64

Highway Traffic Operations

Andrzej P. Tarko
Purdue University

64.1 Introduction

Highway traffic operations are influenced by the behavior of drivers. A highway can be used by a finite number of vehicles, and the driver perceived safe distances between vehicles determine this limit. For a given speed, as distances become shorter, more vehicles can use the highway. Both the volume of drivers choosing to use the highway (demand) and the maximum volume that can be served (supply) depend on driver behavior. Congestion results from too many people attempting to reach their destinations at the same time using the same highways. The combination of demand, capacity, and certain infrastructure features determines how drivers perceive the traffic conditions. Transportation agencies strive for economical solutions to congestion that satisfy a majority of highway users.

This chapter provides introductory material to the traffic operations area, but not all aspects of highway traffic operations are discussed in great detail. Rather, this chapter provides (1) an appreciation of field observations as the most reliable source of operations information for existing highways, (2) a basic understanding of the behavior of traffic flows in relation to available methods of traffic and road improvements, and (3) an awareness of the methods available for highway traffic analysis.

The basic characteristics of traffic flows, together with their measurement types, are introduced first, followed by a brief overview of measurement techniques available. The relationships between traffic flow characteristics that allow the converting of collected traffic data into usable information are then explored. Important terms such as *free-flow speed*, *capacity*, and *jam conditions* are introduced. The theory of *shock*

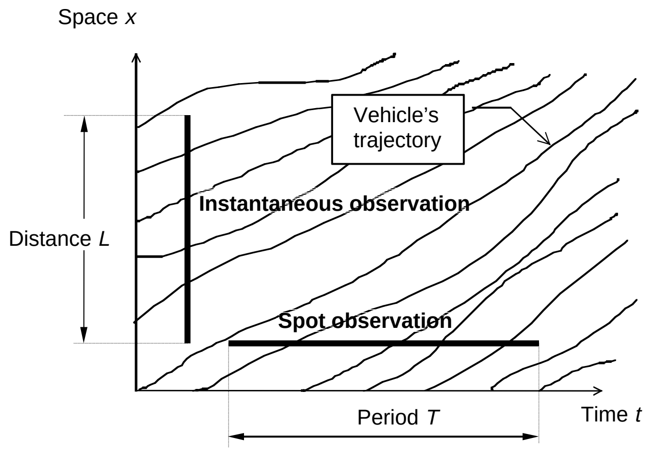

FIGURE 64.1 Spot and instantaneous observations.

waves is presented next, as a means to describe the behavior of sustained traffic queues at highway bottlenecks. Although practicing engineers do not often use shock wave theory, learning the concept contributes to a better understanding of the dynamics of traffic conditions. Queues in random traffic and the relationship between queues and delays are covered within *queuing theory*; and equations for estimating capacity for selected road facilities are provided, as capacity is an important factor in queuing theory. The major characteristics of traffic have been introduced at this point, enabling explanation of the last important concept – *quality of traffic*. *Traffic control* is also briefly reviewed for its importance in improving highway traffic operations.

64.2 Traffic Flow Characteristics and the Fundamental Relationships

Since traffic is strongly influenced by human behavior, even advanced methods of predicting traffic operations are burdened with a considerable degree of uncertainty. The best source of information about traffic operations is the highways themselves. If a traffic engineer wants to learn about traffic conditions on an existing highway, the best method is measurement.

Two pieces of data often collected are counts and speeds. Vehicles can be counted and their speeds measured as they pass a specific location during an observation period. This type of observation is called *spot observation*. There is a second type of observation, *instantaneous observation*, in which vehicles on a designated highway segment are counted and their speeds measured instantly. Figure 64.1 explains the difference between the two types of observations. Spot observations are preferred over instantaneous observations because they are more convenient and less expensive.

This section introduces four traffic flow characteristics: traffic density and space-mean speed directly estimated with instantaneous observations, and traffic volume and time-mean speed directly estimated with spot observations. The fundamental relationship between volume, speed, and density is then introduced, which allows the measuring of all four traffic flow characteristics with preferred spot observations.

Traffic Flow Characteristics

Instantaneous observations render the number of vehicles on some highway segment at a given time of measurement. This count can be converted to the so-called *traffic density* D:

$$D = \frac{N}{L \cdot n} \tag{64.1}$$

where D is measured in vehicles per mile per one lane, L is the length of the segment expressed in miles, and n is the number of traffic lanes. Typically, density is measured in each direction separately. The speeds of vehicles at the moment of measurement can be averaged to calculate the so-called *space-mean speed* S_L:

$$S_L = \frac{1}{N}\sum_{i=1}^{N} S_i \tag{64.2}$$

where S_L = calculated based on N instantaneous observations
S_i = the speed of vehicle i on the highway segment

Spot observations render the number of vehicles passing a spot during an observation period. This count can be converted to *traffic volume* V:

$$V = \frac{N}{T} \tag{64.3}$$

where V is the volume typically expressed in vehicles per hour, estimated based on the count N observed for period T (in hours). The speeds of vehicles are sometimes measured in spot observations, and the average speed, called *time-mean speed*, is calculated as follows:

$$S_T = \frac{1}{N}\sum_{i=1}^{N} S_i \tag{64.4}$$

where S_T is the time-mean speed calculated based on N spot measurements and S_i is the speed of vehicle i included in the observation. The time-mean and space-mean speeds measured for the same traffic are typically not equal. This phenomenon will be explained later in this chapter.

Spot observations must be planned in accordance with the time of day, week, and year that are to be represented by the measurements. Due to the considerable random and periodic variability of traffic, the observation period must be adequate for the purpose. Traffic observations in short intervals may vary randomly; thus a short observation period of seconds or minutes is not recommended to represent an entire hour. On the other hand, traffic fluctuations over the period of an entire day make a several-hour traffic observation nonrepresentative for any particular hour within the observation period as well as for the entire day. It is prudent to observe traffic for 1 hour to obtain hourly traffic estimates and for 1 day for daily traffic estimates. Of course, if the periodic fluctuation is known, then even hourly observations can be converted to daily observations by using appropriate conversion factors. *The Manual of Traffic Engineering Studies* (Robertson et al., 1994) and McShane et al. (1998) provide more information on traffic variability, adequate observation periods, and various traffic studies.

Fundamental Traffic Flow Relationship

Let us assume that all vehicles move at the same speed S. The time headway between two consecutive vehicles h is the distance x between these two vehicles divided by speed S: $h = x/S$. If someone measures the time headways and the distances between consecutive vehicles and calculates the average values $\bar{h}$ and $\bar{x}$, the relationship between the average values will be preserved: $\bar{h} = \bar{x}/S$. It is also true that $1/\bar{h} = S/\bar{x}$. Please notice that the reverse of average intervehicle time is volume V and that the reverse of average intervehicle distance is density D. Thus, we can write

$$V = S \cdot D \tag{64.5}$$

Now, let us assume that a traffic flow consists of k classes of vehicles. All vehicles of class i move at speed S_i. If V_i is the volume of i-class vehicles and k_i is the density of i-class vehicles, then we can claim that $V_i = S_i \cdot D_i$. Since the total volume is

$$V = \sum_{i=1}^{k} V_i$$

and the total density is

$$D = \sum_{i=1}^{k} D_i$$

we can say that

$$V = \sum_{i=1}^{k} S_i \cdot D_i = D \sum_{i=1}^{k} S_i \frac{D_i}{D}$$

Since

$$\sum_{i=1}^{k} S_i \frac{D_i}{D}$$

is the space-mean speed S_L (densities are obtained in instantaneous observations), it is proven that

$$V = S_L \cdot D \tag{64.6}$$

This relationship is called the fundamental traffic flow relationship.

Estimating Space-Mean Speed and Traffic Density from Spot Observations

The source of discrepancy between the time-mean and space-mean speeds can be demonstrated by considering a segment of highway with two classes of vehicles: slow and fast (Fig. 64.2 and Table 64.1). The difference between the two speeds is caused by the higher fraction of fast vehicles in total volume than in total density. Although the difference in the two speeds is negligible in typical traffic conditions, it may be considerable in a particular situation, such as when some vehicles are stopped. Zero speeds significantly reduce the average space-mean speed, but they are not included in the time-mean speed, since the volume of the vehicles with zero speed is zero.

Spot observations render speeds of vehicles that pass a particular spot. Straight averaging gives the time-mean speed. To estimate the space-mean speed, let us first consider vehicles having the same speed. Let N_i be a count of vehicles with speed S_i. The corresponding volume can be estimated as $V_i = N_i/T$.

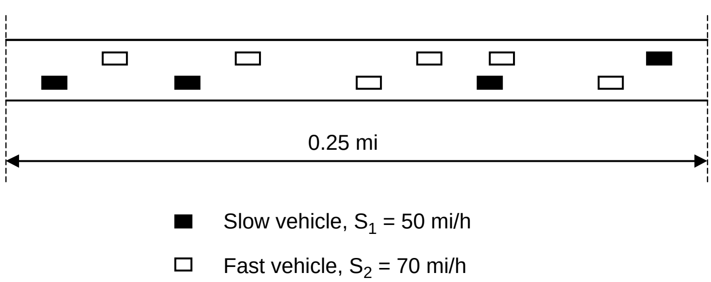

FIGURE 64.2 Example of a highway section for speeds comparison.

TABLE 64.1 Comparison of the Space-Mean and Time-Mean Speeds

Variable	Equation	Calculation
	Space-Mean Speed	
Density of slow vehicles	$D1 = N_1/L$	4/0.25 = 16 veh./mi.
Density of fast vehicles	$D2 = N_2/L$	6/0.25/2 = 24 veh./mi.
Total density	D	16 + 24 = 40 veh./mi.
Fraction of slow vehicles on segment	D_1/D	8/20 = 0.40
Fraction of fast vehicles on segment	D_2/D	12/20 = 0.60
Space-mean speed on segment	$S_1 \cdot D_1/D + S_2 \cdot D_2/D$	50(0.40) + 70(0.60) = 62.0 mi./hr
	Time-Mean Speed	
Volume of slow vehicles	$V_1 = S_1 \cdot D_1$	50 _ 16 = 800 veh./hr
Volume of fast vehicles	$V_2 = S_2 \cdot D_2$	70 _ 24 = 1680 veh./hr
Total volume	$V = V_1 + V_2$	800 +1680 = 2480 veh./hr
Fraction of slow vehicles in volume	V_1/V	800/2480 = 0.323
Fraction of fast vehicles in volume	V_2/V	0.677
Time-mean speed at spot on segment	$S_T = S_1 \cdot V_1/V + S_2 \cdot V_2/V$	50(0.323) + 70(0.677) = 63.5 mi./hr

Note: veh. = vehicle.

Using the relationship derived for flows with the same speed, $V_i = S_i\, D_i$, we can claim that $D_i = V_i/S_i$ or $D_i = N_i/(S_i \cdot T)$. Since the space-mean speed is

$$S_L = \frac{1}{D}\sum_{i=1}^{k} S_i D_i$$

we can say that

$$S_L = \frac{\sum_{i=1}^{k} S_i \cdot N_i / \left(S_i T\right)}{\sum_{j=1}^{k} N_i / \left(S_j T\right)}$$

Since T and some speeds S_i cancel out, the formula becomes

$$S_L = \frac{\sum_{i=1}^{k} N_i}{\sum_{j=1}^{k} N_i / S_j} = \frac{N}{\sum_{j=1}^{k} N_i / S_j}$$

The result will not change if each vehicle constitutes its own class, even if some other vehicles have the same speed. The final formula is

$$S_L = \frac{N}{\sum_{j=1}^{N} \frac{1}{S_j}} \tag{64.7}$$

The density near the spot can be estimated using the fundamental traffic flow equation, $D = V/S_L = (N/T)/S_L$, or, after using Eq. (64.7),

$$D = \frac{1}{T} \sum_{j=1}^{N} \frac{1}{S_j} \tag{64.8}$$

Please notice that spot observations may not include vehicles with zero speeds.

64.3 Measuring Techniques

Instantaneous Observations

An instantaneous observation requires an observer to be at a sufficient elevation to see all vehicles on an observed highway segment. A tall building located near the highway segment is the least expensive solution. However, in many cases tall buildings are not available, so aerial photography is used. An equipped aircraft takes two photographs in short succession and the vehicle speeds are estimated from their shifts along the highway segment. Instantaneous observations are expensive and are not often used. As has been shown, spot observations may be used to estimate space-mean speed and density.

Spot Observations

In manual techniques, human observers count and classify vehicles and sometimes measure their speeds. This technique is accurate but expensive. Machine measurements are less expensive, and numerous techniques and technologies are available. This section briefly overviews several commonly used alternatives.

Vehicles are detected by devices called detectors, which utilize various physical phenomena such as perturbation of electromagnetic or magnetic fields, changes of pressure in rubbers tubes, generation of electrical field in piezoelectric materials, detection of energy reflected or generated by a vehicle, and the Doppler phenomenon caused by a vehicle. The types of energy used in vehicle detection include almost all ranges of electromagnetic and acoustic waves. The most popular detectors are electromagnetic loops installed in the pavement, and video detectors that use the visible range of electromagnetic waves are becoming popular.

Regardless of the detection technology used, the most popular technique is based on a detection zone, which is the spot on the pavement selected when setting a detector. Some detectors count only vehicles, while others measure the time vehicles are present in the detection zone. Most detectors can use more than one detection zone. The dimensions of detection zones vary from very small (microsensors) to large enough to span across several lanes and over a long distance. The detection zone dimensions depend on the type of detector and its purpose. If traffic measurements are the purpose, the detection zone should be small. A traditional detection zone is a 6-foot-long square or hexagon that covers a single traffic lane.

Let us first consider a detector that uses a single detection zone and is able to measure the time when vehicles are present in the detection zone. Figure 64.3 presents two vehicles passing the detection zone, represented by two trajectories: the front bumpers and the rear bumpers. The presented detector is capable of returning counts and vehicle presence information. Typical detectors calculate so-called *detector occupancy*, which is the percent of time when the detection zone was occupied by vehicles. If an observation period is 5 minutes, 20% detector occupancy means that for a total time of 1 minute during the observation period one or more vehicles were present in the detection zone. Typically, detectors return data in consecutive intervals. Volume V is calculated with Eq. (64.3). The following equations calculate space-mean speed S_L and density D from count N and detector occupancy B, obtained for an interval of length T:

$$S_L = \frac{100 \cdot N \cdot \left(l_v + l_d\right)}{T \cdot B} \tag{64.9}$$

$$D = \frac{B}{100 \cdot \left(\bar{l}_v + l_d\right)} \tag{64.10}$$

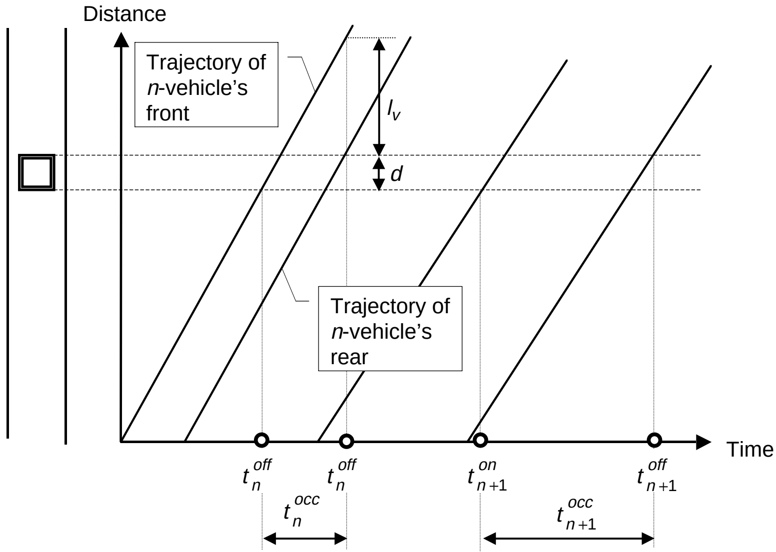

FIGURE 64.3 Measurements with a single detector.

where l_v stands for vehicle length and l_d stands for detection zone length. These lengths should be in miles and T should be in hours to obtain speed in miles/hour and density in vehicles/mile/lane. Detector occupancy B is expressed in percent. Equation (64.9) can be derived by starting with individual vehicle speeds $S_i = (l_{vi} + l_d)/t_i$, where l_{vi} is the length of vehicle i and t_i is the detector occupancy time caused by vehicle i. Equation (64.7), used together with the definition of detector occupancy, yields Eq. (64.9). Equation (64.10) is obtained from Eq. (64.9) by using the fundamental flow relationship shown in Eq. (64.6).

The main disadvantage of single detection zones is that the vehicle length is not measured and therefore has to be assumed. Detectors with a single detection zone are used mainly to count vehicles and to evaluate the congestion level through detector occupancy — a quantity directly related to traffic density, as seen in Eq. (64.10).

Detectors with two detection zones are much more reliable in measuring vehicle speeds. The known distance between the leading edges of two detection zones divided by the travel time between the two edges estimates the speed of a vehicle. The vehicle length is estimated from its speed and the time it occupies the detection zone. This information is used to classify vehicles.

Cumulative Counts

A new technique based on cumulative counts at spots was proposed for freeway traffic by Cassidy and Windover (1995), using count detectors located along a freeway section at a frequency of approximately 1/2 mile. Detectors are required in all traffic lanes, and vehicle counting starts as the same vehicle passes the consecutive spots. Cumulative counts over an extended period can be used to estimate the number of vehicles present between two spots (segment occupancy), travel times between two spots, vertical queues, and delays (see Fig. 64.4).

64.4 Relationships between Volume, Speed, and Density

Traffic flow characteristics — volume, speed, and density — do not vary independently from one another. The fundamental traffic flow equation ties these three characteristics together and allows for calculating the quantity of one of them when the other two are known. The fundamental equation was derived from the basic properties of traffic flows without any assumptions about driver behavior. An additional equation is needed in order to calculate two unknown flow characteristics from one known.

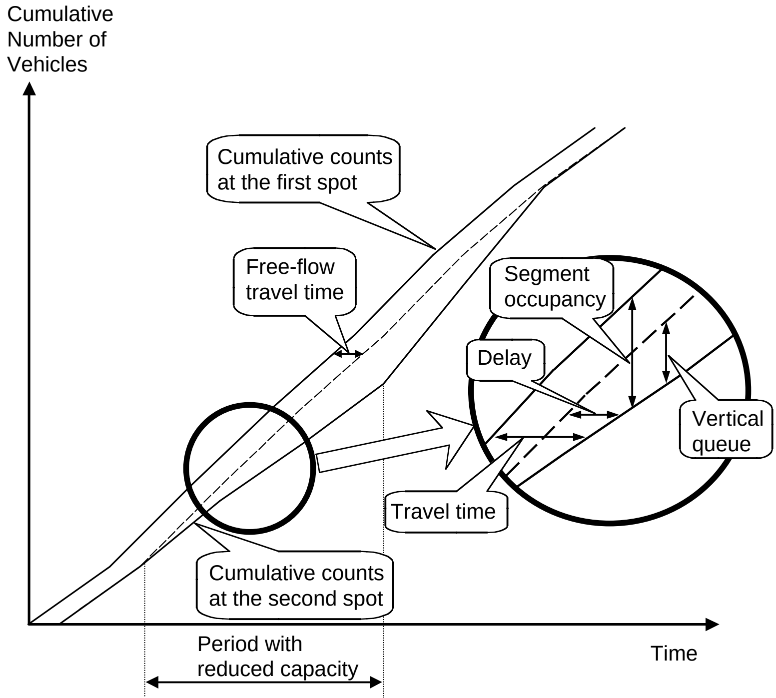

FIGURE 64.4 Measurements with cumulative counts.

Theories

Multiple theories furnish a second equation derived from assumptions about how drivers respond to traffic situations or by assuming analogies between traffic and flows (Gartner et al., 1995). For example, Pipes (1953) postulates a linear relationship between the speed and the distance between vehicles. If one assumes also that drivers are identical and consistent, that they do not pass other vehicles, and that particular traffic conditions exist long enough for drivers to adjust their speeds, the following speed–density relationship can be derived:

$$S = V_M \left(\frac{1}{D} - \frac{1}{D_J} \right) \tag{64.11}$$

The speed of vehicle flow S depends on maximum volume V_M, traffic density D, and maximum density D_J, called the jam density. Equation (64.11) inherits the weakness of the crude assumptions about driver behavior. The attempt to apply this equation to low-density conditions fails, since the speed prediction is unrealistically high.

Among other theories, one that is considered reasonable for all traffic conditions is called Greenshields' equation (Greenshields, 1935). It assumes that drivers adjust their speed to the speed of preceding vehicles, but unlike Eq. (64.11), this response weakens with the square distance between vehicles. The equation is

$$S = S_F \cdot \left(1 - \frac{D}{D_J} \right) \tag{64.12}$$

The theory says that vehicles move at their maximum speed, called free-flow speed S_F, when the highway is empty ($D = 0$), and the speed drops linearly with an increase in density and reaches zero at

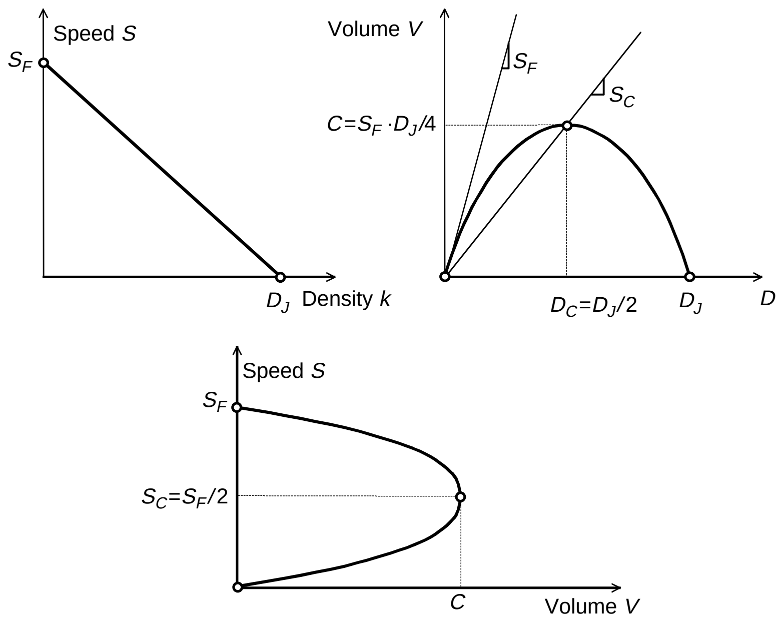

FIGURE 64.5 Greenshields' traffic flow relationships.

jam density D_J. The relationship between volume and density is a parabolic curve obtained after using Eq. (64.5) for a flow of vehicles with identical speeds:

$$V = D \cdot S_F \cdot \left(1 - \frac{D}{D_J}\right) \tag{64.13}$$

Figure 64.5 shows three relationships between volume, speed, and density derived from both the Greenshields' and fundamental traffic flow equations. These relationships indicate the existence of traffic conditions with maximum flow *C*. This value is considered the capacity of the segment, with traffic characterized by the curves. Also, free-flow and jam conditions are shown on these graphs.

Recent observations of freeway traffic and simplifications in traffic modeling have led to a piecewise linear relationship between volume and density (Daganzo, 1992; Newell, 1993): $V = \min\{S_F \cdot D, V_M \cdot (1 - D/D_J)\}$. Surprisingly, this relationship can be easily obtained from Eq. (64.11) by using free-flow speed S_F as an upper bound for speed *S* and then applying the fundamental relationship in Eq. (64.5).

Greenshields' equation and others are quite useful in understanding the effect of driver behavior on traffic conditions. One plausible result is that drivers tend to slow down when density increases; another is that capacity is reached at a particular density value. If density grows beyond the capacity, the loss in volume caused by the speed decrease exceeds the gain caused by the density increase, and the volume drops. Capacity is determined by the spacing drivers maintain between vehicles. Drivers select spacing perceived as safe for the speed that persists in the capacity conditions. The second half of the volume–density curve and the bottom half of the speed–volume curve describe traffic forced to move slower by some flow constraint located downstream. Drivers slow down accordingly and reduce their distances between vehicles. Jam density indicates the distance drivers prefer between vehicles when stopped in a queue. A comprehensive overview of flow relationships can be found in May (1990) and Gardner et al. (1995).

Field Observations

The presented traffic flow relationships were derived from assumptions that fall short of the true complexity of traffic. Drivers are not identical; they drive different vehicles, change lanes, and behave inconsistently.

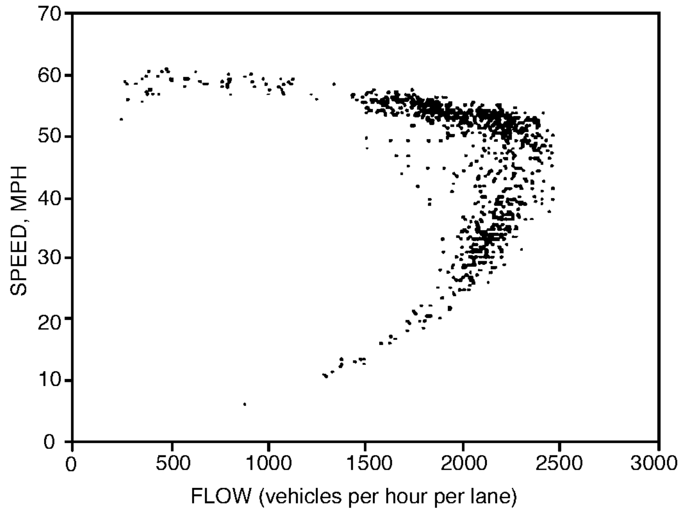

FIGURE 64.6 Example detector observations. (From Banks, J.H., *Transp. Res. Rec.*, 1225, 53, 1989.)

It should not be a surprise that the trends in field observations do not fully follow theories. Figure 64.6 presents example results of spot observations (Banks, 1989). One striking departure from the Greenshields' model is the small speed reduction for quite a wide range of volumes, which is caused mainly by the opportunity for passing on multilane highways. As long as passing is possible, drivers maintain their preferred speeds, and the average flow speeds do not indicate any effect of growing volume or density. Speed declines noticeably when traffic volume approaches capacity.

Other departures from the existing traffic flow theories are two values of capacity reported by Hall and Agyemang-Duah (1991). It seems that traffic moving smoothly for a while can reach quite high values, which indicates that drivers accept short distances between vehicles. On the other hand, when traffic discharges from a standing queue, its volume does not reach the capacity observed for moving flows.

As mentioned earlier, so-called forced traffic is observed where the density of traffic is larger than the density value in capacity conditions. Dense traffic is caused by a downstream obstruction, which is called *bottleneck,* the effects of which are apparent when traffic volumes without a bottleneck would be higher than the bottleneck capacity. Bottlenecks are caused by a reduced number of lanes, activities that distract drivers, or traffic controls such as stop signs and traffic signals.

Measurements representing capacity and near-capacity conditions are more scattered than those representing other conditions. These measurement points, lying mostly inside the expected speed–volume curve, represent transient conditions where traffic is changing from free-flowing to forced or from forced to free-flowing. Flows in transient conditions consist of drivers who rapidly change speeds to adjust to new conditions. Because the flow relationships discussed here represent conditions in which drivers have already reached their preferred speeds and spacing, transient conditions are not well described with these relationships.

The widely used *Highway Capacity Manual* (HCM) contains methods of predicting the capacity of highway segments and bottlenecks and their impacts on traffic conditions (Transportation Research Board, 2000). The chapter on freeway segments uses speed–volume relationships generalized from field observations (Fig. 64.7). These curves apply to specific conditions, called ideal. The HCM provides a means to convert the actual traffic volume in actual conditions to its equivalent value under ideal conditions. The speed–volume and fundamental relationships allow for the calculating of the speed and density from known traffic volume persisting on an analyzed freeway segment. It should be mentioned that the bottom part of the curve for congested and forced traffic is not shown, since the HCM focuses on conditions that are acceptable to motorists. Further details on highway capacity are presented in Section 64.6.

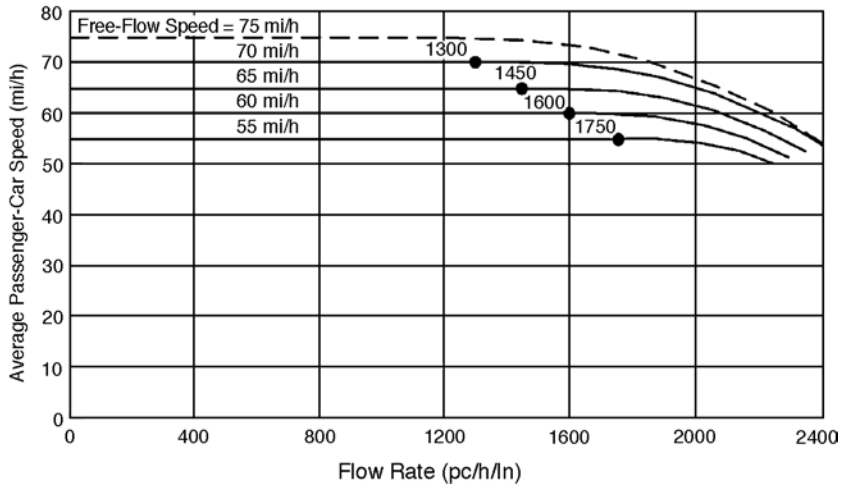

FIGURE 64.7 Speed–volume relationships for ideal conditions. (From Transportation Research Board, *Highway Capacity Manual*, Special Report 209, National Research Council, Washington, D.C., 2000.)

64.5 Queues and Delays at Bottlenecks

Points on the earlier introduced flow relationship curves represent traffic conditions ranging from free-flow to jam. The points represent a stable situation when drivers' speeds and distances have already been adjusted and these conditions are called steady state. The HCM methods for freeways and other types of road segments apply to steady-state traffic that persists during the busiest 15 minutes of design or rush hour. This is acceptable for uniform road sections without bottlenecks, since actual 15-minute traffic measurements indeed can be approximately steady. A single steady state cannot properly represent highway operations where traffic conditions change rapidly. Such changes are observed upstream of a highway bottleneck, where the congested traffic of vehicles in the queue meets the noncongested traffic of arriving vehicles.

Shock Waves

The boundary between congested and noncongested traffic where vehicles slow down rapidly is called a *shock wave*. A shock wave is also created by vehicles that accelerate rapidly after being released from a queue, for example, during a green signal. A shock wave can travel forward or backward. The end of a growing queue moves backward while the end of a dense column of vehicles behind a truck can move forward.

Speed w of a shock wave depends on the characteristics of traffic states that meet each other at the shock wave location:

$$w = \frac{V_1 - V_2}{D_1 - D_2} \tag{64.14}$$

A truck blocking traffic is the example presented in Fig. 64.8. The truck's speed is lower than the flow speed in capacity conditions and passing is not possible. The straight sections on the volume–density curve connect points represent distinct traffic states. The slopes of these sections equal the corresponding shock waves separating the two traffic states (Eq. (64.14)). The solid lines on the time-space diagram are the shock waves' trajectories separating different traffic states. The time-space areas bounded by the shock waves' trajectories are arriving traffic, behind-truck traffic, free-flow traffic in front of the truck, and

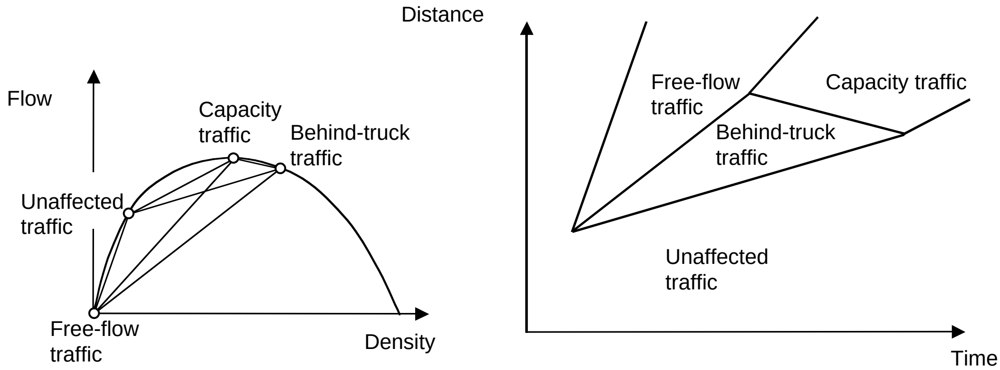

FIGURE 64.8 Shock waves caused by a slow vehicle.

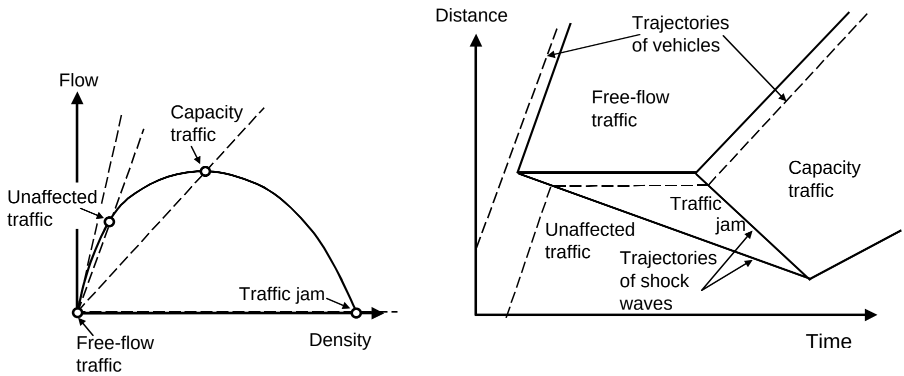

FIGURE 64.9 Shock waves caused by traffic interruption.

capacity conditions. The only backward shock wave is the shock wave of accelerations that separates the capacity conditions from the behind-truck conditions. The capacity conditions appear after the truck exits the road and when vehicles at the front of the dense column accelerate to reach the speed and the spacing between vehicles desired for the capacity conditions.

When the truck's speed is lower than the speed of unaffected vehicles and is higher than the capacity speed, vehicles decelerate when joining the traffic behind the truck and accelerate back to the original speed after being released in the front of the column. The wave of accelerations is forward. Capacity state does not occur since its appearance would require vehicles released from the front of the column to decelerate to reach the capacity speed, which is against the tendency of drivers to travel at their desired speeds if possible. More about the theory of shock waves can be found in Lighthill and Whitham (1955), Richards (1955), Gartner et al. (1995), and in most textbooks on traffic flow theory.

A slow-moving truck can be viewed as a moving bottleneck, while a jam caused by immobile obstructions, such as traffic signals or temporary road closures, is called a stationary bottleneck. Traffic directly upstream of an obstruction does not move and accumulates in a queue at jam density. When the obstruction ends, the traffic released in the front of the queue accelerates and reaches capacity state because the zero speed of traffic behind the obstruction is apparently lower than the capacity speed. Figure 64.9 presents all the traffic states and shock waves associated with a temporary obstruction.

The dimensions of the triangle with jam conditions can be calculated using Eq. (64.14) and other equations. The triangle's side parallel to the time axis is equal to the interruption time t_I, while the slope of the triangle's side representing the wave of decelerations (boundary between unaffected and jam traffic) is

$$w_{AJ} = \frac{-V}{D_J - D_A}$$

while the slope of the triangle's side representing the wave of accelerations (boundary between jam and capacity traffic) is

$$w_{JC} = \frac{-C}{D_J - D_C}$$

The two shock waves travel the same distance to meet at the point where the queue completely dissipates. This condition can be written as $w_{AJ}\ (t_I + t_D) = w_{JC} \cdot t_D$, where t_D is the time the queue needs to dissipate. This time can be calculated from the equation obtained by solving the above equation for t_D:

$$t_D = \frac{w_{AJ} \cdot t_I}{w_{JC} - w_{AJ}} \tag{64.15}$$

Distance l_Q, measured from the location of the interruption to the last vehicle stopped, is the distance traveled by the acceleration shock wave: $l_Q = t_D \cdot w_{JC}$. Let us use this equation and Eq. (64.15) to calculate l_Q.

$$l_Q = \frac{w_{AJ} \cdot w_{JC} \cdot t_I}{w_{JC} - w_{AJ}} \tag{64.16}$$

A negative value of distance means that the distance is measured backward. To avoid negative values, the absolute values of speeds can be used in Eq. (64.16) and in the equations that follow. The total number of vehicles stopped can be estimated as $N_S = l_Q \cdot D_J$, which reflects the fact that each affected vehicle stops once in the jam that persists along distance l_Q. Using this equation and Eq. (64.16) gives

$$N_S = \frac{w_{AJ} \cdot w_{JC} \cdot t_I \cdot D_J}{w_{JC} - w_{AJ}} \tag{64.17}$$

The longest stopping time equals the interruption time. Since the stopping time decreases linearly with time, the average stopping delay is $\bar{d} = 1/2\ t_I$. Since the total stopping time is the product of the average stopping time and the number of stopped vehicles, or $D = \bar{d} \cdot N_S$, the total delay is

$$D = \frac{w_{AJ} \cdot w_{JC} \cdot t_I^2 \cdot k_J}{2\left(w_{JC} - w_{AJ}\right)} \tag{64.18}$$

The concept of shock waves provides a convenient bridge between the theory of uninterrupted traffic flows and the theory of traffic operations at bottlenecks. Although rarely used in traffic engineering practice, the shock wave theory contributes to an understanding of how the relationships between traffic flow characteristics affect the behavior of traffic queues. The two main weaknesses of the shock wave theory are that (1) the calculations are cumbersome and, more importantly, (2) the random fluctuation of traffic is not addressed. The two approaches that follow, cumulative counts and equations based on the queuing theory, address the weaknesses of the shock wave theory.

Cumulative Counts

Consideration of the detailed behavior of a queue, as done in the shock wave theory, is not required to predict the delays caused by a bottleneck. A vehicle delay measured at a spot is nothing more than the time when a vehicle passes the spot minus the time when the vehicle would pass the spot if the bottleneck

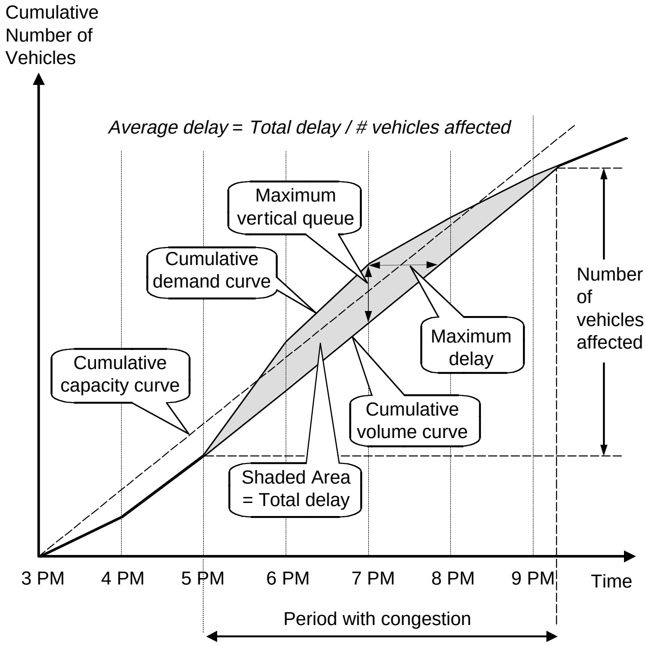

FIGURE 64.10 Predicting queues and delays with cumulative counts.

did not exist. This definition of delay gives a good approximation of the complete effect of the bottleneck if the spot mentioned in the definition is selected at the bottleneck or somewhere downstream of it.

The two passage times, affected and unaffected, can be easily determined by using two cumulative curves that represent *traffic demand* and bottleneck capacity. Traffic demand is the traffic volume that would be observed at the bottleneck location if the bottleneck did not exist. Figure 64.10 presents the two curves for an example bottleneck. The upper thin solid curve illustrates the total number of vehicles that would pass the spot at any time t if there were no bottleneck, while the thin dashed line represents the cumulative capacity. The slope of the first curve is the demand rate, while that of the second line is the capacity rate. At 5 p.m., the demand rate exceeds the capacity rate, which initiates congestion. The capacity line then shifts down to cross the demand line at 5 p.m. The congestion ends at 9:20 p.m., when the demand and capacity lines cross each other again. The actual flow rate passing the spot is the demand line between 3 and 5 p.m., the shifted capacity line between 5 and 9:20 p.m., and the demand line after 9:20 p.m.

The area between the demand line and the shifted capacity line provides information about the delay and the extent of congestion. The horizontal separation between the two curves for vehicle n is the delay of vehicle n as defined earlier in this section. The total area equals the total delay caused by the bottleneck. To obtain the average delay per affected vehicle, this value can be divided by the total number of vehicles affected.

The vertical separation at time t is simply the number of vehicles that would pass the spot by time t but could not due to the bottleneck. This number, sometimes called "vertical queue," is not the actual queue. The actual queue is made up of vehicles that would pass the spot if the bottleneck did not exist (vertical separation of cumulative curves) and of additional vehicles that have already joined the queue.

Cumulative curves simplify delay calculations but cannot provide the length of the congested section easily. The shock wave theory should be considered for this task. Newell (1993) has shown that the cumulative curve method and the shock wave theory, sometimes called the LWR theory (Lighthill and Whitham, 1955; Richards, 1956), are equivalent, since one can be derived from the other. Convenience and desired results determine which theory is used.

Queuing Theory Equations

The shock wave and cumulative counts theories are deterministic. They do not consider random fluctuation of vehicle volume and density, and both theories claim zero queues and delays as long as the traffic volume does not exceed the capacity. Actually, short-term traffic fluctuations may cause congestion conditions, particularly when traffic volume approaches capacity. The average delay for an extended period is not zero. Deterministic approximation is satisfactory and can be used for traffic volumes that are well below or well above bottleneck capacity.

The queuing theory deals with random queues and is useful in analyzing random traffic at highway bottlenecks. Many good introductions to this theory are available (for example, Cooper, 1981; Bunday, 1996). The queuing theory describes the behavior of queues of customers waiting for service in a server where customers spend some time, called service time. Service time can be fixed or it can vary randomly. Customers arrive at the server at a specified rate with a specified variability, and if the server is occupied at the time of arrival, a customer joins a queue that is governed by queue discipline and maximum length. The length of the queue and the time spent in the queue depend on the arrival and service rates and on the random variability of arrivals and service. Time spent in the queue is considered a waste of time or a delay, and the time spent in the system is the total time spent in the queue and in the server. The primary measures of queue performance are the likelihood that the server is busy, average queue size, average number of customers in the system, average time spent in the queue, and average time spent in the system. The classical queuing theory assumes that the arrival rate is lower than the service rate. The average queue can be converted to average delay using Little's formula:

$$Average\ Delay = Average\ Queue / Arrival\ Rate. \tag{64.19}$$

It is important to fully understand how the concepts used in the queuing theory translate to the concepts used to describe traffic operations at highway bottlenecks. Customers arriving at the server in the queuing theory correspond to unaffected traffic arriving at the bottleneck. A service time is a time headway between vehicles passing a bottleneck when a queue persists upstream of the bottleneck.

There is an important difference in defining "queue" in queuing theory and in highway operations theory. According to the queuing theory, a customer being served is not in a queue. An attempt to identify a server at a highway bottleneck can yield only one answer: a vehicle is served if it is immediately upstream of the bottleneck. Spending time in the first position before the bottleneck is as undesirable as spending time at any other position in the queue. Therefore, in order to calculate the average queue, a traffic engineer should use an equation that returns the average number of customers in the system. The same applies to calculating delays for a traffic engineer: the average delay is the same as the average time spent in the system for a queuing theorist. It is also important to mention that the so-called average queue is in fact an average vertical queue, as explained for cumulative counts. After all, the queuing theory was developed to model queues of telephone calls, which have no physical dimensions.

Bottleneck with Uninterrupted Traffic

Let us analyze a queue at a highway bottleneck where the width of lanes or the number of lanes is reduced. The arriving traffic is often random (Poisson) and time headways between consecutive vehicles passing the bottleneck during capacity conditions are relatively stable, at least compared to the time between arrivals. Thus, the arrival of vehicles is random, while the service is deterministic. If the entire cross section at the bottleneck is considered a single server, a suitable equation for this case to calculate average vehicle queue is

$$Q = \frac{2x - x^2}{2(1-x)} \tag{64.20}$$

where x is the V/C ratio. It should be remembered that Eq. (64.20) is used in the queuing theory to estimate the average number of customers in the system. Where volume equals 3240 vehicles/hour and capacity is 3600 vehicles/hour, the V/C ratio is $x = 3240/3600 = 0.900$, and the average vertical queue is estimated as

$$Q = \frac{2x - x^2}{2(1-x)} = \frac{2 \cdot 0.9 - 0.9^2}{2 \cdot (1-0.9)} = 4.95 \text{ vehicles}$$

Average delay d is calculated with Little's formula:

$$d = Q/V = 4.95/(3240/3600) = 5.5 \text{ sec}$$

The main weakness of the queuing theory is that it fails when traffic demand equals or exceeds capacity. This weakness has been overcome by developing time-dependent versions of queuing equations that allow the traffic demand to exceed the capacity for a finite time T (Catling, 1977). These equations return average delays and queues. A time-dependent version of Eq. (64.20) is

$$d = \frac{3600}{C} + 900T\left[x - 1 + \sqrt{(x-1)^2 + \frac{4 \cdot x}{CT}}\right] \tag{64.21}$$

where T is the time period to which the values of x and capacity C apply. Time T is expressed in hours and capacity C in vehicles/hour. Equation (64.21) returns 5.4 seconds for the previous example and for T set at 1 hour. For the same bottleneck, Equation (64.21) predicts a delay of 31.0 seconds if the demand equals the capacity for one hour ($x = 1$, $T = 1$ hour). The deterministic cumulative counts would predict zero delay. On the other hand, for x much higher than 1, the results produced by the time-dependent queuing equation and by the cumulative counts should be similar.

Unsignalized Intersections

A traffic flow approaching unsignalized intersections on a road with a stop sign must yield to traffic on the main road. This subordination reduces the capacity of the traffic flow considerably and creates a bottleneck with queues and delays. The service time varies considerably, since vehicles at the first position in the queue (server) have to wait for a sufficient gap between vehicles on the main road. Some vehicles may wait briefly, while others may wait much longer. Unlike the previous case of a reduced cross section, service at unsignalized intersections is random. Queues at unsignalized intersections are influenced by random arrivals and random service. An equation offered by the queuing theory applies only to volumes that do not exceed capacity. The HCM recommends a time-dependent version of the equation, which can deal with volumes higher than capacity:

$$d = \frac{3600}{C} + 900T\left[x - 1 + \sqrt{(x-1)^2 + \frac{8 \cdot x}{CT}}\right] + 5 \tag{64.22}$$

The term $3600/C$ is the average service time or the average time spent by vehicles at the first position in the queue. The second term is the average time spent in the queue at positions other than the first one. According to the HCM, the third term is the delay component associated with vehicle deceleration to stop on the approach and acceleration to return to the previous speed. Volume V, expressed in vehicles/hour, represents traffic during a period of analysis T, expressed in hours. Typically, the busiest 15-minute portion of the rush hour is used. However, if congestion is expected to start earlier and to last longer than the worst 15 minutes, then T should be the duration of congestion. Refer to Eq. (64.3), which converts a count in period T to the hourly rate, and check the section on cumulative counts to recall how the congested period can be determined.

Signalized Intersections

The most severe bottlenecks are observed at signalized intersections. It is not that traffic signals create them but rather that signals are used at busy intersections where elimination of the bottlenecks by building an interchange would be too expensive. Traffic signals interrupt flows, causing queues even when the arriving volume is well below capacity. The case of a single traffic interruption was analyzed earlier with the help of shock waves. The HCM recommends quite an elaborate time-dependent equation for signalized bottlenecks:

$$d = d_1 \cdot PF + d_2 + d_3 \tag{64.23}$$

$$d_1 = \frac{0.5 \cdot c \cdot (1 - g/c)^2}{1 - (g/c) \cdot [\min\{x, 1\}]} \tag{64.24}$$

$$d_2 = 900T \left\{ x - 1 + \sqrt{(x-1)^2 + \frac{8 \cdot k \cdot I \cdot x}{C \cdot T}} \right\} \tag{64.25}$$

where c = the average cycle length (seconds)
g = the average effective green signal (seconds)
x = the degree of saturation, $x = V/C$
T = the period of analysis (hours)
C = the capacity (vehicles/hour)
PF = the adjustment factor for signal progression
k = the incremental delay factor, 0.5 for signals that do not adjust to traffic
I = the filtering/metering adjustment factor, 1 in most cases

The first term, d_1, calculates delay if there were no randomness in traffic. This delay component, called *deterministic delay*, has been derived assuming the "bunched" and nonrandom service of vehicles during a green signal. It can be easily derived using the cumulative counts method. Adjustment factor *PF* incorporates the progression effect of upstream signals that influence time when the column of vehicles released from the upstream signals arrives at the considered bottleneck. Progression can increase or decrease delay at the bottleneck (Hillier and Rothery, 1967; Fambro et al., 1991). The second component, called *overflow delay*, d_2, is derived by modifying the queuing equation for random arrivals and deterministic service and by adding two adjustment factors, *k* and *I*, that incorporate the effect of signals actuation (responsive to traffic) and the reducing effect of upstream signals on traffic randomness. The last term, called *unmet demand delay*, d_3, includes the effect of the initial queue of vehicles present at the beginning of the period of analysis *T*.

64.6 Highway Capacity

Capacity is the maximum traffic volume a highway section can serve. It is a strong factor of traffic queues and delays at bottlenecks, and the equations for delays presented in the preceding sections include capacity.

The capacity of a highway lane is a direct result of the willingness of drivers to maintain a certain distance from the preceding vehicles given the prevailing speed along the highway. Apparently, the main factor of highway capacity is driver perception of risk associated with following another vehicle. In distracting conditions, such as work zones or adverse weather, drivers increase their distances between vehicles and even reduce their speeds. Capacity reduction results from this behavior (Krammes and Lopez, 1994; Ibrahim and Hall, 1994). Some bottlenecks along highway segments experience capacity reduction caused by local factors such as steep grades, narrow cross sections, or even disabled vehicles that do not block the traveled way but distract drivers.

Another bottleneck type, or spots with reduced capacity, is an intersection approach. The capacity of an approach that is controlled by stop signs at an intersection and where a crossing road has priority depends primarily on the quantity and size of intervehicle gaps on the main road. The perception of a safe gap by the drivers from the minor road to pass through the intersection is also important. The capacity of an approach controlled by traffic signals depends on how much green signal is given to the approach per hour and whether traffic can freely discharge from the queue during a green signal or must yield to other traffic. The three mentioned cases of highway sections, stop-controlled approaches to intersections, and signal-controlled approaches to intersections will be presented shortly. The HCM provides an extensive presentation of the methods applicable to these and other road facilities.

Freeway Sections

The HCM provides a method of estimating the capacity of freeway roadways between interchanges. Figure 64.7 presents the speed–volume curves determined from field measurements for ideal conditions: no heavy vehicles, level terrain, and drivers familiar with the roadway (Schoen et al., 1995). Volume is given per lane, and reasonable weather conditions are presumed by definition. These curves end at different capacity values for different free-flow speeds.

There are several situations in which the capacity for ideal conditions is not suitable. Drivers who are unfamiliar with the road drive more cautiously and reduce capacity. Heavy vehicles, particularly on steep grades, tend to move slower and observe greater distances from other vehicles. The capacity of the freeway roadway strongly depends on the number of lanes. The capacity C_0 for ideal conditions can be adjusted to actual conditions by multiplying it by the number of lanes n, the adjustment factor for heavy vehicles f_{HV}, and the adjustment factor for driver population (familiarity with road) f_p:

$$C = C_0 \; n \; f_{HV} \; f_p \tag{64.26}$$

The adjustment for heavy vehicles depends on the percent of heavy vehicles, hilliness of the terrain, and type of heavy vehicle (trucks, buses, recreational vehicles).

The HCM provides separate methods applicable to other types of road sections, including multilane highways and two-lane rural roads; different capacity factors are considered for these roads.

Unsignalized Intersections

Unsignalized intersections include two-way stop-controlled, all-way stop-controlled, and roundabouts. Only two-way stop-controlled intersections will be presented here to discuss specific human behavior, called *gap acceptance*, and its effect on the bottleneck's capacity.

The previous section on queues at unsignalized intersections points out that vehicles at the first position in the queue must wait for a sufficient gap between vehicles on the main road to cross or merge into the priority traffic stream. The number and size of gaps between priority vehicles depend on the volume of priority vehicles. The more vehicles that are on the major road, the less opportunity there is to cross the road. On the other hand, drivers of subordinate vehicles decide which gaps are sufficiently long, and this decision is a strong capacity factor. Research has indicated that different drivers accept different gaps, and even the same driver accepts various gaps in similar conditions (Kittelson and Vandehey, 1991). To simplify the theory, a minimum time gap acceptable to an average and consistent driver represents the decisions of the overall driver population. The minimum acceptable gap is called *critical gap* t_g. When the critical gap is shorter, more vehicles can enter an intersection per hour. An additional factor that reduces the bottleneck capacity is the time between departures of consecutive vehicles that utilize a single long gap. Called *follow-up time* t_f, it represents the time a vehicle uses to move from second position to first in the queue, to make sure that the way is clear, and to pass the stop bar. The HCM uses the following equation to calculate the capacity of a single queue of vehicles crossing a priority single stream:

$$C = V_c \cdot \frac{\exp\left(-V_c \cdot t_g / 3600\right)}{1 - \exp\left(-V_c \cdot t_f / 3600\right)} \tag{64.27}$$

where V_c is the total volume of priority vehicles, called *conflicting volume* and expressed in vehicles/hour; t_g is the critical gap in seconds; and t_f is the follow-up time in seconds. Critical gap t_g depends on the type of maneuver performed by subordinate vehicles, the number of lanes on the main road, the traffic speed on the main road, the percent of heavy vehicles, the approach grade, and other factors. In addition, follow-up time depends on the percent of heavy vehicles. The details of calculating the critical gaps and the follow-up times can be found in the HCM.

A driver on a minor road can cross a main road only if the gap between arriving priority vehicles is sufficiently long and there are no queues of vehicles turning left from the main road. Values returned by Eq. (64.27) reflect the availability of acceptable gaps in the arriving priority traffic, but the equation does not include the blocking effect by queues formed by priority vehicles. The values obtained from Eq. (64.27) are multiplied by the portion of time when such queues are not present. Calculations for all the minor movements at an intersection are performed in proper order, starting with the left turns on a main road and ending with the left turns from a minor road. Additional factors, such as close signalized intersections, short lanes, and a median on a major road, are covered by the HCM.

Signalized Intersections

Let us consider the simplest case of through vehicles that use a group of traffic lanes controlled by traffic signals of known and preset lengths. The maximum through flow that can pass the traffic signal is called *lane group capacity C*. Vehicles arriving during a red signal form a queue. If the queue is sufficiently long, it discharges during green and yellow signals at a high rate, somewhat lower than the maximum flow rate measured in uninterrupted streams. The queue discharge rate will be called *saturation flow rate S*. The lane group capacity would equal the saturation flow rate ($C = S$) if a green signal were displayed for an entire hour. Since a typical green signal is much shorter but displayed multiple times during 1 hour, the lane group capacity is a portion of the saturation flow rate:

$$C = S \cdot \frac{g}{c} \tag{64.28}$$

where g is the *effective green signal* and c is the *signal cycle* — time measured between the beginnings of two consecutive green signals for the same lane group. An effective green signal g is equal to the displayed green plus yellow signals, reduced by the lost times during green and yellow signals. A beginning portion of green is lost because of driver reaction time and vehicle inertia. A portion of yellow is lost because some drivers seeing the yellow signal decide to stop their vehicles even when some of them could pass the stop bar before the red signal starts. Since the total lost time during green and yellow signals approximates the yellow signal length, in many cases the effective green time equals the displayed green time.

As indicated by Eq. (64.28), the capacity of a signalized lane group depends on the saturation flow, the effective green signal, and the signal length. Despite the simplicity of Eq. (64.28), the complexity of the calculations is considerable. The saturation flow rate, even for through vehicles, depends on many factors such as the number of lanes, presence of heavy vehicles, lane width, roadway grade, bus stops, parking activities, and location of the intersection in the town. The presence of turning vehicles on the lane group complicates the situation tremendously. The blocking effect of left turns and right turns depends on the volume of these vehicles and on their interactions with pedestrians and other vehicles. The saturation flow rate for a lane group becomes dependent on other streams and even on traffic signals. The HCM provides a special procedure to deal with the effect of left turns.

TABLE 64.2 Service Measures Recommended by the HCM

Highway Facility	Service Measure
Freeway basic segment	Traffic density
Freeway–ramp intersection	Traffic density
Weaving section	Travel speed
Nonfreeway multilane highway	Traffic density
Two-lane highway	Travel speed, percent time spent following
Urban street	Travel speed
Two-way stop-controlled intersection	Average delay
All-way stop-controlled intersection	Average delay
Signalized intersection	Average delay

Source: Transportation Research Board, *Highway Capacity Manual*, Special Report 209, National Research Council, Washington, D.C., 2000.

The lengths of effective green signals are relatively easy to determine for pretimed traffic signals because the displayed signals are fixed and often given. The situation becomes much more difficult when traffic signals are actuated and follow random fluctuations of traffic. They depend on varying traffic, the type of signal controller, and the signal settings (minimum and maximum greens and fixed change periods between consecutive greens). The HCM recommends using average signal lengths obtained by measurement if field observations are possible. Otherwise, analytical procedures must be used. An analysis of green signal demand by lane groups indicates the critical lane groups that determine the average green signal and cycle lengths. These values are then used in Eq. (64.28) to calculate the capacity of lane groups.

64.7 Traffic Quality

Highways should be designed and traffic should be managed and controlled to meet motorists' expectations, namely, to travel sufficiently fast between bottlenecks and to experience reasonably short queues and delays at bottlenecks. To fulfill these expectations, traffic engineers need a method of evaluating present and future highway operations that returns results consistent with this perception. To accomplish such consistency, traffic evaluation criteria must measure the performance of highway operations that are important to motorists. These measures are called *service measures*; Table 64.2 presents service measures recommended by the HCM for selected highway facilities. It should be emphasized that the HCM focuses on traffic conditions from free-flow conditions up to capacity conditions. Congested traffic conditions resulting from the demand exceeding the capacity for a considerable time are unacceptable to motorists and are not given much attention in the HCM.

The traffic quality on a freeway segment is evaluated based on traffic density, while traffic at signalized and unsignalized intersections is evaluated based on average delays. Indeed, the density of below-capacity traffic determines the freedom of drivers to select their own speeds and to change lanes to pass other vehicles. The traffic between free-flow and capacity conditions has been divided into five *levels of quality* (LOS) with assigned letters *A* for near free-flow conditions and *E* for near-capacity conditions. A single level *F* was reserved for congested and unstable traffic. The threshold values of density for the six levels of quality are given in Table 64.3. These values are determined for the ideal conditions, as described in the capacity section. The HCM provides a method of converting the actual volume into the equivalent volume V_E for ideal conditions:

$$V_E = \frac{V}{n \cdot PHF \cdot f_{HV} \cdot f_p} \tag{64.29}$$

where n is the number of lanes, PHF converts the hourly volume V into the flow rate that represents the busiest 15 minutes during design or rush hour, f_{HV} is an adjustment factor for heavy vehicles, and f_p is

TABLE 64.3 Level of Service Criteria for Freeway Basic Segments

Level of Service	Density Range (pc/mi./lane)
A	0.0–12
B	>12–18
C	>18–26
D	>26–35
E	>35–45
F	>45

Source: Transportation Research Board, *Highway Capacity Manual*, Special Report 209, National Research Council, Washington, D.C., 2000.

TABLE 64.4 Level of Service Criteria for Intersections

	Delay Range(s)	
Level of Service	Unsignalized Intersections	Signalized Intersections
A	0–10	0–10
B	>10–15	>10–20
C	>15–25	>20–35
D	>25–35	>35–55
E	>35–50	>55–80
F	>50	>80

Source: Transportation Research Board, *Highway Capacity Manual*, Special Report 209, National Research Council, Washington, D.C., 2000.

an adjustment factor for driver population. A measured or predicted free-flow speed is used to select or interpolate a speed–volume curve in Fig. 64.7. The actual speed for ideal conditions is determined from the obtained speed–volume curve. Then Eq. (64.5) is used to calculate the density. This value, when compared to the threshold values in Table 64.3, reveals the level of service for the freeway segment.

Determination of the quality of service at intersections is straightforward. The average delay can be measured or, if direct observations are not possible, capacity must be calculated as described in the capacity section and the average delay estimated or predicted using proper equations from the bottleneck section. The last step is to compare the measured or calculated average delays with the critical values shown in Table 64.4.

The general procedure for any highway facility can be summarized through the following procedure:

1. Estimate the service measure through field observations and go to step 3.
2. If observations are not possible, then:
   A. Collect data needed in calculations.
   B. Calculate the capacity.
   C. Calculate the service measure.
3. Determine LOS.

National and local design and traffic control standards specify target LOS for various facilities and situations that are believed to reflect motorists' expectations. A set of target LOS values for a freeway segment is given in Table 64.5 (American Association of State Highway and Transportation Officials, 2001). These target values indicate the perception that motorists expect much better conditions on arterial highways in an easy terrain than on nonarterial roads in a difficult terrain.

TABLE 64.5 Design Levels of Service Recommended by AASHTO

	Area and Terrain Type			
Functional Class	Rural Level	Rural Rolling	Rural Mountainous	Urban and Suburban
Freeway	B	B	C	C
Arterial	B	B	C	C
Collector	C	C	D	D
Local	D	D	D	D

Source: American Association of State Highway and Transportation Officials, *A Policy on Geometric Design of Highways and Streets,* 4th ed., Washington, D.C., 2001.

64.8 Traffic Control

Traffic signs and signals, if properly used, increase highway operation effectiveness and safety. The *Manual of Uniform Traffic Control Devices* (MUTCD) specifies standards for the design and use of signs and signals between intersections and at intersections (Federal Highway Administration, 2001). Additional guidance is provided in the *Traffic Control Devices Handbook* (Federal Highway Administration, 1983). The first decision necessary regarding traffic control for an intersection is the type of control desired. Several different types of control are available, including no control, all-way stop control, two-way stop control, circular traffic (roundabouts), and traffic signals. The MUTCD and supplementary state documents provide guidance and warrants for using particular types of control. Detailed engineering studies are often required to decide whether stop control or signal control is warranted. Among the available types of control for intersections, traffic signal control has the strongest impact on traffic operations. Installation of traffic signals is warranted if traffic delays, traffic costs, or crash hazards can be reduced. Other considerations, such as uniformity of control along routes, are also evaluated.

Installation of new signals and modernizing existing ones involve traffic signal design and setting. Signal control can be limited to a single intersection or it can include coordination of a sequence of signals along an arterial, even in a street network. In the simplest signal controller, signals change according to a preset program designed in advance, based on historical traffic data. This type of controller cannot accommodate traffic signals to new traffic patterns and to short-term traffic fluctuations automatically. Signal setting requires periodic updates based on newer data collected by traffic engineers. Computer methods of traffic signal optimization are available to traffic engineers.

Modern traffic controllers can accommodate signals to traffic changes. To update signals to the current demand for green signal, vehicles are detected by local detectors on approaches to intersections or by system detectors at strategic spots between intersections. Local detection is used by signal controllers at individual intersections, while system detection is sent to a center for strategic control decisions. Detection techniques and technologies are described in the measuring techniques section; inductive loop detectors (Kell et al., 1990) and video detection are those most frequently used for signal actuation.

Vehicle detection with local detectors extends the current green signal or places a call for a new green signal if the red signal is displayed. This local signal adjustment may be available to all lane groups at the intersection (full actuation) or only to selected ones (semiactuation). If signal control is coordinated between intersections, then local signal adjustment is still allowed but must be limited so that the coordination of green signals along coordinated roads is preserved. All coordinated intersections operate on a common background cycle. Actuated and coordinated signals form a complex system that needs to be properly set. Arterial and area-wide traffic control systems have a large number of parameters, including background cycle, offsets, green minimums, unit green extensions, green maximums, permission periods, force-off points, and many others (Kell and Fullerton, 1991).

Some of the optimization methods developed for pretimed signals can translate pretimed solutions into coordinated actuated signals settings, but there are serious doubts about the quality of these solutions when applied to traffic control system actuation. The main source of concern is that rather simple traffic models do not properly reflect dynamic and random traffic processes. Traffic stream interactions with signal controllers through the vehicle detection component cannot even be modeled in some of these methods. On the other hand, there are microsimulation computer packages capable of simulating the complex traffic processes and interactions with the control component, but they do not include a signal optimizer. These methods are useful for testing solutions but not for obtaining them. A combined approach is recommended where an optimization method is utilized first to generate an approximate solution, which is then tested using simulation and subsequently fine-tuned if needed. Commercial computer packages that integrate these two steps are available, whereby a user introduces the needed data once and then runs the optimizer. The solution with the input data is then transferred seamlessly from the optimizer to the microsimulation component for testing and further tuning.

References

American Association of State Highway and Transportation Officials, *A Policy on Geometric Design of Highways and Streets,* 4th ed., Washington, D.C., 2001.

Banks, J.H., Freeway speed-flow-concentration relationships: more evidence and interpretations, *Transp. Res. Rec.*, 1225, 53, 1989.

Bunday, B.D., *An Introduction to Queueing Theory,* John Wiley & Sons, New York, 1996.

Cassidy, M.J. and Windover, J., Methodology for assessing dynamics of freeway flow, *Transp. Res. Rec.,* 1484, 73, 1995.

Catling, I., A time-dependent approach to junction delays, *Traffic Eng. Control*, 11, 520, 1977.

Cooper, R.B., *Introduction to Queueing Theory*, Books on Demand, 1981.

Daganzo, C., The Cell Transmission Model: Part I: A Simple Dynamic Representation of Highway Traffic, Report UCB-ITS-PRR-93-7, Institute of Transportation Studies, University of California at Berkeley, 1992.

Fambro, D.B., Chang, E.C.P., and Messer, C.J., Effects of the Quality of Traffic Signal Progression on Delay, NCHRP Report 339, Transportation Research Board, National Research Council, Washington, D.C., 1991.

Federal Highway Administration, *Traffic Control Devices Handbook,* U.S. Department of Transportation, Washington, D.C., 1983.

Federal Highway Administration, *Manual on Uniform Traffic Control Devices: Millennium Edition,* U.S. Department of Transportation, Washington, D.C., 2001, http://mutcd.fhwa.dot.gov/.

Gartner, N.H., Messer, C.J., and Rathi, A.K., Eds., Monograph on Traffic Flow Theory, TRB Special Report, Transportation Research Board, National Research Council, Washington, D.C., 1995, http://www-cta.ornl.gov/cta/research/trb/tft.html.

Greenshields, B.D., A study of traffic capacity, *Highway Res. Board Proc.*, 14, 1935.

Hall, F.L. and Agyemang-Duah, K., Freeway capacity drop and the definition of capacity, *Transp. Res. Rec.*, 1320, 91, 1991.

Hillier, J.A. and Rothery, R., The synchronization of traffic signals for minimum delays, *Transp. Sci.*, 1, 81, 1967.

Ibrahim, A.T. and Hall, F.L., Effect of adverse weather conditions on speed-flow-occupancy relationships, *Transp. Res. Rec.*, 1457, 184, 1994.

Kell, J.H. and Fullerton, I.J., *Manual of Traffic Signal Design,* Institute of Transportation Engineers, Prentice Hall, Englewood Cliffs, NJ, 1991.

Kell, J.H., Fullerton, I.J., and Mills, M.K., *Traffic Detector Handbook*, IP-90-002, Federal Highway Administration, Washington, D.C., 1990.

Kittelson, W.K. and Vandehey, M.A., Delay effects on driver gap acceptance characteristics at two-way stop-controlled intersections, *Transp. Res. Rec.,* 1320, 154, 1991.

Krammes, R.A. and Lopez, G.O., Updated capacity values for short-term freeway work zone lane closures, *Transp. Res. Rec.,* 1442, 49, 1994.

Lighthill, M.J. and Whitham, G.B., On kinematic waves, I: Flood movement in long rivers, II: A theory of traffic flow on long crowded roads, *Proc. Royal Soc.*, London, 1955.

May, A.D., Jr., *Traffic Flow Fundamentals*, Prentice Hall, Englewood Cliffs, NJ, 1990.

McShane, W.R., Roess, R.P., and Prassar, E.S., *Traffic Engineering*, 2nd ed., Prentice Hall, Upper Saddle River, NJ, 1998.

Newell, G.F., A simplified theory of kinematic waves in highway traffic, I. General theory, II. Queuing at freeway bottlenecks, III. Multi-dimensional flows, *Transp. Res.,* 27B, 281, 1993.

Pipes, L.A., An operational analysis of traffic dynamics, *J. Appl. Phys.*, 24, 274, 1953.

Richards, P.I., Shockwaves on the highway, *Operations Res.*, 4, 42, 1956.

Robertson, H.D., Hummer, J.E., and Nelson, D.C., *Manual of Traffic Engineering Studies,* Institute of Transportation Engineers, Prentice Hall, Englewood Cliffs, NJ, 1994.

Schoen, J. et al., Speed-Flow Relationships for Basic Freeway Sections, Final Report, NCHRP Project 3-45, JHK & Associates, Tucson, AZ, 1995.

Transportation Research Board, *Highway Capacity Manual*, Special Report 209, National Research Council, Washington, D.C., 2000.