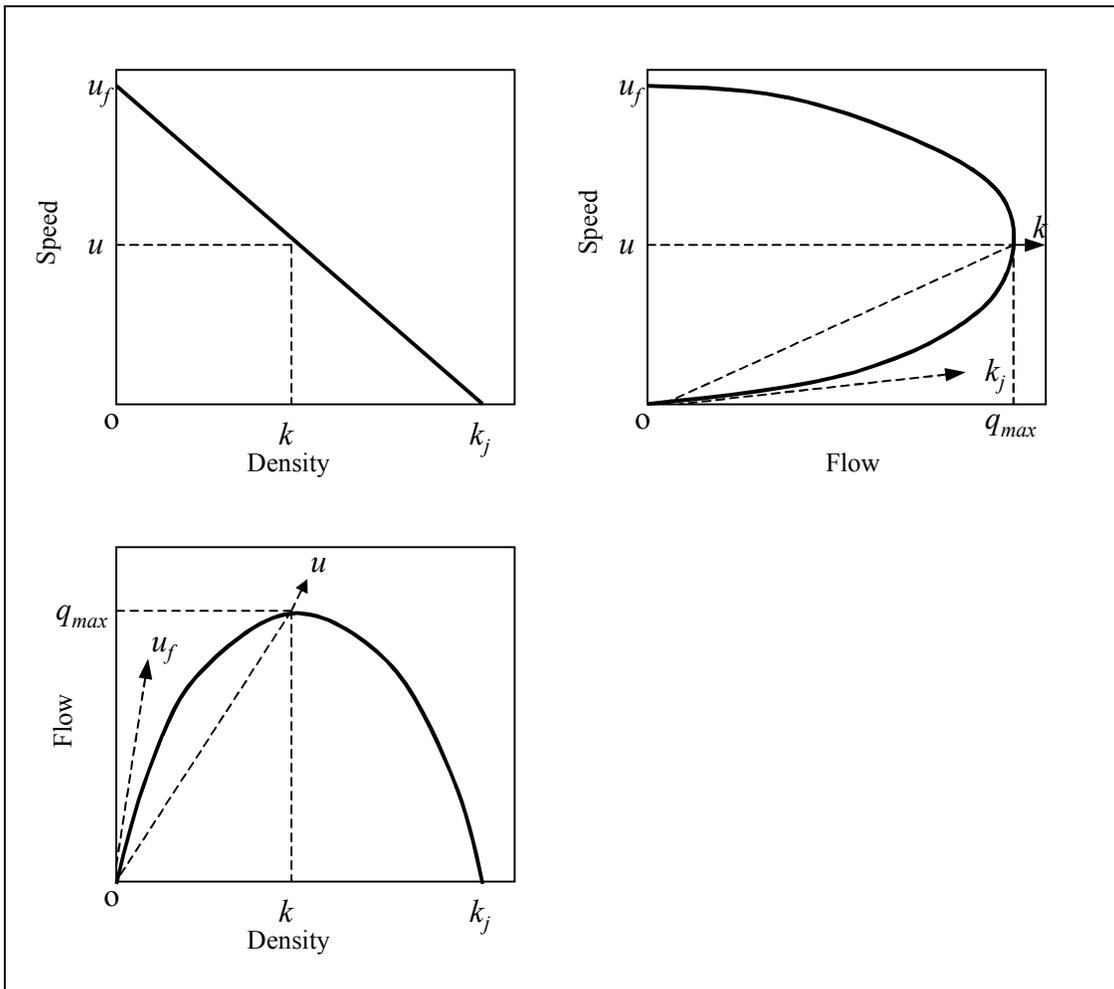


Figure 2.3: Generalized Relationships among Speed, Density, and Rate of Flow.

Drew (1967) introduced an additional parameter β and proposed a family of traffic stream models that could be expressed by Equation 2.28:

$$u = u_f \left(1 - \left(\frac{k}{k_j} \right)^{\frac{\beta+1}{2}} \right) \quad (2.28)$$

where:

β = parameter determining the type of model used.

In Equation 2.28, it is interesting to note that assigning a value of 1 to the parameter β yields the same equation as Greenshields' model (Equation 2.25). Similarly, it can be observed that assigning a value of 0 to the parameter β results in a parabolic speed-density model, while a value of -1 yields a constant model.

In addition to the models discussed above, other macroscopic traffic stream models were proposed Gerlough and Huber (1975), May (1990) and McShane and Roess (1990). Recently, Hall and Gunter (1986) also proposed an inverted V relationship between flow and concentration. Van Aerde (1996) proposed a generalized model that can revert to the Greenshields model, as demonstrated in Equation 2.29.

$$k = \frac{1}{c_1 + \frac{c_2}{u - u_f} + c_3} \quad (2.29)$$

where:

k = density (veh/km),

c_1 = fixed distance headway constant (km),

c_2 = first variable headway constant (km²/h),

c_3 = second variable distance headway constant (h⁻¹),

u = speed (km/h),

u_f = free flow speed (km/h).

2.5.3 MICROSCOPIC CAR FOLLOWING MODELS

The detailed study of driver behavior within traffic streams started in the 1950's, two decades after the development of the first macroscopic traffic stream models. The car-following theory that was developed in these studies assumes the behavior of a motorist following another vehicle can be modeled based on a certain number of highway control rules. Typically, a car-following model is built around the relation of Equation 2.30, which indicates that the response of a driver to a change in traffic conditions is a function of his/her sensitivity to surrounding factors and his/her perception of the change.

$$Response = Sensitivity \times Stimulus \quad (2.30)$$

In the equation, the response could be the acceleration or deceleration of the vehicle ahead. In such case, the stimulus that triggers a reaction from the second driver could then be the change in the relative velocity or relative spacing between the two vehicles. The sensitivity considers for its part the driver's sensitivity to its environment and other factors such as his reaction time and the mechanical response delay of the various components of his vehicle.

Pipes (1953) developed the first microscopic car-following model by relating the velocity of a vehicle to the minimum headway that drivers usually keep with the vehicle in front of them for safety purposes. In his model, the minimum safe distance between a lead and a following vehicle is assumed to be a function of the speed of the following vehicle (in mile per hour) and the length of the vehicle (in feet) in front, as indicated in Equation 2.31. The 1.47 is a conversion factor to convert from mph to ft/s. This equation indicates that the minimum safe distance between two vehicles corresponds at a minimum to one car length and that it increases by one car length for every 10-mile increment in the speed of the following vehicle.

$$d_m = l_n \left(\frac{\dot{x}_{n+1}}{1.47 \times 10} \right) + l_n \quad (2.31)$$

where:

l_n = length of vehicle in front (feet),
 \dot{x}_{n+1} = speed of following vehicle (mile/hour).

Forbes (1958) later improved Pipes' safe distance model by incorporating the time required by a driver to perceive the need to decelerate and apply the brakes. His research lead to the time headway model of Equation 2.32.

$$h_m = \frac{l_n}{\dot{x}_{n+1}} + \Delta t \quad (2.32)$$

where:

Δt = driver's perception-reaction time.

Following the work of Pipes and Forbes, significant advancements in microscopic traffic flow modeling were achieved by General Motors (1959), which used the results of extensive field studies to develop a series of microscopic car-following models that now form the base of modern microscopic traffic flow models. The first model that was developed through this research effort is given by Equation 2.33.

$$\ddot{x}_{n+1}(t + \Delta t) = \alpha[\dot{x}_n(t) - \dot{x}_{n+1}(t)] \quad (2.33)$$

where:

α = sensitivity constant (1/second),
 $\dot{x}_n(t)$ = speed of vehicle n at time t (meters/second),
 $\dot{x}_{n+1}(t)$ = speed of vehicle $n+1$ at time t (meter/second),
 $\ddot{x}_{n+1}(t)$ = acceleration of vehicle $n+1$ at time t (meters/second²).

Later improvements to Equation 2.33 lead to the model of Equation 2.34. In this model, the sensitivity constant was divided by the spacing between the lead and following vehicles to reflect the assumption that the response of the driver of the following vehicle is inversely proportional to the distance between his vehicle and the one in front of him. This change also allowed the

sensitivity constant to be expressed in more practical units. While the sensitivity constant in Equation 2.33 is expressed in units of inverse of time, the same constant in Equation 2.34 is expressed in units of velocity, which relate more directly to the variables used in the model.

$$\ddot{x}_{n+1}(t + \Delta t) = \frac{\alpha}{x_n(t) - x_{n+1}(t)} [\dot{x}_n(t) - \dot{x}_{n+1}(t)] \quad (2.34)$$

where:

α = sensitivity constant (meters/second),

$x_n(t)$ = position vehicle n at time t (meters/second).

The above model was then further refined to yield the model of Equation 2.35. In this model, the driver of the following vehicle is assumed to be more responsive to the relative speed of the lead vehicle with respect to the speed of his own vehicle as he accelerates or decelerates. As a result of this change, the sensitivity constant now became dimensionless.

$$\ddot{x}_{n+1}(t + \Delta t) = \frac{\alpha [\dot{x}_{n+1}(t + \Delta t)]}{x_n(t) - x_{n+1}(t)} [\dot{x}_n(t) - \dot{x}_{n+1}(t)] \quad (2.35)$$

After a generalization of the model of Equation 2.35, a final model was finally developed. This model is expressed by Equation 2.36.

$$\ddot{x}_{n+1}(t + \Delta t) = \frac{\alpha_{l,m} [\dot{x}_{n+1}(t + \Delta t)]^m}{[x_n(t) - x_{n+1}(t)]^l} [\dot{x}_n(t) - \dot{x}_{n+1}(t)] \quad (2.36)$$

In the final General Motors model, it is interesting to note that the car-following model of Equation 2.36 corresponds to Greenberg's macroscopic traffic stream model of Equation 2.26 when values of 0 and 2 are respectively assigned to parameters l and m , respectively.

2.6 DELAY ESTIMATES USING SHOCK WAVE ANALYSIS

Because traffic involves flow, density, and speed, there is a natural tendency to attempt to describe traffic in terms of fluid dynamics. Applying dynamic models to traffic models fluid implies greater concern in the over-all traffic stream behavior than in the interactions between individual vehicles. The first successful attempts at such description were made by Lighthill and Whitham (1955) and Richards (1956), who demonstrated the existence of traffic shock waves and proposed the first theory of one-dimensional waves that could be applied to the prediction of highway traffic flow behavior.

2.6.1 CONTINUITY EQUATION

This section demonstrates the origin of the continuity equation that is the backbone of shock wave analysis. First, consider two traffic counting stations on a one-way link situated in such way that there are no traffic sources or sinks between the stations, and in such way that Station 2 is downstream from Station 1 (Figure 2.4). Second, consider the following definitions:

Δt = duration of simultaneous traffic counting activities at Stations 1 and 2;

Δx = distance between Station 1 and Station 2;

N_i = number of vehicles passing station i during interval Δt ;

q_i = $N_i/\Delta t$ = flow of vehicles passing station i during interval Δt ;

ΔN = $N_2 - N_1$ = difference between departing and arriving flow rates;

Δq = $q_2 - q_1$ = difference between departing and arriving flow rates.

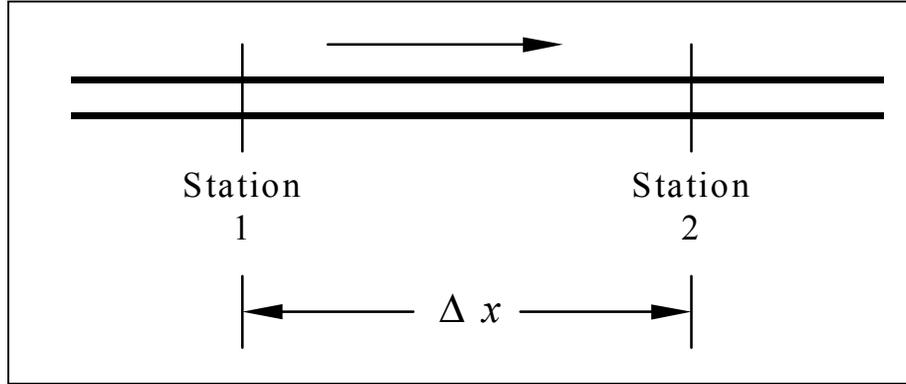


Figure 2.4: Sketch of two closely spaced measuring stations

Now suppose that $N_1 > N_2$, thus implying that there is a buildup of vehicles between Station 1 and Station 2. In this case, Δq would take a negative value and the cumulative buildup of vehicles between the two stations over a period Δt would then correspond to $(-\Delta q)(\Delta t)$. Now, if Δx is of such length that is appropriate to assume that the vehicles are uniformly distributed between the two stations, the increase in concentration Δk of vehicles between the two stations can then be expressed by Equation 2.37 and the buildup of vehicles by Equation 2.38.

$$\Delta k = \frac{-(N_2 - N_1)}{\Delta x} \quad (2.37)$$

$$(\Delta k)(\Delta x) = -\Delta N \quad (2.38)$$

Under the assumption that the number of vehicles is conserved between the two stations, Equation 2.38 can then be rewritten into Equations 2.39 and 2.40.

$$-(\Delta q)(\Delta t) = (\Delta k)(\Delta x) \quad (2.39)$$

$$\frac{\Delta q}{\Delta x} + \frac{\Delta k}{\Delta t} = 0 \quad (2.40)$$

If the medium is now considered continuous and the finite increments are allowed to become infinitesimal, Equation 2.40 thus transforms into Equation 2.41 when the limit theorem is applied.

$$\frac{\partial k}{\partial t} + \frac{\partial q}{\partial x} = 0 \quad (2.41)$$

Equation 2.41 is the continuity expression for a fluid that can also be applied to traffic flows.

2.6.2 CLASSIFICATION OF TRAFFIC SHOCK WAVES

Lighthill and Whitham (1955) were the first to describe a theory of one-dimensional wave motions that could be applied to describe the motion of certain types of fluid and highway traffic flow. Richards (1956) later independently proposed a similar theory for traffic flow. The main postulate of the theory proposed in these two studies is that there exists some functional relation between the traffic flow q and the traffic density k and that this relation could be used to describe the speed at which a change in traffic flow propagates either downstream or upstream from an origin point.

The shock wave model developed Lighthill and Whitman is given by Equation 2.42.

$$SW_{ij} = \frac{q_j - q_i}{k_j - k_i} \quad (2.42)$$

This equation describes the speed at which a change in traffic characteristics propagates along a roadway. Thus, for shock waves propagating downstream, $SW_{ij} > 0$, while $SW_{ij} < 0$ denotes reverse shock waves, or shock waves propagating upstream. Similarly $SW_{ij} = 0$ indicates a stationary. In Equation 2.42, it can also be observed that if the point (q_i, k_i) and (q_j, k_j) are plotted on a flow-density diagram, the magnitude of the slope of line joining the two points then represents the speed of the shock wave between the two traffic states, or SW_{ij} . It must also be observed that in order to perform the calculations, a deterministic macroscopic flow-density relationship similar to the ones discussed in Section 2.5.2 must be used to describe the characteristics of traffic within each traffic area.

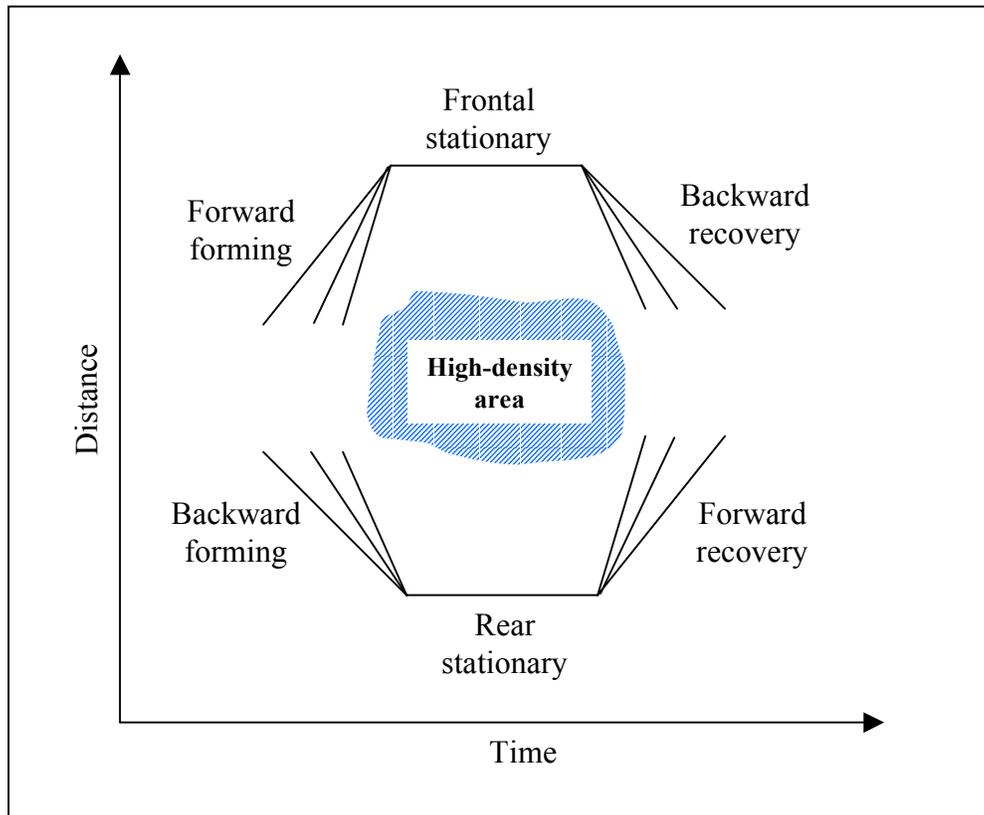


Figure 2.5: Classification of Shock Waves

The time-space diagram of Figure 2.5 illustrates the various types of waves that can be created along a roadway. If the example of a freeway incident is considered, the figure first indicates that two waves can be formed at the beginning of the incident. A first wave would move upstream of the incident, while the second would move downstream. After a certain amount of time, when traffic conditions stabilize, both the frontal and rear waves may become stationary. Then, when the congestion starts to dissipate, both waves would start to move back towards their origin and would eventually disappear.

2.6.3 SHOCK WAVES AT SIGNALIZED INTERSECTION

Shock wave analysis at signalized intersections is a common application because of the concern for the length of queues interfering with upstream flow movements. Examples of application include queues extending out of left-turn pocket lanes into through traffic lanes and queues extending upstream to block adjacent intersections. In addition, shock waves at signalized

intersections can be analyzed if a flow-density relationship is known for the approach to the signalized intersection and if the flow state of the approaching traffic is specified.

Shock waves are generated by the traffic signal, which causes congested conditions to develop near the stop line during the red interval, and capacity conditions to occur in the period during which the queue is discharging at the saturation flow rate. Computation of queue lengths at traffic signal is briefly discussed by Lighthill and Whitham (1955), and analyzed in more detail by Rorbech (1968), who applied the shock wave theory of Lighthill and Whitham to signalized intersections. However, the simple analysis presented in these earlier works is limited to uncongested signalized intersections where queue length computations are not crucial. Furthermore, no particular applications were proposed by Lighthill and Whitham (1955), while Rorbech's (1968) analysis is limited to a linear case and is based on geometric arguments that are not entirely rigorous.

Unlike previous analyses, Stephanopoulos, *et al.* (1979) investigated the dynamics of queue formation and dissipation at isolated signalized intersections by analyzing the vehicle conservation equation along the street. The paper examined the effect of the control variables (cycle length, green and red intervals) and system parameters (arrival rates, capacity) on the length of the stop line queue. A solution to the conservation equation was then obtained by the method of characteristics for general initial and boundary conditions. Michalopoulos, *et al.* (1980) further analyzed the dynamics of traffic downstream of a signalized intersection and on the links between adjacent intersections. Their approach is macroscopic in nature and treats traffic as a continuum fluid. They also demonstrated the existence and behavior of shock waves generated periodically downstream of a traffic signal, and derived analytical expression for describing their propagation along the road.

Finally, Michalopoulos *et al.* (1981a, 1981b) developed a real time traffic signal control algorithm for isolated signalized intersection that minimizes total intersection delays subject to queue length constraints. This model is based on the examination of shock waves generated upstream of the stop lines by the intermittent service of traffic at the signal. The author also

demonstrated that the previously developed traffic models could be employed for deriving analytical expressions for estimating the effective queue size and delays. Subsequently, the results are compared with those of conventional (input-output) models and their differences were identified and explained.

2.6.4 SHOCK WAVES AT FREEWAY INCIDENTS

Freeway incident congestion is a major problem in urban areas. When an incident occurs on a high-volume freeway, a queue usually forms at the location of the incident. This queue, and the resulting congestion, then usually propagates upstream from the scene of the incident, often for several miles, causing significant congestion.

Similarly to the delay caused by the operation of traffic signals, the delay incurred by motorists as a result of freeway incidents can be analyzed using shock wave theory. In achieving this objective, a framework for such analysis was established Messer, Dudek and Friebele (1973). In their study, various traffic states related to incidents were first defined in terms of the fundamental traffic variables of Table 2.4 and Figure 2.6. The shock waves created by a traffic incident were then identified and their velocity estimated using Greenshields' linear traffic flow model (Equation 2.25).

Table 2.4: Dynamic traffic flow mechanisms during incidents

Events	Upstream	Downstream
Before Incidents	Normal flow $q(t), k(t)$	Normal flow $q(t), k(t)$
Incident Occurs	Queuing flow $q_q(t), k_q(t)$	Metered flow $q_m(t), k_m(t)$
Incident Cleared	Capacity flow $q_c(t), k_c(t)$	Capacity flow $q_c(t), k_c(t)$
Traffic Recovered	Normal flow $q(t), k(t)$	Normal flow $q(t), k(t)$

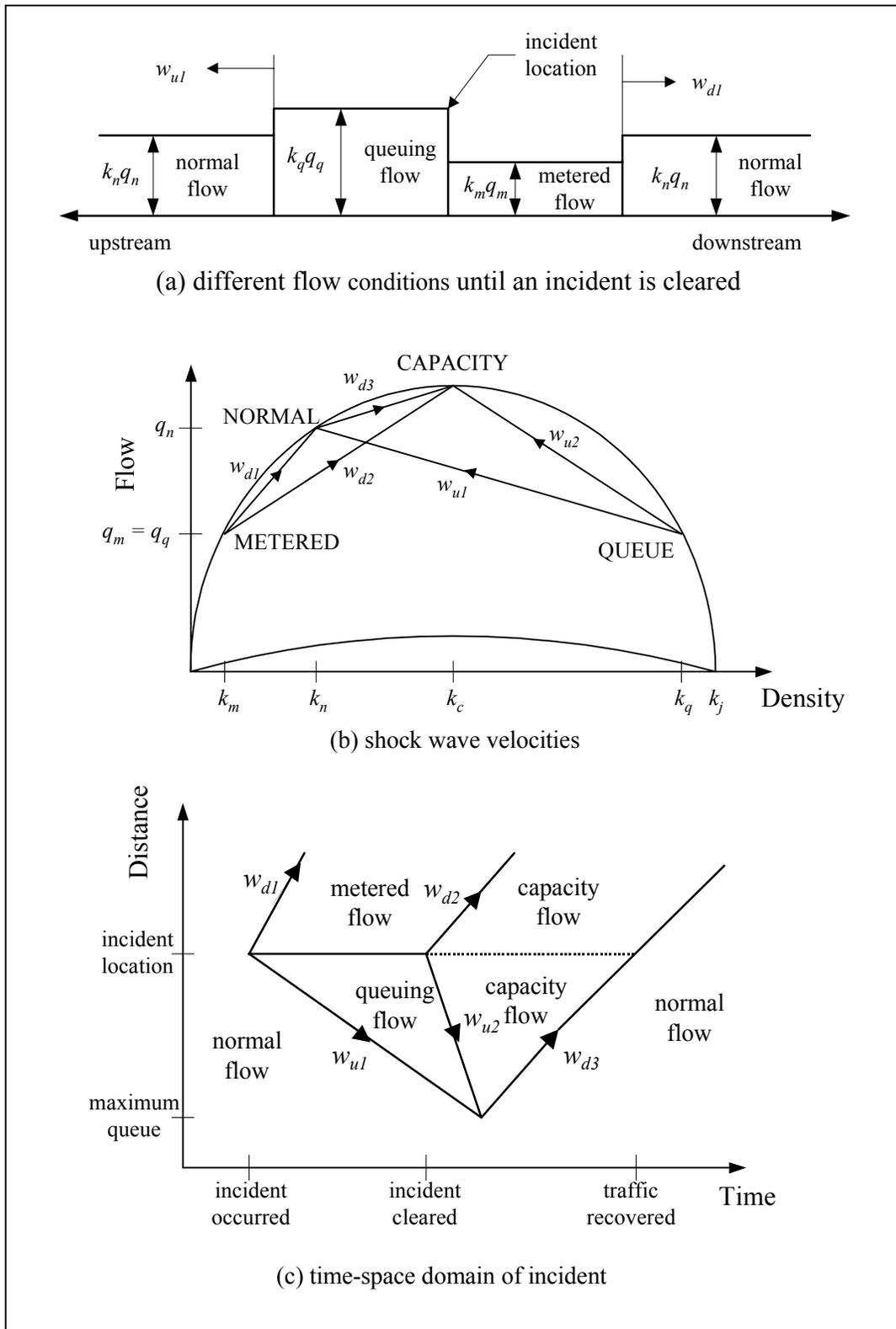


Figure 2.6: Shock wave analysis of incident conditions

In total, five different shock waves were identified, as indicated in the diagrams of Figure 2.6. In the figure, the hatched area is identified the temporal and spatial extent of congestion. Also, note that a subscript d denotes a shock wave traveling downstream, while the subscript u indicates a wave that propagates upstream.

Considering the geometry of the problem, Equations 2.43 and 2.44 can then be derived to determine the maximum queue length and the duration of delay caused by the freeway incident:

$$l_{max} = \frac{w_{u1}w_{u2}}{(w_{u1} - w_{u2})} \cdot T_1 \quad (2.43)$$

$$T_2 = \frac{w_{u1}w_{u2}}{(w_{u1} - w_{u2})} \cdot T_1 + \frac{w_{u1} \cdot (w_{u2} + w_{d3})}{w_{d3} \cdot (w_{u2} - w_{u1})} \cdot T_1 \quad (2.44)$$

where:

- l_{max} = maximum queue length;
- T_1 = duration of an incident;
- T_2 = duration of delay caused by an incident.

Chow (1976) compared the shock wave and queuing analysis methods for calculating total incident delay on a highway section. In his study, he assumed a unique flow-density relationship and derived the equations of total delay. In this case, it was found that the estimated delay was identical in both methods. Chow also concluded that the two methods would have yield different results if he had used a more realistic time-dependent flow-density relationship. Wirasinghe (1978) also used shock wave theory to develop formulas for calculating individual and total delays upstream of incidents. The formulas that were developed through this work are based on areas and densities of regions representing different traffic conditions (mainly congested and capacity regions) that are formed by shock waves in the time-space plot.

More recently, Morales (1987) developed an analytical method that plots the cumulative arrival and departure curves and calculates the cumulative vehicle hours of incident delay and that is now widely used by practitioners and researchers. Al-Deek, Garib and Radwan (1995) extended

this work and proposed a method suitable for the analysis of a corridor equipped with a surveillance system. Using shock wave theory, they first determined the time-space domain of an incident and the freeway links within the domain. Then, instead of using deterministic flow-density relationship, densities were estimated by using the ratio of measured flow to measured speed. After the densities, the longitudinal domain of each link was determined by taking as boundary the midpoint between adjacent detectors and travel time on each link during the incident estimated by dividing the link distance by the speed measurement. By referring to the historical link speeds, the total delay for each link and the total delay caused by an incident were then estimated using Equations 2.45 and 2.46.

$$d_j^i = l_j \cdot \left(\frac{1}{u_j^i} - \frac{1}{u_j^{i,r}} \right) \cdot q_j^i \cdot \Delta t \quad \text{for } u_j^i < u_j^{i,r} \quad (2.45)$$

$$TD = \sum_{i=1}^m \sum_{j=1}^n d_j^i \quad (2.46)$$

where:

d_j^i = delay on link j during time interval i ,

l_j = length of link j ,

u_j^i = speed measurement on link j during time interval i ,

$u_j^{i,r}$ = historical speed on link j during time interval i ,

q_j^i = flow measurement measured at the midpoint on link j during time interval i ,

Δt = data aggregation rate or length of time interval i ,

m = number of time intervals during a duration of delay,

n = number of freeway links within a time-space domain of an incident.

In another study, McShane and Roess (1990) finally suggest that the queuing analysis and shock wave delay methods yield different results when applied to the analysis of a bottleneck. Using a simple numerical example, they demonstrated that the use of deterministic queuing analysis methods might underestimate the overall magnitude of delay when compared to the shock wave methods. Nam (1998) found similar results when analyzing bottlenecks on congested freeways.

In another study, Chin (1996) indicates that both the queuing and shock wave analysis models yield similar results, provided that the problem under consideration is modeled appropriately in each method. However, the validity of his conclusion is weakened by the fact that he used the actual number of vehicles that arrived at the study bottleneck to make its comparison rather than the expected number of vehicles causing its proposed shock wave model to underestimate delays.

2.7 THE INTEGRATION MODEL

2.7.1 MODEL OVERVIEW

The INTEGRATION model was conceived during the mid-1980's as an integrated simulation and traffic assignment model (Van Aerde, 1985). What made the model unique was that its approach utilized the same logic to represent both freeway and signalized links, and that both the simulation and the traffic assignment components were mesoscopic, integrated and dynamic (Van Aerde and Yagar, 1988a and b; Van Aerde and Yagar, 1990). In order to achieve these attributes, traffic flow was represented as a series of individual vehicles that each followed pre-specified macroscopic traffic flow relationships. The combined use of individual vehicles and macroscopic flow theory resulted in the model being considered mesoscopic by some.

During the subsequent decade, the INTEGRATION model evolved considerably from its original mesoscopic roots. This evolution took place through the addition, enhancement and refinement of various new features. Some of these improvements, such as the addition of car-following logic, lane-changing logic, gap acceptance logic, and dynamic traffic assignment routines, have noticeably enhanced the original traffic flow model. Other additions have also extended the model's application domain. Among these additions are the inclusion of features for modeling toll plazas, vehicle emissions, weaving sections, and high-occupancy vehicles. Finally, some features, such as the real-time graphics animation and the extensive vehicle probe statistics, have also been added to simply make the model more user-friendly and easier to calibrate.

The current version of INTEGRATION features a fully microscopic traffic simulation model that tracks the lateral and the longitudinal movements of individual vehicles at a resolution of up to

one deci-second. This microscopic approach permits the analysis of many dynamic traffic phenomena, such as shock waves, gap acceptance, and weaving. These attributes are usually very difficult or infeasible to capture under non-steady state conditions using a macroscopic rate-based model. As an example, average gap acceptance curves cannot typically be utilized at permissive left turns if the opposing flow rate varies from one cycle to the other or within a particular cycle. These curves cannot also be used if the size of the acceptable gap varies as a function of the length of time during which a driver has been waiting to find an acceptable gap. In another example, most macroscopic models cannot model platoon progression between adjacent traffic signals if the intersections are operated on a common signal cycle or on a cycle that is a multiple of a reference cycle.

In addition to the above features, the current INTEGRATION model is not restricted to hold departure rates, signal timings, incident severity, and traffic routings at a constant setting for any particular common time period. This implies that the model can consider virtually continuous time varying traffic demands, routings, link capacities and traffic controls without the need to pre-define an explicit common time-slice duration between these processes. Consequently, instead of treating each of the above model attributes as a sequence of steady-state conditions, as needs to be done in most macroscopic rate-based models, all of these attributes can be changed on a virtually continuous basis over time.

The microscopic approach used in INTEGRATION also permits considerable flexibility in terms of representing spatial variations in traffic conditions. For example, while most rate-based models consider traffic conditions to be uniform along a given link, INTEGRATION permits the density of traffic to vary continuously along the link. Such dynamic density variation along an arterial link permits the representation of platoons departing from traffic signals and the associated propagation of shock waves in an upstream or downstream direction, or both directions.

Finally, it is important to note that the microscopic features of the model have been carefully calibrated to allow it to capture the macroscopic traffic features that traffic engineers usually use

when conducting traffic analysis. As an example, the model provides link speed-flow relationships, multi-path equilibrium traffic assignment, uniform, random and over-saturation delay estimates, as well as weaving and ramp capacities. Over the years, the main challenge in the design of the INTEGRATION model has been to ensure that these important macroscopic features automatically remain an emergent behavior, arising from the more fundamental microscopic model rules that are needed to represent the system dynamics using a single integrated approach. In addition, it should be noted that the INTEGRATION model assumes instantaneous driver perception and reaction times.

As a result, the current microscopic approach of the INTEGRATION model is considered as a means to an end, rather than as an end in itself. While the choices that have been made over the years significantly increase the memory and computational requirements of the model, the current model is perceived to yield some critical improvements in the accuracy with which it can represent the dynamics of traffic conditions at an operational level of detail.

2.7.2 MODELING SIGNALIZED INTERSECTIONS

This section describes in some detail the modeling of signalized intersections within the INTEGRATION model. More specifically, this section covers the following topics:

- Modeling of signal cycles;
- Shock waves at traffic signals;
- Estimation of uniform, random and oversaturation delay;
- Gap acceptance modeling for left-turning vehicles;
- Signal coordination between adjacent intersections

2.7.2.1 MODELING OF SIGNAL CYCLES

Within INTEGRATION, a signalized link is identical in virtually all respects to a freeway link. The only exception is that the exit privileges of this link may periodically be suspended, and that the free-speed and saturation flow rates usually take on slightly lower values (Rakha, *et. al.*, 1993).

On a signalized link, the suspension of exit privileges is set to occur when the traffic light indicates an effective red. This suspension of exit privileges is modeled by assuming that a vehicle is positioned just beyond the stop line at the end of each lane each time the traffic light is red. Since vehicles must obey at all times the link's car-following logic, the addition of a *virtual* vehicle just beyond the intersection stop line thus creates a reduction in the perceived headway of the first vehicle that is currently approaching the intersection. This reduction in perceived headway then causes the approaching vehicle to slow down as its headway to the traffic signal decreases and to eventually come to a complete stop just upstream of the stop line. Subsequent vehicles then automatically queue upstream of the first vehicle in a horizontal queue, where the minimum spacing of vehicles in this horizontal queue is governed by the user specified jam density.

When the effective green indication commences, the *virtual* vehicle that was introduced at the stop line of each lane is removed. As a result, the first vehicle in queue faces an uninterrupted headway down the next downstream link and started to accelerate. Other vehicles then subsequently start to accelerate as soon as their perceived headway with the vehicle in front become larger.

2.7.2.2 SHOCK WAVES AT TRAFFIC SIGNALS

When a traffic signal turns red, the stopping of approaching vehicle creates a shock wave that corresponds to the tail of the growing stop line queue. Since shock wave theory applies to both freeways and arterials, the rate at which the tail of the queue moves upstream along the link can be determined in a standard fashion. This rate is equal to the ratio of the "arrival rate at the tail of the queue", divided by the "net difference between the density of the queued vehicles and the density of the arriving traffic". The dynamic nature of the model's car-following logic also permits the rate at which the queue grows to vary dynamically with the varying arrival rate as a function of time during the cycle.

At the beginning of a green interval, the initial acceleration of the first vehicle in a queue, together with the subsequent impact of the model's car-following logic on any additional vehicles, causes two additional shock waves to form concurrently, as illustrated in Figure 2.7. The first shock wave that is created moves downstream from the intersection stop line. It consists of the front of the surge of traffic that crosses the stop line at saturation flow. The second shock wave moves upstream from the stop line. It develops when queued vehicles start to accelerate as the vehicles ahead of them accelerate. This backward moving shock wave constitutes the dividing lines between those vehicles that are still stationary and those vehicles that have begun to accelerate to a speed associated with the saturation flow rate. The speed of this second shock wave is again a function of the speed-flow characteristics of the link.

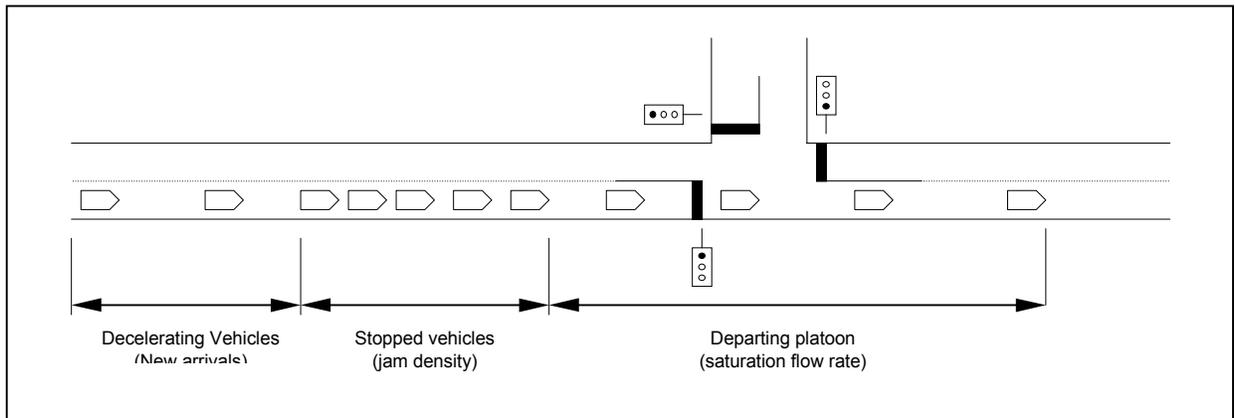


Figure 2.7: Traffic flow dynamics at a traffic signal several seconds into a green phase

2.7.2.3 UNIFORM, RANDOM AND OVER-SATURATED DELAY

In INTEGRATION, the complex problem of analytically estimating the expected delays associated with the random and varying demands is effectively circumvented as the uniform, random and oversaturation delays are an emergent behavior of the model. Instead of being computed based on an explicit mathematical formula, the delays are computed by directly comparing the actual travel time of individual vehicles along a link with an expected travel time reflecting free-flow conditions, i.e., reflecting the absence of traffic signals.

2.7.2.4 GAP ACCEPTANCE MODELING FOR LEFT-TURNING VEHICLES

One of the most complex modeling tasks in estimating the capacity of signalized intersections is the treatment of permissive left turns and right turns on red. Within INTEGRATION, a microscopic gap acceptance model is utilized to reflect the impact of opposing flows on opposed left turners and right turners on red (Velan and Van Aerde, 1996). This opposition is automatically customized by the model at each intersection by means of a built-in gap acceptance logic. This logic specifies which opposing movements are in conflict with the movement of interest and also determines which of the turning movements are opposed within a shared lane or shared link. Based on this information, the model can then automatically provide opposition to left turners when the opposing flow link discharges concurrently. However, the model can also automatically allow the discharge rate to revert back to the unopposed saturation flow rate whenever the opposed movement is given a protected phase.

The incorporation of gap acceptance logic within INTEGRATION permits the model to evaluate the impact of protected versus permissive left turn phases. This logic also allows the model to consider the impact of leading versus lagging greens, and the impact of the duration of the left turn phase in great detail. In addition, the gap acceptance logic can work concurrently with the queue spill-back model to determine when (or if) vehicles in a left turn bay spill back into the through lanes, or conversely, to determine when the through lanes spill back to cut off entry into the left turn bays. In addition, the combination of lane striping, used to allocate certain lanes to exclusive traffic movements, and the selective opposition of vehicles permits the implicit computation of shared lane saturation flow rates.

2.7.2.5 SIGNAL COORDINATION

The INTEGRATION model is capable of evaluating the impact of alternative signal coordination effects. This evaluation is made possible by the fact that the pattern of discharged vehicles is preserved downstream of signalized intersections as platoons of vehicles travel toward the next intersection. Naturally, this preservation of discharging patterns is subject to the dispersion of traffic that results from the variability of individual vehicle speeds, and to the inflow and outflow of vehicles at mid-block, non-signalized intersections.

More importantly, however, the INTEGRATION model is not constrained to operate all traffic signals on a common cycle length. Consequently, it is possible to explicitly evaluate the impact of a lack of coordination at the boundary of two control areas, or to evaluate the impact of taking a signalized intersection out of a coordinated area. In particular, the latter evaluation might be carried out to explore the relative benefits of placing one intersection under some form of critical intersection control.

Signal timing plans are optimized using the methods of Webster and Cobbe (1966) as implemented in the Canadian Capacity Guide (ITE, 1984). This method is similar, but not identical to, the method described in Chapter 9 of the 1994 Highway Capacity Manual.

2.7.3 MEASURES OF EFFECTIVENESS

Since vehicle movements within the INTEGRATION model are predicted using speed-flow and car-following relationships, the model does use an explicit link travel time function to evaluate traffic flow performance in a fashion similar to most macroscopic or planning oriented traffic assignment models. Instead, link travel time emerges as the weighted sum of the speeds that vehicles experienced as they traversed each link segment. This distinction introduces in INTEGRATION both a level of complexity and level of accuracy not present in most other simulation models.

Specifically, the dynamic temporal and spatial interactions of shock waves, which form upstream of a traffic signal, or along a congested freeway link, are such that the final link travel time is neither a simple function of the inflow nor the outflow of the link. Instead, the travel time is a complex product of the traffic flow time series and associated dynamics along the entire link. In addition, the temporal interactions of this flow with the signal timings and flow oppositions at the end of these links are also considered. In this case, the strength of a microscopic approach is that there is no need, beyond the basic car-following/lane-changing/ gap-acceptance logic, for any further analytical expressions to estimate either uniform, oversaturation, coordination, random, left-turn or queue spill-back delay. While such complexity precludes the simplicity of a

functional relationship, such as the Bureau of Public Roads relationship, it does permit two distinct travel times to be properly considered for the same flow level, depending on whether forced or free-flow conditions prevail. It also allows the model to deal much more readily with the concurrent presence of multiple vehicle/driver types on the same link.

2.7.4 EVALUATION OF CAPACITY

The value of saturation flow or capacity is specified in the model's input data files in vehicles per hour per lane for each link in the network, as are the number of lanes per link. The product of the saturation flow per lane and the number of lanes then yield the overall capacity of the link in vehicles per hour.

The inverse of the capacity is the time headway between successive vehicles passing a single point. Similarly, the capacity time headway is considered to be the minimum allowable time headway between successive vehicles. In a simulated network, this minimum time headway is enforced both at the entrance and exit of a links and vehicles are not allowed to proceed at time headways smaller than this minimum value.

2.8 COMMENTS

Traffic operations are highly variable and dynamic. The constituent factors that combine to create an observable traffic stream on a highway are many and difficult to quantify. Elements such as driver characteristics, roadway geometries, vehicle type and weather all affect the operations that are measured. No traffic stream is ever exactly the same as any other. In this spatially and temporally dynamic environment, the current analytical methods in practice are severely limited in predicting delay, number of stops and queue length estimates at signalized intersections. This limitation is one of the main impetuses behind the research included in this dissertation.