BASIC PRINCIPLES OF INTERSECTION SIGNALIZATION

Introduction

In previous lecture, various options for intersection control were presented and discussed. Warrants for implementation of traffic control signals at an intersection, presented in the Manual on Uniform Traffic Control Devices, provide general and specific criteria for selection of an appropriate form of intersection control, t many intersections, .the combination of traffic volumes, potential conflicts, overall safety of operation, efficiency of operation, and diver convenience lead to a decision to install traffic control signals. The operation of signalized intersections is often complex, involving competing vehicular and pedestrian movements. Appropriate methodologies for design and timing of signals and for the operational analysis of signalized intersections require the behavior of divers and pedestrians at a signalized intersection to be modeled in a form that can be easily manipulated and optimized. This lecture discusses some of the fundamental operational characteristics at a signalized intersection and the ways in which they may be effectively modeled. (In next lecture, these principles are applied to a signalized intersection design and timing process for pretimed signals. In next lectures they are augmented and combined into overall models of signalized intersection operations. The particular model presented in Chapter 24 is that of the Highway Capacity Manual. [This chapter focuses on four critical aspects of signalized intersection operation:

1. Discharge headways, saturation flow rates, and lost times
2. Allocation of time and the critical-lane concept
3. The concept of let-turn equivalency
4. Delay as a measure of service quality

Other aspect of signalized intersection operation are also important, and the Highway Capacity Manual analysis model addresses many of them. These four, however, are central to understanding traffic behavioral signalized intersections and are highlighted here.

Terms and Definitions

Traffic signals are complex devices that can operate in a variety of deferent modes. A number of key terms and definitions should be understood before pursuing a more substantive discussion.

Components of a Signal Cycle

The following terms describe portions and sub portions of a signal cycle. The most fundamental unit in signal design and timing is the cycle, as defined here.

1. Cycle. A signal cycle is one complete rotation through all of the indications provided. In general, every legal vehicular movement receives a “green” indication during each cycle, although there are some exceptions to this rule.

2. Cycle length. The cycle length is the time (in seconds) that it takes to complete one full cycle of indications. It is given the symbol "C."
3. Interval. The interval is a period of time during which no signal indication changes. It is the smallest unit of time described within a signal cycle. There are several types of intervals within a signal cycle:
   - Change interval. The change interval is the “yellow" indication for a given movement It is part of the transition from " green " to "red," in which movements about to lose " green " are given a " yellow" signal while all other movements have a " red" signal. It is timed to allow a vehicle that cannot safely stop when the “green “is withdrawn to enter the intersection legally. The change interval is given the symbol “y_i" for movement(s) i.
   - Clearance interval. The clearance interval is also part of the transition from “green “to "red" for a given set of movements. During the clearance interval, all movements have a "red" signal. It is timed to allow a vehicle that legally enters the intersection on 'yellow" to safely cross the intersection before conflicting flows are released. The clearance interval is given the symbol “ar_i” (for " all red”) for movement(s) i.
   - Green interval. Each movement has one green"' interval during the signal cycle. During a green interval, the movements permitted have a "green" light while all other movements have a "red" light. The green interval is given the symbol "G" for movement(s) i.
   - Red interval. Each movement has a red interval during the signal cycle. All movements not permitted have a "red" light while those permitted to move have a "green" light In general, the red interval overlaps the green, yellow, and all red intervals for all other movements in the intersection. The red interval is given the symbol “R_i" for movement(s) i. Note that for a given movement or set of movements, the “red" signal is present during both the clearance (all red) and red intervals.

4. Phase. A signal phase consists of a green interval plus the change and clearance intervals that follow it. It is a set of intervals that allows a designated movement or set of movements to flow and to be safely halted before release of a conflicting set of movements.

Types of Signal Operation

The traffic signals at an individual intersection can operate on a pretimed basis or may be partially or fully actuated by arriving vehicles or pedestrians sensed by detectors.

1. Pretimed operation. In pretimed operation, the cycle length, phase sequence, and timing of each interval ar constant. Each cycle of the signal follows the same predetermined plan. Modem signal controllers allow different pretimed settings to be established. An internal clock is used to activate the appropriate timing for each defined time period. In such cases, it is typical to have at least an am peak, a pm peak, and an off-peak signal timing.
2. Semi-actuated operation. In semi-actuated operation, detectors are placed on the minor approach (es) to the intersection; there are no detectors on the major street. The light is green for the major street at all times except when a "call" or actuation is noted on one of the minor approaches. Then, subject to limitations such as a minimum major-street green. The green is transferred to the minor street. The green returns to the major street when the maximum minor-street green is reached or when the detector senses there is no further

demand on the minor street Semi-actuated operation is often used where the primary reason for signalization is "interruption of continuous traffic, " as discussed in previous lecture.

3. Full actuated operation. In full actuated operation, every lane of every approach must be monitored by a detector. Green time is allocated in accordance information from detectors and programmed "rules” established in the controller for capturing and retaining the green. In full actuated operation, the cycle length, sequence of phases, and green time split may vary from cycle to cycle. Next lectures presents more detailed descriptions of actuated signal operation along with a methodology for timing such signals.

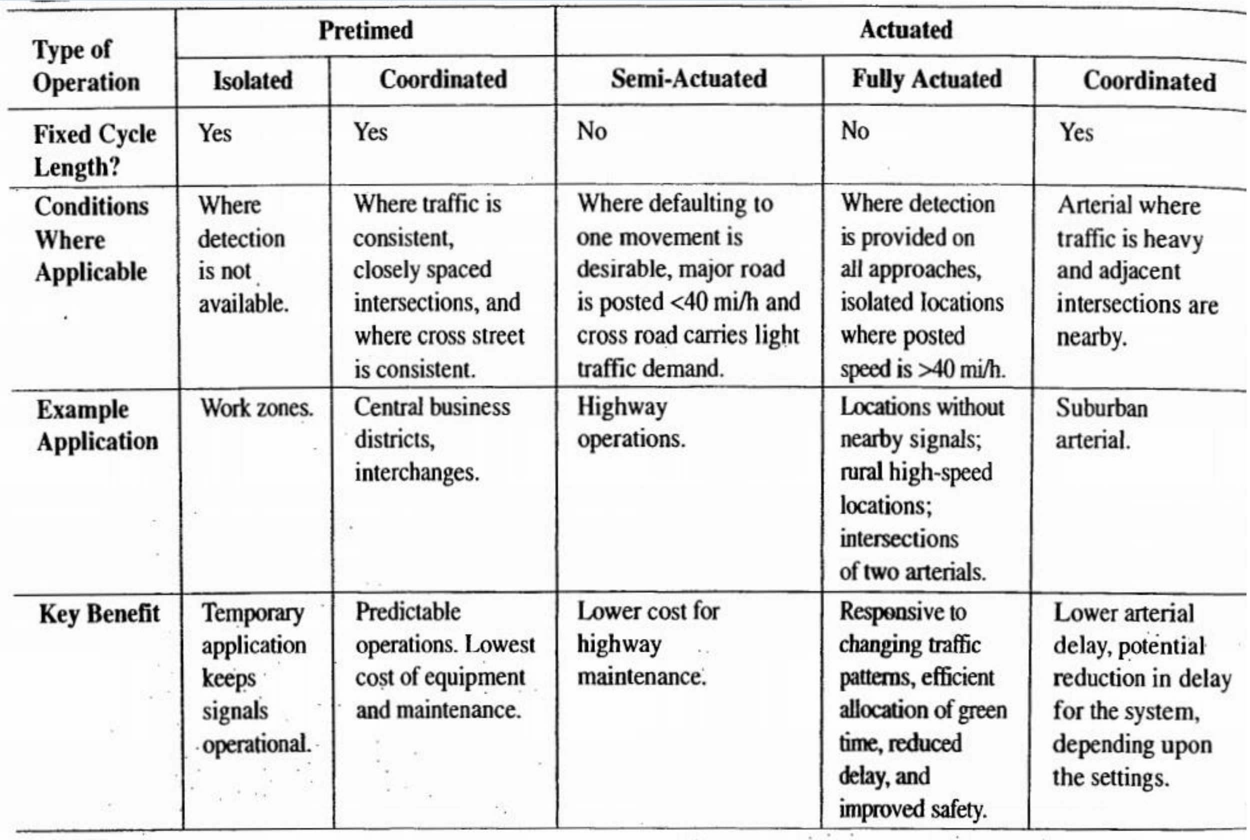
In most urban and suburban settings, signalized intersections along arterials and in arterial networks are close enough to have a significant impact on adjacent signalized intersection operations. In such cases, it is common to coordinate signals into a signal system. When coordinated, such systems attempt to keep vehicles moving through sequences of individual signalized intersections without stopping for as long as possible. This is done by controlling the "offsets" between adjacent green signals; that is, the green at a downstream signal initiates "x" seconds after its immediate upstream neighbor. Coordinated signal systems must operate on a common cycle length because of sets cannot be maintained from cycle to cycle if cycle lengths vary at each intersection. Coordination is provided using a variety of technologies:

1. Master controllers. A "master controller" provides a linkage between a limited set of signals. Most such controllers can connect from 20 to 30 signals along an arterial or in a network. The master controller provides fixed settings for each offset between connected signals. Settings can be changed for defined periods of the day.
2. Computer control. In a computer-controlled system, the computer acts as a “supersized " master controller, coordinating the timings of a large number (hundreds) of signals. The computer selects or calculates an optimal coordination plan based on input from detectors placed throughout the system. In general, such selections are made only once in advance of an am or PM peak period. The nature of a system transition from one timing plan to another is sufficiently disruptive to be avoided during peak-demand periods in a traditional system. Individual signals in a computer-controlled system generally operate in the pretimed mode.
3. Adaptive traffic control systems (ATCS). Since the early 1990s, there has been rapid development and implementation of "adaptive" traffic control systems. In such systems, both individual intersection signal timings and offsets are continually modified in real time based on advanced detection system inputs. In many cases, such systems use actuated controllers at individual intersections. Even though the system sill requires a fixed cycle length (which can be changed periodically based on detector input), the allocation of green within a fixed cycle length has been found to be useful in reducing delay and travel times. A critical part of adaptive traffic control systems is the underlying logic of software used to monitor the system and continually update timing patterns. A number of software systems are in use, and the list of products is increasing each year. Some of the more popular systems (in 2009) include SCOOT (Split Cycle Offset Optimization Technique), SCATS (Sydney Coordinated Adaptive Traffic System), RHODES (Real-Time Hierarchical Optimized Distributed Effective System), OPAC (Optimization Policies for Adaptive Control), and ACS-Lite (Adaptive Control System-Lite). In addition to the standard features of signal coordination, such systems usually also incorporate other

features, such as bus priority, emergency vehicle priority, traffic gating, and incident detection.

Table 1 summarizes the various types of individual signal controllers with key characteristics and guidelines on their most common uses. Dramatic changes have occurred in the use of traffic signal control technology over the past two decades. Before 1990, all coordinated traffic signal systems on arterials and in networks used pretimed signal controllers exclusively. Today, actuated controllers are regularly coordinated, although, as shown in Table 20.1, they lose one of their principal variable features: cycle length. To coordinate signals, cycle lengths must be common during any given time period, so that the offset between the initiation of green at an upstream intersection and the adjacent downstream intersection is constant for every cycle. Pretimed signals, because they are the cheapest to implement and maintain, are still a popular choice where demands are relatively constant throughout major periods of the day. Where demand levels (and relative demands for various movements) vary significantly during all times of the day, actuated signals are the most likely choice for use. Even when coordinated and using a constant cycle length, the allocation of green times among the defined phases can significantly reduce delay.

Table 1. Signal Controllers and Types of Intersection Control.



Type of Operation	Pretimed		Actuated		
	Isolated	Coordinated	Semi-Actuated	Fully Actuated	Coordinated
Fixed Cycle Length?	Yes	Yes	No	No	Yes
Conditions Where Applicable	Where detection is not available.	Where traffic is consistent, closely spaced intersections, and where cross street is consistent.	Where defaulting to one movement is desirable, major road is posted <40 mi/h and cross road carries light traffic demand.	Where detection is provided on all approaches, isolated locations where posted speed is >40 mi/h.	Arterial where traffic is heavy and adjacent intersections are nearby.
Example Application	Work zones.	Central business districts, interchanges.	Highway operations.	Locations without nearby signals; rural high-speed locations; intersections of two arterials.	Suburban arterial.
Key Benefit	Temporary application keeps signals operational.	Predictable operations. Lowest cost of equipment and maintenance.	Lower cost for highway maintenance.	Responsive to changing traffic patterns, efficient allocation of green time, reduced delay, and improved safety.	Lower arterial delay, potential reduction in delay for the system, depending upon the settings.

(*Source:* Koonce, P., et al., *Traffic Signal Timing Manual,* Final Report, FHWA Contract No. DTFH61-98-C-00075, Kittelson and Associates Inc, Portland, OR, June 2008, Table 5-1, p. 5-3.)

Treatment of Left Turns (and Right Turns)

The modeling of signalized intersection operation would be straightforward if let turns did not exist. Left turns at a signalized intersection can be handled in one three ways:

1. Permitted let turns. A "permitted" left turn movement is one that is made across an opposing flow of vehicles. The diver is permitted to cross through the opposing low but must select an appropriate gap in the opposing traffic stream through which to turn. This is the most common form of left-turn phasing at signalized intersections, used where let-tum volumes are reasonable and where gaps.in the opposing flow are adequate to accommodate let turns safely.
2. Protected left turn. A “protected “left turn movement is made without an opposing vehicular flow. The signal plan protects let-tuning vehicles by stopping the opposing through movement. This requires that the let turns and the opposing through flow be accommodated in separate signal phases and leads to multiphase (more than two) signalization. In some cases, let nuns are “protected “by geometry or regulation. Let turns from the stem of a T-intersection, for example, face no opposing low because there is no opposing approach to the intersection. Let turns from a one-way street similarly do not face an opposing flow.
3. Compound let turns. More complicated signal timing can be designed in which left turns are protected for a portion of the signal cycle and are permitted in another portion of the cycle. Protected and permitted portions of the cycle can be provided in any order. Such phasing is also referred to as protected plus permitted or permitted plus protected, depending on the order of sequence.

The permitted let turn movement is very complex. It involves the conflict between a left turn and an opposing through movement. The operation is affected by the left-tum flow rate and the opposing flow rate, the number of opposing lanes, whether left turns flow from an exclusive left-tum lane or rom a shared lane, and the details of the signal timing. Modeling the interaction among these elements is a complicated process, one that often involves iterative elements.

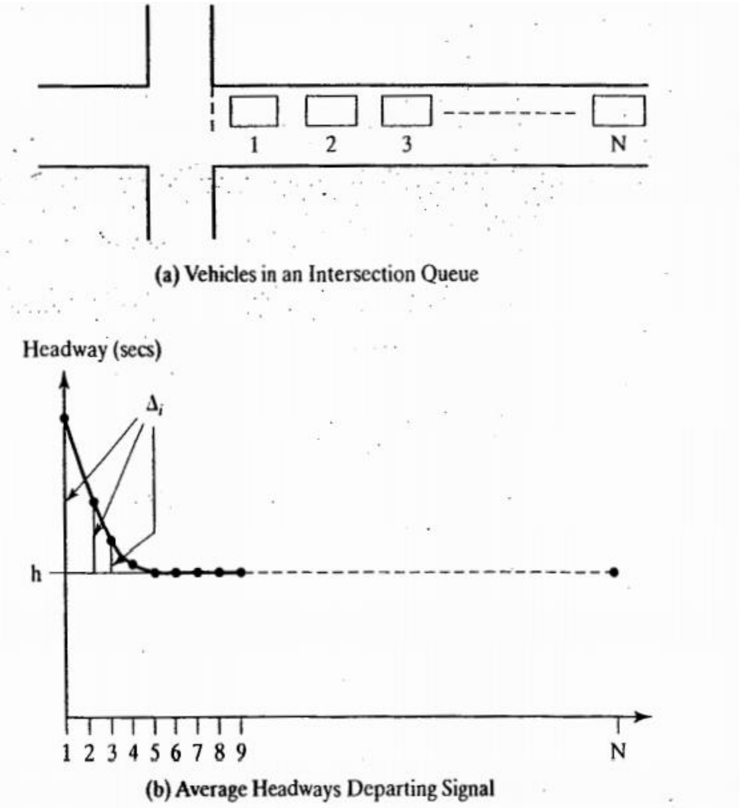
The terms protected and permitted may also be applied to right turns. In this case, however, the conflict is between the right-turn vehicular movement and the pedestrian movement in the conflicting crosswalk. The vast majority of right turf at signalized intersections are handled on a permitted basis Protected right turns generally occur at locations where there are overpasses or underpasses provided for pedestrians. At these locations, pedestrians are prohibited from making surface crossings; barriers are often required to enforce such a prohibition.

Discharge Headways, Saturation Flow, Lost Times, and Capacity

The fundamental element of a signalized intersection is the periodic stopping and restarting of the traffic stream. Figure 1 illustrates this process. When the light turns GREEN, there is a queue of stored vehicles that were stopped during the preceding RED interval, waiting to be discharged. As the queue of vehicles moves, headway measurements are taken as follows:

- The first headway is the time lapse between the initiation of the GREEN signal and the time that the font wheels of the first vehicle cross the stop line.
- The second headway is the time lapse between the time that the first vehicle's front wheels cross the stop line and the time that the second vehicle's front wheels cross the stop line.
- Subsequent headways are similarly measured.
- Only headways through the last vehicle in queue (at the initiation of the GREEN light) are considered to be operating under "saturated" conditions.

If many queues of vehicles are observed at a given location and the average headway is plotted versus the queue position of the vehicle, a trend similar to that shown in Figure 1 (b) emerges.
The first headway is relatively long. The first driver must go through the full perception-reaction sequence, move his or her foot from the brake to the accelerator, and accelerate through the intersection. The second headway is shorter because the second driver can overlap the perception-reaction and acceleration process of the first driver. Each successive headway is a little bit smaller than the last. Eventually, the headways tend to level out. This generally occurs when queued vehicles have fully accelerated by the time they cross the stop line. At this point, a stable moving queue has been established.



(a) Vehicles in an Intersection Queue

(b) Average Headways Departing Signal

Figure 1: Row Departing a Queue at a Signalized Intersection

Saturation Headway and Saturation Flow Rate

As noted, average headways tend toward a constant value. In general, this occurs from the fourth or fifth headway position. The constant headway achieved is referred to as the saturation headway because it is the average headway that can be achieved by a saturated, stable moving queue of vehicles passing though the signal. It is given the symbol "h," in units of s/veh.
It is convenient to model behavior at a signalized inter- section by assuming that every vehicle (in a given lane) consumes an average of "h" seconds of green time to enter the intersection. If every vehicle consumes "h" seconds of green time and if the signal were always green, then "s" vehicles per hour could enter the intersection. This is referred to as the saturation flow rate:

$$\mathrm{s} = \frac{3600}{\mathrm{h}} \qquad 1$$

Where:
s = saturation low rate, vehicles per hour of green per lane (veh/hg/ln).
h= saturation headway, seconds/vehicle (s/veh).

Saturation flow rate can be multiplied by the number of lanes provided for a given set of movements to obtain a saturation / flow rate for a lane group or approach.
The saturation low rate, in effect, is the capacity of the approach lane or lanes if they were available for use all of the time (i.e., if the signal were always GREEN). The signal, of course, is not always GREEN for any given movement Thus, some mechanism (or model) for dealing with the cyclic starting and stopping of movements must be developed.

Start-Up Lost Time

The average headway per vehicle is actually greater than "A" seconds. The first several headways are, in fact, larger than "h" seconds, as illustrated in Exhibit 1 (b). The first three or four headways involve additional time as drivers react to the GREEN signal and accelerate. The additional time involved in each of these initial headways (above and beyond "h" seconds) is noted by the symbol Δ_i (for headway i).
These additional times are added and referred to as the start-up lost time:

$$l_i = \sum_i \Delta_i \qquad 2$$

Where:
l_i = start-up lost time, s/phase.
Δ_i = incremental headway (above "h" seconds) for vehicle i, s.

Thus it is possible to model the amount of GREEN ti required to discharge a queue of "n" vehicles as:

$$T_n = l_1 + nh \qquad 3$$

Where:

T_n = GREEN time required to move queue of "«" vehicles through a signalized intersection, s.
l_i = start-up lost time, s/phase.
n = number of vehicles in queue.
h= saturation headway, s/veh.

Although this particular model is not of great use, it does illustrate the basic concepts if saturation headway and startup lost times. The start-up lost time is thought of as a period of time that is "lost" to vehicle use. Remaining GREEN time, however, may be assumed to be usable at a rate of h s/veh.

Clearance Lost Time

The start-up lost time occurs every time a queue of vehicles starts moving on a GREEN signal. There is also a lost time associated with stopping the queue at the end of the GREEN signal. This time is more difficult to observe in the field because H requires that the standing queue of vehicles be large enough to consume all of the GREEN time provided. In such a situation. The clearance lost time,l_2 is defined as the time interval between the last vehicle's front wheels crossing the stop line and the initiation of the GREEN for the next phase. The clearance lost time occurs each time a flow of vehicles is stopped.

Total Lost Time and the Concept of Effective GREEN Time

If the start-up lost time occurs each time a queue starts w move and the clearance lost Time occurs each time the flow of vehicles stops, then. For each GREEN phase:

$$t_L = l_1 + l_2 \qquad 4$$

Where:
t_L = total lost time per phase, s/phase.
All other variables are as previously defined.

The concept of lost times leads to the concept of effective green time. The actual signal goes through a sequence of intervals for each signal phase:

- Green
- Yellow
- All-red
- Red

They "yellow" and "all-red" intervals are a transition between GREEN and RED. This must be provided because vehicles cannot stop instantaneously when the light changes. The "all-red" is a period of time during which all lights in all directions are red. During the RED interval for one set of movements, another set of movements goes through the green, yellow, and all-red intervals. These intervals are defined in next lectures.

In terms of modeling, there are really only two time periods of interest: effective green time and effective red time. For any given set of movements, effective green time is the amount of time that vehicles can move (at a rate of one vehicle every h seconds). The effective red time is the amount cannot move (at a rate of one vehicle every h seconds). Effective green time is related to actual green time as follows:

$$g_i = G_i + Y_i - t_{Li} \qquad 5$$

Where:

g_i= effective green time for movement(s) i, s
G_i = actual green time for movements) i, s
Y_i = sum of yellow and all red intervals for movement(s) i, (Y_i =y_i + ar_i)
y_i = yellow interval for movements) i, s
ar_i = all-red interval for movement(s) i, s
t_{Li} = total lost time for movement(s) i, s

This model results in an effective green time that may be fully used by vehicles at the saturation flow rate (i.e., at an average headway of h s/veh).

Capacity of an intersection Lane or Lane Group

The saturation flow rate(s) represents the capacity of an interaction lane or lane group assuming that the light is always GREEN. The portion of real time that is effective green is defined by the "green ratio," the ratio of the effective green " me to the cycle length of the signal (g/Q). The capacity of an intersection lane or lane group may then be computed as:

$$c_i = s_i({}^{g_i}/_{C}) \qquad 6$$

Where:

q = capacity of lane or lane group i, veh/h
s_i = saturation flow rate for lane or lane group i, veh/hg
g_i = effective green time for lane or lane group i, s
C = signal cycle length, s.

A Sample Problem

These concepts are best illustrated using a sample problem. Consider a given movement at a signalized intersection with the following known characteristics:

- Cycle length, C = 60 s
- Green time, G = 27 s
- Yellow plus all-red time, Y=4 s
- Saturation headway, h = 2.4 s/veh
- Start-up lost time, l_1 = 2.0 s
- Clearance lost time, l_2 = 2.0 s

For these characteristics, what is the capacity (per lane) for this movement?

The problem will be approached in two different ways. In first, a ledger of time within the hour is created. Once the amount of time per hour used by vehicles at the saturation flow rate is established, capacity can be found by assuming that this time is used at a rate of one vehicle every h seconds. Because the characteristics stated are given on a per phase basis, these would have to be converted to a per hour basis. This is easily done knowing the number of signal cycles that occur within an hour. For a 60-second cycle, there are 3,600/60 = 60 cycles within the hour. The subject movements will have one GREEN phase in each of these cycles. Then:

- Time in hour: 3, 600 s
- RED time in hour: (60 - 27 -4) X 60 = 1740 s
- Lost time in hour: (2. 0 + 2.0) X 60 =240 s
- Remaining time in hour: 3600 – 1740 – 240 = 1620 s

The 1,620 remaining seconds of time in the hour represent the amount of time that can be used at a rate of one vehicle every h seconds, where h = 2.4 s/veh in this case. This number was calculated by deducting the periods during which no vehicles (in the subject movements) are effectively moving. These periods include the RED time as well as the start-up and clearance lost times in each signal cycle. The capacity of this movement may then be computed as:

$$c = \frac{1620}{2.4} = 675\frac{veh}{hr}/ln$$

A second approach to this problem uses Equation 6, with the following values:

$$s = \frac{3600}{2.4} = 1500\frac{veh}{hr}/ln$$

$$g = 27 - 4 + 4 = 27\,s$$

$$c = 1500\left({}^{27}\!/_{60}\right) = 675\frac{veh}{hr}/ln$$

The two results are, as expected, the same. Capacity is found by isolating the effective green time available to the subject movements and by assuming that this time is used at the saturation flow rate (or headway).

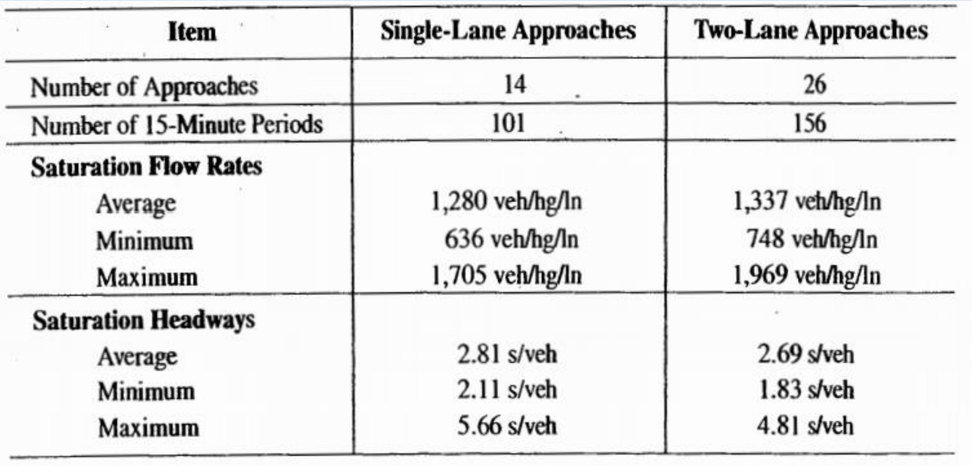
Notable Studies on Saturation Headways, Flow Rates, and Lost Times

For purposes of illustrating basic concepts, subsequent sections of this chapter assume that the value of saturation flow rate (or headway) is known. In reality, the saturation flow rate varies widely with a variety of prevailing conditions, including lane widths, heavy-vehicle presence, and approach grades, parking conditions near the intersection, transit bus presence, vehicular and pedestrian low rates, and other conditions.

The first significant studies of saturation low were conducted by Bruce Greenshields in the 1940s. His studies resulted in an average saturation flow rate of 1,714 veh/hg/ln and a start-up lost time of 3.7 seconds. The study, however, covered a variety of intersections with varying underlying characteristics. A later study in 1978 reexamined the Greenshields hypothesis; it resulted in the same saturation low rate (1,714 veh/hg/ln) but a lower start-up lost time of 1.1 seconds. The latter study had data from 175 intersections covering a wide range of underlying characteristics.

A comprehensive study of saturation low rates at intersections in five cities was conducted in 1987-1988 to determine the effect of opposed left turns. It also produced a good deal of data on saturation low rates in general. Some of the results are summarized in Table 2.

Table 2: Saturation Flow Rates from a Nationwide Survey



Item	Single-Lane Approaches	Two-Lane Approaches
Number of Approaches	14	26
Number of 15-Minute Periods	101	156
Saturation Flow Rates		
Average	1,280 veh/hg/ln	1,337 veh/hg/ln
Minimum	636 veh/hg/ln	748 veh/hg/ln
Maximum	1,705 veh/hg/ln	1,969 veh/hg/ln
Saturation Headways		
Average	2.81 s/veh	2.69 s/veh
Minimum	2.11 s/veh	1.83 s/veh
Maximum	5.66 s/veh	4.81 s/veh

These results show generally lower saturation flow rates (and higher saturation headways) than previous studies. The data, however, reflect the impact of opposed left turns, truck presence, and a number of other “nonstandard" conditions, all of which have a significant impeding effect. The most remarkable result of this study, however, was the wide variation in measured saturation flow rates, both over time at the same site and from location to location. Even when underlying conditions remained fairly constant, the variation in observed saturation low rates at a given location was as large as 20% to 25%. In a doctoral dissertation) using the same data, Prassas demonstrated that saturation headways and low rates have a significant Stochastic), component, making calibration of stable values difficult.

The study also isolated saturation flow rates for " ideal" conditions, which include all passenger cars, no turns, level grade, and 12-foot lanes. Even under these conditions, saturation low rates varied from 1,240 pc/hg/In to 2,092 pc/hg/ln for single-lane approaches and from 1,668 pc/hg/ln to 2,361 pc/hg/ln for multilane approaches. The difference between observed saturation low rates at single and multilane approaches is also interesting. Single-lane approaches have a number of unique characteristics that are addressed in the Highway Capacity Manual model for analysis of signalized intersections (see Chapter 24).

Current standards in the Highway Capacity Manual use an ideal saturation flow rate of 1,900 pc/hg/ln for both single and multilane approaches. This ideal rate is then adjusted for a variety of prevailing conditions. The manual also provides default values for lost times. The default value for start-up lost time l_1 is 2.0 seconds. For the clearance lost time (l_2). The default value varies with the "yellow" and all-red" timings of the signal:

$$l_2 = y + ar - e \qquad 7$$

<u>Where:</u>

l_2 = clearance lost time, s
y = length of yellow interval, s

ar = length of all-red interval, s
e = encroachment of vehicles into yellow and all-red, s
A default value of 2.0 s is used for e.

The Critical-Lane and Time –Budget Concepts

In signal analysis and design, the " critical-lane" and "time budget" concepts are closely related. The time budget, in its simplest form, is the allocation of time to various vehicular and pedestrian movements at an intersection through signal control. Time is a constant: There are always 3,600 seconds in an hour, and all of them must be allocated. In any given hour, time is “budgeted" to legal vehicular and pedestrian movements and to lost times.
The “critical lane” concept involves the identification of specific movements that will control the timing of a given signal phase. Consider the situation illustrated in Figure 2. A simple two-phase signal controls the intersection.

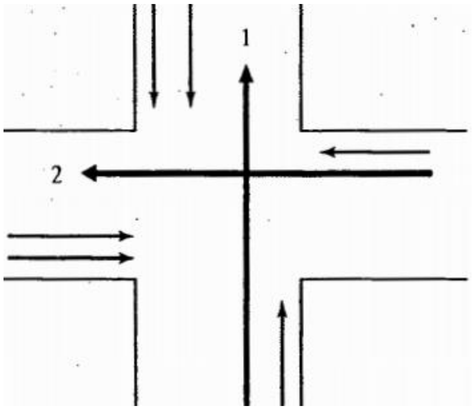


Figure 2: Critical Lanes Illustrated

Thus all E-W movements are permitted during one phase, and all N-S movements are permitted in another phase. During each of these phases, there are four lanes of traffic (two in each direction) moving simultaneously. Demand is not evenly distributed among them; one of these lanes will have the most intense traffic demand. The signal must be timed to accommodate traffic in this lane-the "critical lane" for the phase.

In the illustration of Figure 2, the signal timing and design must accommodate the total demand flows in lanes 1 and 2, because these lanes have the most intense demand, if the signal accommodates them, all other lanes will be accommodated as well. Note that the critical lane is identified as the lane with the most intense traffic demand, not the lane with the highest volume.
This is because many variables are affecting traffic flow.

A lane with many left-turning vehicles, for example, may require more time than an adjacent lane with no turning vehicles but a higher volume. Determining the intensity of traffic demand in a lane involves accounting for prevailing conditions that may affect flow in y that particular lane. y In establishing a time budget for the intersection of Figure 2, time would have to be allocated to four elements:

- Movement of vehicles in critical lane 1
- Movement of vehicles in critical lane 2
- Start-up and clearance lost times for vehicles in critical lane 1
- Start-up and clearance lost times for vehicles in critical lane 2

This can be thought of in the following way: Lost times are not used by any vehicle. When deducted from total time, remaining time is effective green time and is allocated to critical-lane demands-in this case, in lanes 1 and 2. The total amount of effective green time, therefore, must be sufficient to accommodate the total demand in lanes 1 and 2 (the critical lanes). These critical demands must be accommodated one vehicle at a time because they cannot move simultaneously.

The example of Figure 2 is a relatively simple case. In general, the following rules apply to the identification of critical lanes:

a. There is a critical lane and a critical-lane flow for each discrete signal phase provided.
b. Except for lost times, when no vehicles move, there must be one and only one critical lane moving during every second of effective green time in the signal cycle.

c. Where there are overlapping phases, the potential combination of lane flows yielding the highest sum of critical lane flows while preserving the requirement t of item (b) identified critical lanes.

The maximum Sum of Critical –Lane Volumes: One View of Signalized Intersection Capacity

It is possible to consider the maximum possible sum of critical-lane volumes to be a general measure of the "capacity" of the intersection. This is not die same as the traditional view of capacity presented in the Highway Capacity Manual, but it is i a useful concept to pursue.

By definition, each signal phase has one and only one critical lane. Except for lost times in the cycle, one critical lane is always moving. Lost times occur for each signal phase and represent time during which no vehicles in any lane are moving. The maximum sum of critical-lane volumes may, therefore, be found by determining how much total lost time exists in the hour. The remaining time (total effective green time) may then be divided by the saturation headway.

To simplify this deviation, it is assumed the total lost time per phase (t_L) is a constant for all phases. Then, the total lost time per signal cycle is:

$$L = N \times t_L \qquad 8$$

Where:

L = lost time per cycle, s/cycle

t_L= total lost lime per phase (sum of $l_1 + l_2$) s/phase

N = number of phases in the cycle

The total lost time in an hour depends on the number of cycles occurring in the hour:

$$L_H = L\left(\frac{3600}{C}\right) \qquad 9$$

Where:
L_H = lost time per hour, s/hr
L = lost time per cycle, s/cycle
C = cycle length, s

The remaining time within the hour is devoted to effective green time for critical-lane movements:

$$T_G = 3600 - L_H \qquad 10$$

Where:
T_G = total effective green time in the hour, s

This time may be used at a rate of one vehicle every h seconds, where h is the saturation headway:

$$V_c = \frac{T_G}{h} \qquad 11$$

Where:
V_c = maximum sum of critical-lane volumes, veh/h
h= saturation headway, s/veh

Merging Equations 8 through 11, the following relationship emerges:

$$V_c = \frac{1}{h}\left[3600 - Nt_L(\frac{3600}{C})\right] \qquad 12$$

All variables are as previously defined.
Consider the example of Figure 2 again. If the signal at this location has two phases, a cycle length of 60 seconds, total lost times of 4 s/phase, and a saturation headway of 2.5 s/veh, the maximum sum of critical-lane flows (the sum of flows in lanes 1 and 2) is:

$$V_c = \frac{1}{2.5}\left[3600 - 2 * 4 * (\frac{3600}{60})\right] = 1248 \text{ veh/hr} \qquad 13$$

The equation indicates there are 3,600/60 = 60 cycles in an hour. For each of these, 2 * 4 = 8 s of lost time is experienced, for a total of 8 * 60 = 480 s in the hour. The remaining 3,600 - 480 = 3,120 s may be used at a rate of one / vehicle every 2.5 seconds.

If Equation 12 is plotted, an interesting relationship between the maximum sum of critical-lane volumes (V_c), cycle length (C), and number of phases (N) may be observed, as illustrated in Figure 3.
As the cycle length increases, the "capacity" of the intersection also increases. This is because of lost time, which are constant per cycle. The longer the cycle length, the fewer cycles there are in an hour. This leads to less lost time in the hour, more effective green time in the hour, and a higher sum of critical-lane volumes. Note, however, that the relationship gets flatter as cycle length

increases. As a general rule, increasing the cycle length may result in small increases in capacity. However, capacity can rarely be increased significantly by only increasing the cycle length. Other measures, such as adding lanes, are often also necessary.

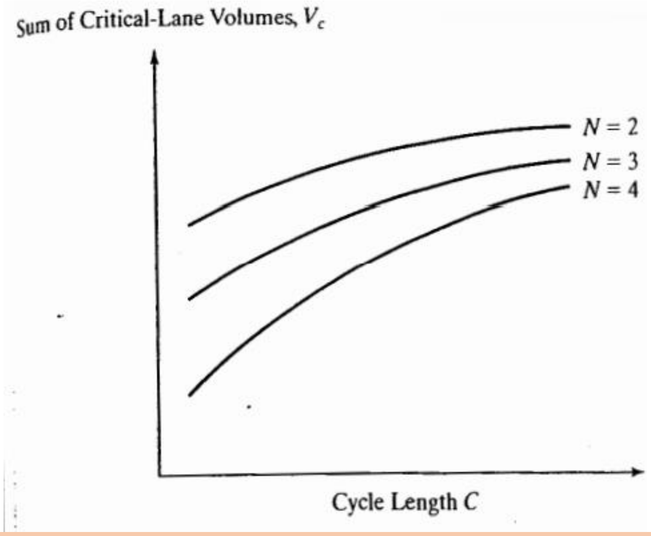


Figure 3: Maximum Sum of Critical-Lane Volumes Plotted

Capacity also decreases as the number of phases increases. This is because for each phase, there is one full set of lost times in the cycle. Thus a two-phase signal has only 1 two sets of lost times in the cycle, and a three-phase signal has three.
These trends provide insight but also raise an interesting question: Given these trends, it appears that all signals should have two phases and that the maximum practical I cycle length should be used in all cases. After all, this combination would, apparently, yield the highest "capacity" for the intersection.

Using the maximum cycle length is not practical unless truly needed. Having a cycle length that is considerably longer than what is needed causes increases in delay to drivers and passengers. The increase in delay is because there will be times when vehicles on one approach are waiting for the green while there is no demand on conflicting approaches. Shorter cycle lengths yield less delay. Further, there is no insensitive to maximize the cycle length. There will always be 3600 seconds in the hour, and increasing the cycle length to accommodate increasing demand over time is quite simple, requiring only a resetting of the local signal controller. The shortest cycle length consistent with a v/c ratio in the range of 0. 80 to 0.95 is generally used to produce optimal delays. Thus the view of signal capacity is quite different from that of pavement capacity. When deciding on the number of lanes on a freeway (or on an intersection approach), it is desirable to build excess capacity (i.e., achieve a low v/c ratio). This is because once built, it is unlikely that engineers will get an opportunity to expand the facility for 20 or more years, and adjacent land development may make such expansion impossible. The 3,600 seconds in an hour, however, are immutable and retiming the signal to allocate more of them to effective green time is a simple task requiring no field construction.

Finding an Appropriate Cycle Length

If it is assumed that the demands on an intersection are known and the critical lanes can be identified, then Equation 12 could be solved using a known value of V_c to find a minimum acceptable cycle length:

$$C_{min} = \frac{Nt_L}{1-(\frac{V_c}{3600/h})} \qquad 13$$

Thus, if in the example of Figure 2, the actual sum of critical-lane volumes was determined to be 1, 000 veh/h, the minimum feasible cycle length would be:

$$C_{min} = \frac{2*4}{1-(\frac{1000}{3600/2.5})} = 26.2 \text{ s}$$

The cycle length could be reduced, in this case, from the given 60 seconds to 30 seconds (the effective minimum cycle length used). This computation, however, assumes that the demand (V) is uniformly distributed throughout the hour and that every second of effective green time will be used. Neither of these assumptions is very practical. In general, signals would be timed for the flow rates occurring in the peak 15 minutes of the hour. Equation 13 could be modified by dividing V_c by a known peak-hour factor (PHF) to estimate the flow rate in the worst 15-minute period of the hour. Similarly, most signals would be timed to have somewhere between 80% and 95% of the available capacity actually used. Due to the normal stochastic variations in demand on a cycle by-cycle and daily basis, some excess capacity must be provided to avoid failure of individual cycles or peak periods on a specific day. If demand, V_c, is also divided by the expected utilization of capacity (expressed in decimal form), then this is also accommodated. Introducing these changes transforms Equation 13 to:

$$C_{des} = \frac{Nt_L}{1-(\frac{1000}{\left(\frac{3600}{h}\right)*PHF*(\frac{v}{c})})} \qquad 14$$

<u>Where:</u>
C_{des} = desirable cycle length, s
PHF - peak hour factor.
v/c = desired volume to capacity ratio.
All other variables are as previously defined.

Returning to the example if the PHF is 0.95 and it is desired to use no more than 90% of available capacity during the peak 15-minute period of the hour, then:

$$C_{des} = \frac{2*4}{1-(\frac{1000}{\left(\frac{3600}{2.5}\right)*0.95*0.9})} = \frac{8}{0.188} = 42.6\ s$$

In practical terms, this would lead to the use of a 45-second cycle length.
The relationship between a desirable cycle length, the sum of critical-lane volumes, and the target v/c ratio is quite interesting and is illustrated in Figure 4.
Figure 4 illustrates a typical relationship for a specified number of phases, saturation headway, lost times, and PHF. If a vertical is drawn at any specified value of Vc (sum of critical-lane volumes), it is clear that the resulting cycle length is very sensitive to the target v/c ratio. Because the curves for each v/c ratio are eventually asymptotic to the vertical, it is not always possible to achieve a specified v/c ratio.

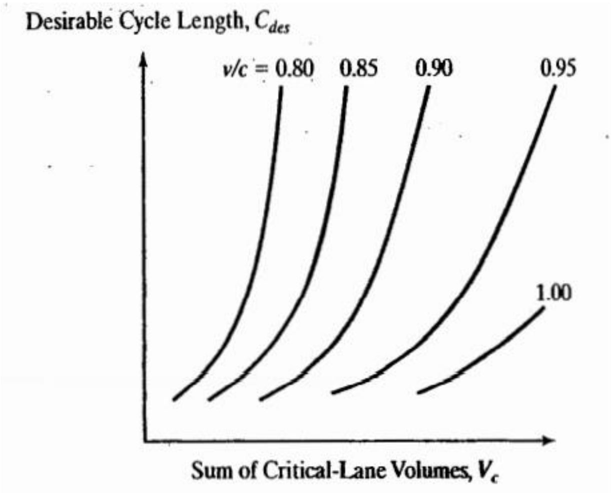


Figure 4: Desirable Cycle Length versus Sum of Critical-Lane Volumes.

Consider the case of a three-phase signal, with t_L = 4 s/phase, a saturation headway of 2. 2 s/veh, a PHF of 0. 90, and Vc = 1,200 veh/h. Desirable cycle lengths will be computed for a range of target v/c ratios varying from 1. 00 to 0.80.

$$C_{des} = \frac{3*4}{1-(\frac{1200}{\left(\frac{3600}{2.2}\right)*0.95*1.00})} = \frac{12}{0.182} = 64.8\ s \rightarrow 65\ s$$

$$C_{des} = \frac{3 * 4}{1 - (\frac{1200}{\left(\frac{3600}{2.2}\right) * 0.95 * 0.95})} = \frac{12}{0.1432} = 84.3\ s\ \rightarrow 85\ s$$

$$C_{des} = \frac{3 * 4}{1 - (\frac{1200}{\left(\frac{3600}{2.2}\right) * 0.95 * 0.90})} = \frac{12}{0.0947} = 126.7\ s \rightarrow 130\ s$$

$$C_{des} = \frac{3 * 4}{1 - (\frac{1200}{\left(\frac{3600}{2.2}\right) * 0.95 * 0.85})} = \frac{12}{0.0414} = 289.9\ s \rightarrow 290\ s$$

$$C_{des} = \frac{3 * 4}{1 - (\frac{1200}{\left(\frac{3600}{2.2}\right) * 0.95 * 0.80})} = \frac{12}{-0.0185} = -648.6\ s$$

For this case, reasonable cycle lengths can provide target v/c ratios of 1.00 or 0.95. Achieving v/c ratios of 0.90 or 0.85 would require long cycle lengths beyond the practical limit of 120 seconds for pretimed signals. The 130-second cycle needed to achieve a v/c ratio of 0.90 might be acceptable for an actuated signal location, or in some extreme cases warranting a longer pretimed signal cycle. However, a v/c ratio of 0.80 cannot be achieved under any circumstances. The negative cycle length that results signifies there is not enough time within the hour accommodate the demand with the required green time plus the 12 seconds of lost time per cycle. In effect, more than 3,600 seconds would have to be available in the hour to accomplish this.

A Sample problem

Consider the intersection shown in Figure 5. The critical directional demands for this two-phase signal are shown with other key variables. Using the time-budget and critical-lane concepts, determine the number of lanes required for each of the critical movements and the minimum desirable cycle length that could be used. Note that an initial cycle length is specified but will be modified as part of the analysis.

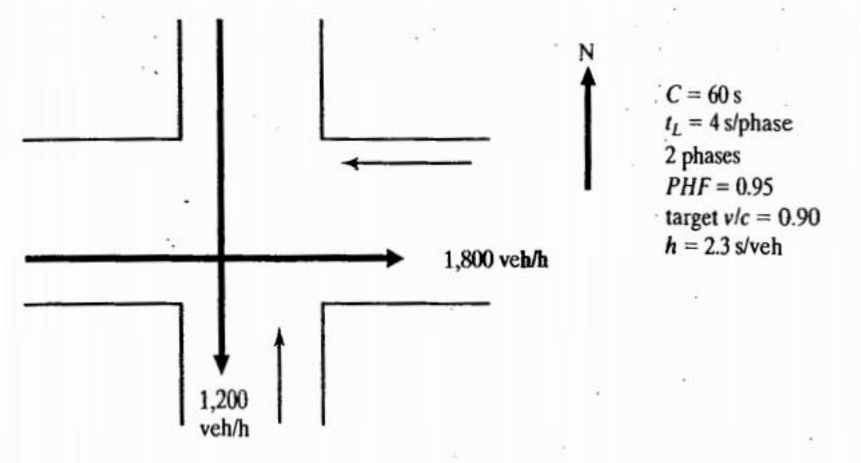


Figure 25: Sample Problem Using the Time-Budget and Critical-Lane Concepts

Assuming that the initial specification of a 60-second cycle is correct and given the other specified conditions, the maximum sum of critical lanes that can be accommodated is computed using Equation 12:

$$V_c = \frac{1}{2.3}\left[3600 - 2 * 4 * (\frac{3600}{60})\right] = 1357\ veh/hr$$

The critical SB volume is 1,200 veh/h, and the critical EB volume is 1,800 veh/h. The number of lanes each must be divided into is now to be determined. Whatever combination is .used, the sum of the critical-lane volumes for these two approaches must be below 1,357 veh/h. Figure 20.6 shows a number of possible lane combinations and the resulting sum of critical-lane volumes. As you can see from the scenarios of Figure 6, to have a sum of critical-lane volumes less than 1,357 veh/h, the SB approach must have at least two lanes, and the EB approach must have three lanes. Realizing that these demands probably reverse in the other peak hour (am or pm), the N-S artery would probably require four lanes, and the E-W artery six lanes.

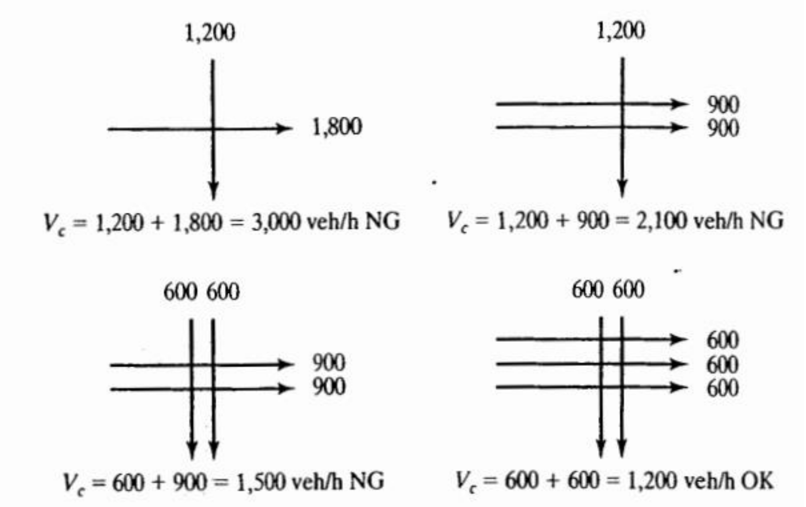

1,200
1,800
V_c = 1,200 + 1,800 = 3,000 veh/h NG
1,200
900
900
V_c = 1,200 + 900 = 2,100 veh/h NG
600 600
900
900
V_c = 600 + 900 = 1,500 veh/h NG
600 600
600
600
600
V_c = 600 + 600 = 1,200 veh/h OK

Figure 6: Possible Lane Scenarios for Sample Problem

This is a very basic analysis, and it would have to be modified based on more specific information regarding individual movements, pedestrians, parking needs, and other factors.
If the final scenario is provided, Vc is only 1, 200 veh/h. It is possible that the original cycle length of 60 seconds could be reduced. A desirable cycle length may be computed from Equation 14:

$$C_{des} = \frac{2 * 4}{1 - (\frac{1200}{\left(\frac{3600}{2.3}\right) * 0.95 * 0.90})} = 77.7\ s \rightarrow 80\ s$$

The resulting cycle length is larger than the original 60 seconds because the equation takes both the PHF and target v/c ratios into account. Equation 12 for computing the maximum value of does not; it assumed full use of capacity (v/c = 1.00) and no peaking within the hour. In essence, the (2 ×3) lane design proposal should be combined with an 80-second cycle length to achieve the desired results.
This problem illustrates the critical relationship between number of lanes and cycle lengths. Clearly, other scenarios would produce desirable results. Additional lanes could be provided in either direction, which would allow the use of a shorter cycle length. Unfortunately, for many cases, signal timing is considered with a fixed design already in place. Only where right-of-way is available or a new intersection is being constructed can major changes in the number of lanes be considered. Allocation of lanes to various movements is also a consideration. Optimal solutions are generally found more easily when the physical design and signalization can be treated in tandem.
If, in the problem of Figure 5, space limited both the EB and SB approaches to two lanes, the resulting V_C would be 1,500 veh/h. Would it be possible to accommodate this demand by lengthening the cycle length? Again, Equation 14 is used:

$$C_{des} = \frac{2 * 4}{1 - (\frac{1500}{\left(\frac{3600}{2.3}\right) * 0.95 * 0.90})} = -66.1s \rightarrow NG$$

The negative result indicates no cycle length can accommodate a Vc of 1500 veh/h at this location.

The Concept of Left-Turn (and Right-Turn) Equivalency

The most difficult process to model at signalized intersection is the left turn. Left turns are made in several different modes using different design elements. Let turns may be made from a lane shared with through vehicles (shared-lane operation) or from a lane dedicated to left-turning vehicles (exclusive-lane operation). Traffic signals may allow for permitted or protected left turns, or some combination of the two.

Whatever the case, however, a let-turning vehicle will consume more effective green time traversing the intersection than will a similar through vehicle. The most complex case is that of a permitted left turn made across an opposing vehicular flow from a shared lane. A left-turning vehicle in the shared lane must wait for an acceptable gap in the opposing flow. While waiting, the vehicle blocks the shared lane and other vehicles (including through vehicles) in the lane are delayed behind it. Some vehicles will change lanes to avoid the delay while others are unable to and must wait until the left-turner successfully completes the turn

Many models of the signalized intersection account for this in terms of through vehicle equivalents" (i.e., how many through vehicles would consume the same amount of effective green time traversing the stop-line as one left-turning vehicle?). Consider the situation depicted in Figure 7.

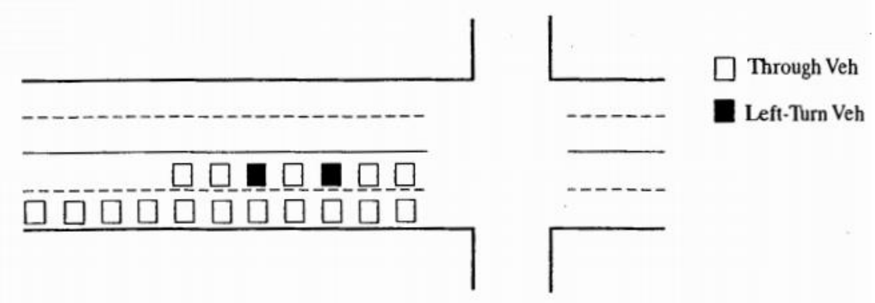


Figure 7: Sample Equivalence Observation on a Signalized Intersection Approach

If both the left lane and the right lane were observed, an equivalence similar to the following statement could be determined: *In the same amount of time, the left lane discharges in' through vehicles and two let-turning vehicles while the right lane discharges eleven through vehicles*. In terms of effective green time consumed, this observation means that 11 through vehicles are equivalent to 5 through vehicles plus 2 let turning vehicles. If the left-tum equivalent is defined as E_{LT}:

$11=5+2\ E_{LT}$

$E_{LT} = (11-5)/2 = 3.0$

Note that this computation holds only for the prevailing characteristics of the approach during the observation period- The left-turn equivalent depends on a number of factors, including how left turns are made (protected, permitted, compound), the opposing traffic flow and the number of opposing lanes.

Figure 8 illustrates the general form of the relationship for through vehicle equivalents of permitted left turns.

The left-turn equivalent, E_{LT}, increases as the opposing (low increases. For any given opposing flow, however, the equivalent decreases as the number of opposing lanes is increased from one to three. This latter relationship is not linear because the task of selecting a gap through multilane opposing traffic is more difficult than selecting a gap through single-lane opposing traffic. Further, in a multilane traffic stream, vehicles do not pace each other side by side, and the gap distribution does not improve as much as the per-lane opposing flow deceases. To illustrate the use of let-turn equivalents in modeling, consider the following problem:

An approach to a signalized intersection has two lanes, permitted left-tum phasing, 10% left-turning vehicles, and a left-tum equivalent of 5.0. The saturation headway for through vehicles is 2.0 s/veh, Determine the equivalent saturation flow rate and headway for all vehicles on this approach.

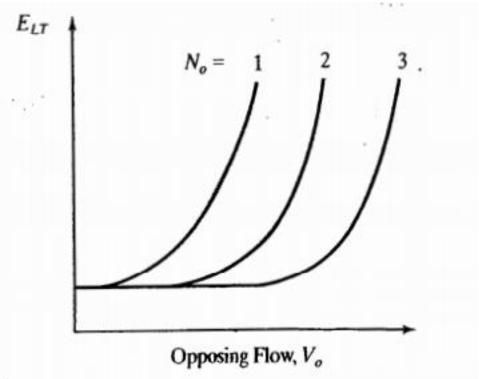


Figure 8: Relationship among Let-Turn Equivalents, Opposing Flow, and Number of Opposing Lanes

The first way to interpret the left-tum equivalent is that each left-tuning vehicle consumes 5.0 times the effective green time as a though vehicle. Thus, for the situation described, 10% of the traffic stream has a saturation headway of 2.0 × 5.0 = 10.0 s/veh, and the remainder (90%) has a saturation headway of 2.0 s/veh. The average saturation headway for all vehicles, therefore, is:

$$h = (0.1 \times 10.0) + (0.9 \times 2.0) = 2.80\ s/veh$$

This corresponds to a saturation flow rate of:

$$s = \frac{3600}{2.80} = 1286 \text{ veh/hr/ln}$$

A number of models, including the Highway Capacity Manual approach, calibrate a multiplicative adjustment factor that converts an ideal (or through) saturation flow rate to a saturation flow rate for prevailing conditions:

$$S_{prev} = S_{ideal} \times f_{LT}$$

$$f_{LT} = \frac{S_{prev}}{S_{ideal}} = \frac{3600/h_{prev}}{3600/h_{ideal}} = \frac{h_{ideal}}{h_{prev}} \qquad 15$$

Where:

S_{prev} = satuation flow rate under prevailing conditions, veh/hg/ln
S_{ideal} = satuaion flow rate under ideal condiions, veh/hg/ln
f_{LT} = let-tum adjustment factor hided
h_{ideal}=saturation headway under ideal condiions, s/veh
h_{prev} = satunion headway under prevailing conditions, s/veh

In effect, in the first solution, the prevailing headway,S_{prev}, was computed as follows:

$$f_{LT} = (P_{LT}E_{LT}h_{ideal}) + [(1 - P_{LT})h_{ideal}] \qquad 16$$

Combining Equations 15 and 16:

$$f_{LT} = \frac{h_{ideal}}{(P_{LT}E_{LT}h_{ideal}) + [(1 - P_{LT})h_{ideal}]}$$

$$f_{LT} = \frac{1}{P_{LT}E_{LT}+(1-P_{LT})} = \frac{1}{1+P_{LT}(E_{LT}-1)} \qquad 17$$

The problem posed may now be solved using a left-turn adjustment factor. Note that the saturation headway under ideal conditions is 3,600/2.0 = 1,800 veh/hg/ln Then:

$$f_{LT} = \frac{1}{1 + 0.10(5 - 1)} = 0.714$$

$$S_{prev} = 1800 \times 0.714 = 1286 \text{ veh/hr/ln}$$

This, of course, is the same result.

It is important that the concept of left-turn equivalence be understood. Its use in multiplicative adjustment factors often obscures its intent and meaning. The fundamental concept, however, is unchanged- -the equivalence is based on the fact that the effective green time consumed by a left-tuning vehicle is f_{LT} times the effective green time consumed by a similar through vehicle.
A similar case can be made or describing the effects of right turns. Right tums are typically made through a conflicting pedestrian flow in the crosswalk to the immediate right of the approach.
Like left turns, this interaction causes right tums to consume more effective green time than through movements. An equivalent, E_{RT}, is used to quantify these effects and is used in the same manner as described for left-turn equivalents.

Signalized intersection and other types of equivalents as well. Heavy-vehicle and local bus equivalents have similar meanings and result in similar equations. Some of these have been discussed in previous lecture, and others will be discussed in next lecture.

Delay as a Measure of Effectiveness

Signalized intersections represent point locations within a surface street network. As point locations, the measures of operational quality or effectiveness used for highway sections are not relevant. Speed has no meaning at a point, and density requires a section of some length for measurement. A number of measures have been used to characterize the operational quality of a signalized intersection, the most common of which are:

- Delay
- Queuing
- Stops

These measures are all related. Delay refers to the amount of time consumed in traversing the intersection – the difference between the arrival time and the departure time, where these may be defined in a number of different ways.
Queuing refer to the number of vehicles forced to queue behind the stop line during a RED signal phase; common measures include the average queue length or a percentile queue length. Stops refer to the percentage or number of vehicles that must stop at the signal.

Types of Delay

The most common measure used to describe operational quality at a signalized intersection is delay, with queuing and/or stops often used as a secondary measure. Although ii is possible to measure delay in the field, it is a difficult process, and different observers may make judgments that could yield different results. For many purposes, it is, therefore, convenient to have a predictive model for the estimate of delay. Delay, however, can be quantified in many different ways. The most frequently used forms of delay are defined as follows:

1. Stopped-time delay. Stopped-time delay is defined as the time a vehicle is stopped in queue while waiting to pass through the intersection; average stopped time delays the average for all vehicles during a specified time period.
2. Approach delay. Approach delay includes stopped time delay but adds the time loss due to deceleration from the approach speed to a stop and the time loss due to reacceleration back to the desired speed. Average approach delay is the average for all vehicles during a specified time period.
3. Time-in-queue delay. Time-in-queue delay is the total time from a vehicle joining an intersection queue to its discharge across the STOP line on departure. Again average time-in-queue delay is the average or all vehicles during a specified time period.
4. Travel time delay. This is a more conceptual value. It is the difference between the driver's expected travel time through the intersection (or any roadway segment and the actual time

taken. Given the difficulty in establishing a "desired" travel time to traverse an intersection, this value is rarely used, other than as a philosophical concept.

5. Control delay. The concept of control delay was developed in the 1994 Highway Capacity Manual and is included in the current HCM. It is the delay caused by a control device, either a traffic signal or a STOP sign. It is approximately equal to time-in queue delay plus the acceleration-deceleration delay component.

Figure 9 illustrates three of these delay types for a single vehicle approaching a RED signal. Stopped-time delay for this vehicle includes only the time spent stopped at the signal, it begins when the vehicle is fully stopped and ends when the vehicle begins to accelerate. Approach delay includes additional time losses due to deceleration and acceleration. It is found by extending the velocity slope of the approaching vehicle as if no signal existed; the approach delay is the horizontal (time) difference between the hypothetical extension of the approaching velocity slope and the departure slope after full speed is achieved. Travel time delay is the difference in time between a hypothetical desired velocity line and the actual vehicle path. Time-in-queue delay cannot be effectively shown using one vehicle because it involves joining and departing a queue of several vehicles. Delay measures can be stated for a single vehicle, as an average for all vehicles over a specified time period, or as an aggregate total value for all vehicles over a specified time period. Aggregate delay is measured in total vehicle-seconds, vehicle-minutes, or vehicle-hours for all vehicles in the specified time interval. Average individual delay is generally stated in terms of s/veh for a specified time interval.

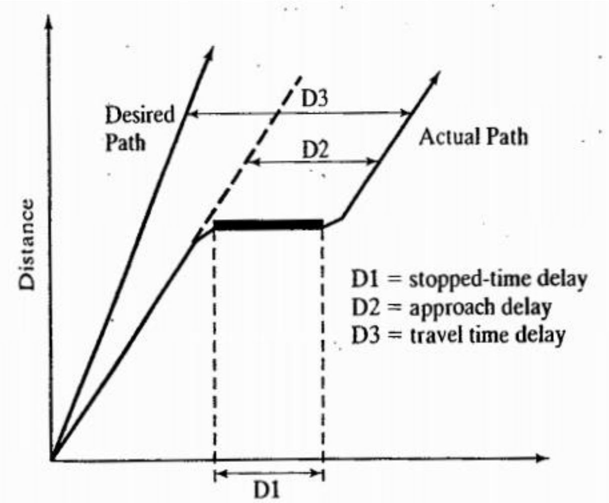


Figure 9: Illustration of Delay Measures

Basic Theoretical Models of Delay

Virtually all analytic models of delay begin with a plot of cumulative vehicles arriving and departing versus time at a given signal location. The time axis is divided into periods of effective green and effective red as illustrated in Figure 10. Vehicles are assumed to arrive at a uniform rate of flow of v vehicles per unit time, seconds in this case. This is shown by the constant slope of the arrival curve. Uniform arrivals assume that the inter-vehicle arrival time between vehicles is a constant. Thus, if the arrival flow rate, v is 1.800 vehs/h. then one vehicle arrives every 3600/1800=2.0 s.

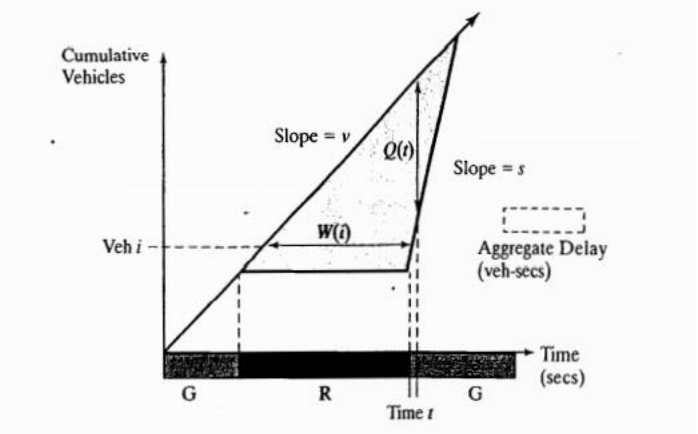


Figure 10: Delay, Waiting Time, and Queue Length Illustrated.

Assuming no preexisting queue, vehicles arriving when the light is GREEN continue through the intersection (i.e., the departure curve is the same as the arrival curve). When the light turns RED, however, vehicles continue to arrive, but none depart. Thus the departure curve is parallel to the x-axis during the RED interval. When the next effective GREEN begins, vehicles queued during the RED interval depart from the intersection, now at the saturation flow rate, s, in veh/s. For stable operations, depicted here, the departure curve “catches up" with the arrival curve before the next RED interval begins (i.e., there is no residual or unserved queue left at the end of the effective GREEN).

This simple depiction of arrivals and departures at a signal allows the estimation of thee critical parameters:

- The total time that any vehicle (spends waiting in the queue, is given by the horizontal time-scale difference between the time of arrival and the time of departure.
- The total number of vehicles queued at any time t, Q (t), is the vertical vehicle-scale difference between the number of vehicles that have arrived and the number of vehicles that have departed.
- The aggregate delay for all vehicles passing through the signal is the area between the arrival and departure curves (vehicles × time).

Note that because the plot illustrates vehicles arriving in queue and departing from queue, this model most closely represents what has been defined as time-in-queue delay. There are many simplifications that have been assumed, however, in constructing this simple depiction of delay. It is important to understand the two major simplifications:

- The assumption of a uniform arrival rate is a simplification. Even at a completely isolated location, actual arrivals would be random {i.e., would have an average rate over time), but inter-vehicle arrival times would vary around an average rather than being constant. Within coordinated signal systems, however, vehicle arrivals are usually in platoons.
- It is assumed that the queue is building at a point location (as if vehicles were stacked on top of one another). In reality, as the queue grows, the rate at which vehicles arrive at its end is the arrival rate of vehicles (at a point), plus a component representing the backward growth of the queue in space.

Both of these can have a significant effect on actual results. Modem models account for the former in ways that we discuss subsequently. The assumption of a "point queue" however, is imbedded in many modem applications.

Figure 11 show a series of GREEN phases and depicts three different types of operation. It also allows for an arrival function, a (t), that varies while maintaining the departure function, d (t), described previously. Figure 11 (a) shows stable low throughout the period depicted. No signal cycle "fails" (i.e., ends with some vehicles queued during the preceding RED unserved).

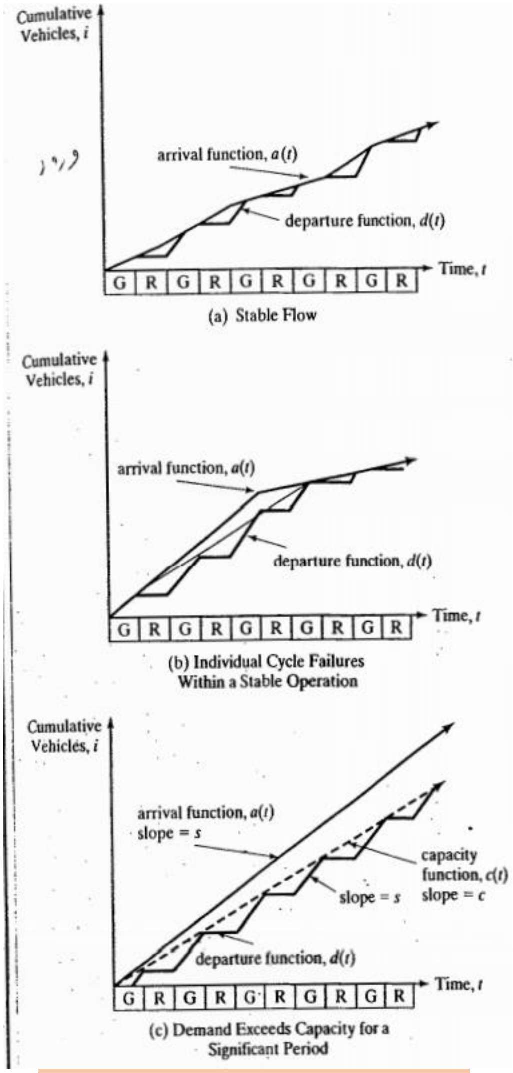


Figure 11: Three Delay Scenarios.

(Source: Adapted with permission of Transportation Research Board, National Research Council, Washington DC, from V. F.Hurdle, "Signalized Intersection Delay Model: A Primer for the Uninitiated," Transportation Research Record 971, pp. 97, 98, 1984.)

During every GREEN phase, the departure function "catches up" with the arrival function. Total aggregate delay during this period is the total of all the triangular areas between the arrival and departure curves. This type of delay is often referred to as "uniform delay".

In Figure 11 (b), some of the signal phases "fail." Ai the end of the second and third GREEN intervals, some vehicles are not served (i.e., they must wait for a second GREEN interval to depart the intersection. By the time the entire period ends, however, the departure function has "caught up" with the arrival function and there is no residual queue let unserved. This case represents a situation in which the overall period of analysis is stable (i.e., total demand does not exceed total capacity). Individual cycle failures within the period however, have occurred. For these periods, there is a second component of delay in addition to uniform delay, it consists of the area between the arrival function and the dashed line, which represents the capacity of the intersection to discharge vehicles and has the slope c. This type of delay is referred to as overflow delay. "Figure 11 (c) shows the worst possible case: Even GREEN interval "fails" for a significant period of time and the residual, or unserved, queue of vehicles continues to grow throughout the analysis period. In this case, the overflow delay component grows over time, quickly dwarfing the uniform delay component.

The latter case illustrates an important practical operational characteristic. When demand exceeds capacity (v/c > 1.00), the delay depends on the length of time that the condition exists. In Figure 11 (b), the condition exists for only two phases. Thus the queue and the resulting overflow delay are limited. In Figure 11 (c), the condition exists for a long time, and the delay continues to grow throughout the oversaturated period.

Components of Delay

In analytic models for predicting delay, three distinct components of delay may be identified:

- Uniform delay is the delay based on an assumption of uniform arrivals and stable low with no individual cycle failures.
- Random delay is the additional delay, above and beyond uniform delay, because flow is randomly distributed rather than uniform at isolated intersections.
- Overflow delay is the additional delay that occurs when the capacity of an individual phase or series of phases is less than the demand or arrival flow rate.

In addition, the delay impacts of platoon flow (rather than uniform or random) have been historically treated as an adjustment to uniform delay. Many modem models combine the random and overflow delays into a single function, which is referred to as "overflow delay," even though it contains both components. The differences between uniform, random, and platoon arrivals are illustrated in Figure 12. As noted, the analytic basis for most delay models is the assumption of uniform arrivals, which are depicted in Figure 12 (a). Even at isolated intersections, however, arrivals would be random, as shown in Figure 12 (b). With random arrivals, the underlying rate of arrivals is a constant, but the inter-arrival times are exponentially distributed around an average. In most urban and suburban cases, where a signalized intersection is likely to be part of a coordinated signal system, arrivals will be in organized platoons that move down the arterial in a cohesive group, as shown in Figure 12 (c). The exact time that a platoon arrives at a downstream signal has an enormous potential effect on delay. A platoon of vehicles arriving at the beginning of the RED forces most vehicles to stop for the entire length of the RED phase. The same platoon of vehicles arriving at the beginning of the GREEN phase may flow through the intersection without any vehicles stopping. In both cases, the arrival flow, v, and the capacity of the intersection, c, are the same. The resulting delay, however, would vary significantly. The existence of platoon arrivals, therefore, necessitates a significant adjustment to models based on theoretically uniform or random flow.

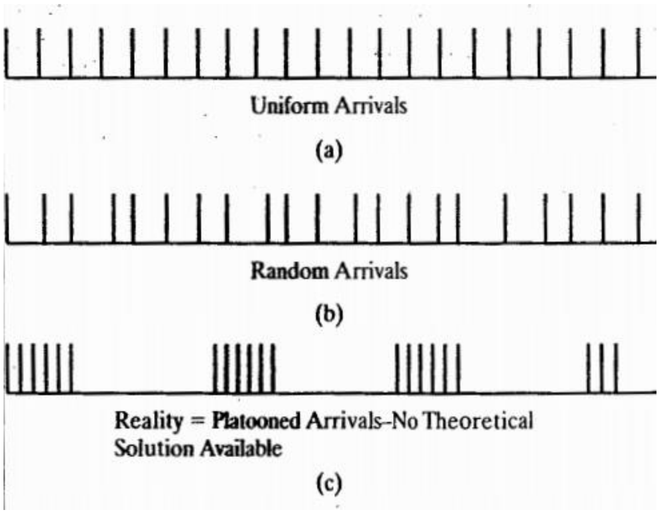

Uniform Arrivals
(a)
Random Arrivals
(b)
Reality = Platooned Arrivals-No Theoretical Solution Available
(c)

Figure 12: Arrival Patterns Compared

Webster's Uniform Delay Model

Virtually every model of delay starts with Webster ' s model of uniform delay. Initially published in 1958, this model begins with the simple illustration of delay depicted in Figure 13, with its assumptions of stable flow and a simple uniform arrival function. As noted previously, aggregate delay can be estimated as the area between the arrival and departure curves in the figure. Thus Webster ' s model for uniform delay is the area of the triangle formed by. The arrival and departure functions. For clarity, this triangle is shown again in Figure 13.
The area of the aggregate delay triangle is simply half the base times the height, or:

$$UD_a = \frac{1}{2}RV$$

UD_a= aggregate uniform delay, veh-sec.
R= length of the RED phase, s
V= total vehicles in queue, veh.

By convention, traffic models-are not developed in terms of RED time. Rather, they focus on GREEN time. Thus Webster substitutes the following equivalence for the length of the RED phase:

$$R = C\;[1 - (^{g}/_{C})]$$

<u>Where</u>:
C = cycle length, s
g = effective green time, s

In words, the RED time is the portion of the cycle length that is not effectively green.
The height of the triangle, V, is the total number of vehicles in the queue. In effect, it includes vehicles arriving during the RED phase, R, plus those that join the end of the queue while it is moving out of the intersection (i. e., during time t_c in Figure 13). Thus determining the time it takes for the queue to clear, t_c, is an important part of the model. This is done by setting the number of vehicles arriving during the period R + tc equal to the number of vehicles departing during the period t_c, or:

$$\mathrm{v}(\mathrm{R}+\mathrm{t_c}) = \mathrm{st_c}$$

$$\mathrm{R}+\mathrm{t_c} = \left(\frac{\mathrm{s}}{\mathrm{v}}\right)\mathrm{t_c}$$

$$\mathrm{R} = \mathrm{t_c}\left[\left(\frac{\mathrm{s}}{\mathrm{v}}\right) - 1\right]$$

$$\mathrm{t_c} = \frac{\mathrm{R}}{\left[\left(\frac{\mathrm{s}}{\mathrm{v}}\right) - 1\right]}$$

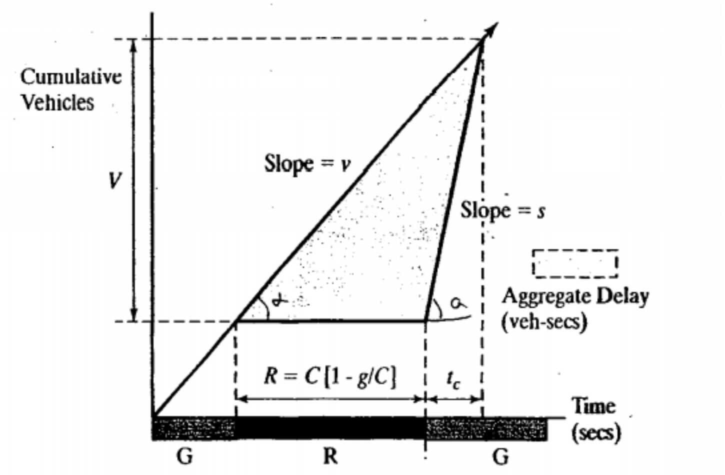


Figure 13: Webster's Uniform Delay Model Illustrated

Then, substituting for tc:

$$V = v(R+t_c) = v\left[R + \frac{R}{\frac{s}{v} - 1}\right] = R\left(\frac{vs}{s-v}\right)$$

And for R:

$$V = C[1-({}^{g}/_{C})]\left[\frac{vs}{s-v}\right]$$

Then, aggregate delay can be stated as:

$$UD_a = \frac{1}{2}RV = \frac{1}{2}C^2[1-({}^{g}/_{C})]\left[\frac{vs}{s-v}\right] \qquad 18$$

Where all variables are as previously defined.

Equation 18 estimates aggregate uniform delay in vehicle-seconds for one single cycle. To get an estimate of average uniform delay per vehicle, the aggregate is divided by the number of vehicles arriving during the cycle, vC. Then:

$$UD = \frac{1}{2}C\,\frac{[1-({}^{g}/_{C})]^2}{[1-{}^{v}/_{s}]} \qquad 19$$

Another form of the equation uses the capacity, c, rather than the saturation low rate, s. noting that s = c/ (g/C), the following form emerges:

$$UD = \frac{1}{2}C\,\frac{[1-({}^{g}/_{C})]^2}{[1-({}^{g}/_{C})({}^{v}/_{c})]} = \frac{0.5C[1-({}^{g}/_{C})]^2}{1-({}^{g}/_{C})X} \qquad 20$$

<u>Where:</u>
UD= average uniform delay per vehicles, s/veh.
C = cycle length, s. g = effective green time, s.
v = arrival low rate, veh/h.
c = capacity of intersection approach, veh/h.
X = v/c ratio, or degree of saturation.

This average includes the vehicles that arrive and depart on green, accruing no delay. This is appropriate. One of the objectives in signalizations is to minimize the number or proportion of vehicles that must stop. Any meaningful quality measure would have to include the positive impact of vehicles that are not delayed.
In Equation 20, note that the maximum value of X (the v/c ratio) is 1.00. As the uniform delay model assumes no overflow, the v/c ratio cannot be more than 1.00.

Modeling Random Delay

The uniform delay model assumes that arrivals are uniform and that no signal phases fail (i.e., that arrival flow is less than capacity during every signal cycle of the analysis period). At isolated intersections, vehicle arrivals are more likely to be random. A number of stochastic models have been developed for this case, including those by Newall, Miller, and Webster. Such models assume that inter-vehicle arrival times are distributed acceding to the Poisson distribution with an underlying average rate of v vehicles/unit time. The models account for both the underlying randomness of arrivals and the fact that some individual cycles within a demand period with v/c < 1.00 could fail due to this randomness. This additional delay is sometimes referred to as "overflow delay," but it does not address situations in which v/c > 1.00 for the entire analysis period. This text refers to additional delay due to randomness as "random delay," RD, to distinguish it from true overflow delay when v/c > 1.00. The most frequently used model for random delay is Webster formulation:

$$RD = \frac{X^2}{2v(1-X)} \tag{21}$$

Where:
RD = average random delay per vehicle, s/veh
X = v/c ratio

This formulation was found to somewhat overestimate delay, and Webster proposed that total delay (the sum of uniform and random delay) be estimated:

$$D = 0.90(UD + RD) \tag{22}$$

Where:
D= sum of uniform and random delay.

Modeling Overflow Delay

“Oversaturation” is used to describe the extended time periods during which arriving vehicles exceed the capacity of the intersection approach to discharge vehicles. In such cases, queues grow, and overflow delay, in addition to uniform delay, accrues. Because overflow delay accounts for the failure of an extended series of phases, it encompasses a portion of random delay as well.
Figure 14 illustrates a time period for which v/c > 1.00. Again, as in the uniform delay model it is assumed the arrival function is uniform. During the period of oversaturation, delay consists of both uniform delay (in the triangles between the capacity and departure curves) and overflow delay (in the growing triangle between the arrival and capacity curves). The formula for the uniform delay component may be simplified in this case because the v/c ratio (X) is the maximum value of 1.00 for the uniform delay component. Then:

$$UD_o = \frac{0.50C\left[1-\left({}^{g}/_{C}\right)\right]^2}{1-\left({}^{g}/_{C}\right)1.00} \tag{23}$$

$$= 0.5\,C\left[1-\left({}^{g}/_{C}\right)\right]$$

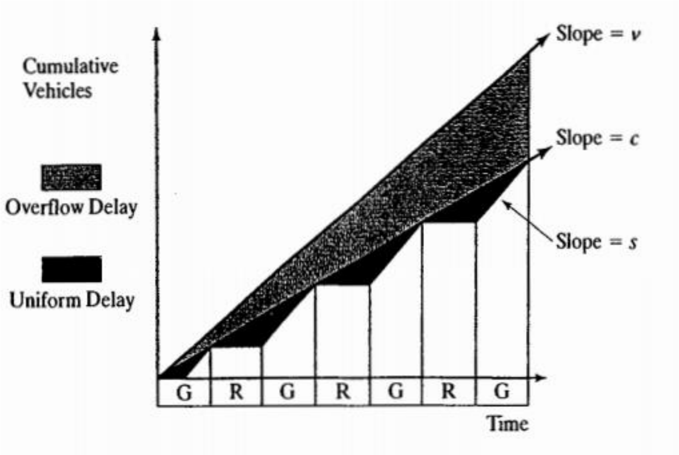


Figure 14: An Oversaturated Period Illustrated.

To this, the overflow delay must be added. Figure 15 illustrates how the overflow delay is estimated. The aggregate and average overflow delay can be estimated as:

$$OD_a = \frac{1}{2}T(vT - cT) = \frac{T^2}{2}(v - c) \qquad 24$$

$$OD = \frac{T}{2}[X - 1]$$

Where:

OD_a = aggregate overflow delay, veh-sec
OD = average overflow delay per vehicle, s/veh
Other parameters are as previously defined.

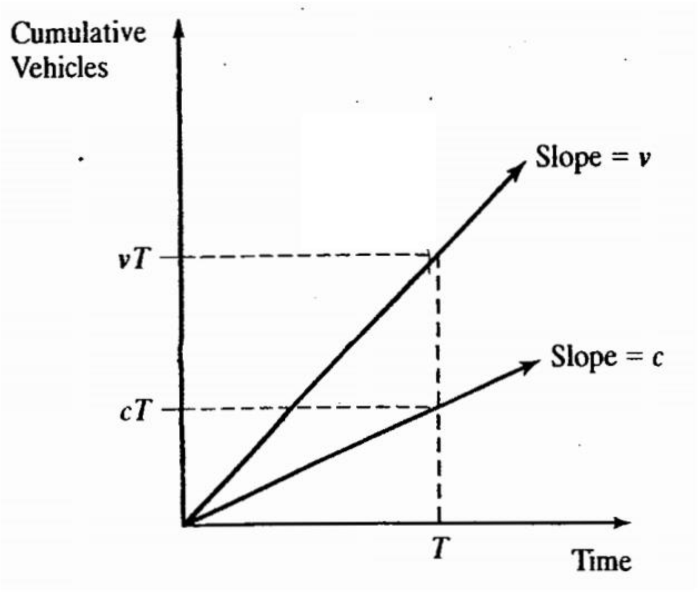


Figure 15: Deviation of the Overflow Delay Formula.

In Equations 24, the average overflow delay is obtained by dividing the aggregate delay by the number of vehicles discharged within time T, cT. Unlike the formulation for uniform delay, where the number of vehicles arriving and the number of vehicles discharged during a cycle were the same, the overflow delay triangle includes vehicle that arrive within time 7 but are not discharged within time T. The delay triangle, therefore, includes only the delay accrued by vehicles though time and excludes additional delay that vehicles still "stuck" in the queue will experience after time T.

Equation 24 may use any unit of time for "T." The resulting overflow delay, OD, will have the same units as specified for T on a per-vehicle basis.

Equations 24 are time dependent (i.e., the longer the period of oversaturation exists, the larger delay becomes).

The predicted delay per vehicle is averaged over the entire period of oversaturation, T. This masks, however, a significant issue: Vehicles arriving early during time T experience far less delay than vehicles arriving later during time T. A model of average overflow delay during a time period T_1 through T_2 may be developed, as illustrated in Figure 16. Note that the delay area formed is a trapezoidal not a triangle.

The resulting model for average delay per vehicle during the time period T_1 through T_2 is:

$$OD = \frac{T_1+T_2}{2}(X-1) \qquad 25$$

Where all terms are as previously defined.

Note that that trapezoidal shape of the delay area results in the T_1+T_2 formulation, emphasizing the growth of delay as the oversaturated condition continues over time. Also, this formulation predicts the average delay per vehicle that occurs during the specified interval, T_1 through T_2 delays to vehicles arriving before time T_1 but discharging after T_2 are included only to the extent of their delay within the specified times, not any delay they may have experienced in queue before T_1 Similarly, vehicles discharging after T_2 do have a delay component after T_2 that is not included in the formulation.

The three varieties of delay-uniform, random, and overflow delay-can be modeled in relatively simple terms as long as simplifying assumptions are made in terms of arrival and discharge flows, and in the nature of the queuing that occurs, particularly during periods of oversaturation. The next section begins to consider some of the complications that raise from tie direct use of these simplified models.

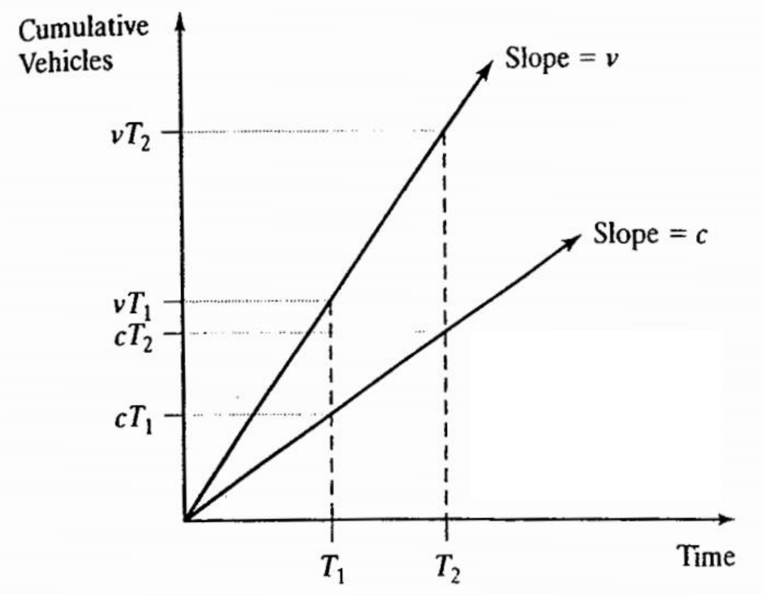


Figure 16: A Model for Overflow Delay between Times T1 andT2.

Inconsistencies in Random and Overflow Delay

Figure 17 illustrates a basic inconsistency in the random md overflow delay models previously discussed. The inconsistency occurs when the v/c ratio (X) is in the vicinity of 1. 00. When the v/c ratio is below 1.00, a random delay model is used because there is no "overflow" delay in this case. Webster's random delay model (Equation 22), however, contains the term (l-X) in the denominator. Thus as X approaches a value of 1. 00, random delay increases asymptotically to an infinite value. When the v/c ratio (X) is greater than 1.00, an overflow delay model is applied. The overflow delay model of Equation 24, however, has an overflow delay of 0 when X= 1.00, and increases uniformly with increasing values of X thereafter.

Neither model is accurate in the immediate vicinity of vie = 1.00. Delay does not become infinite at v/c = 1. 00. There is no true "overflow" at v/c = 1.00, although individual cycle failures due to random arrivals do occur. Similarly, the overflow model, with overflow delay = 0. 0 s/veh. at v/c = 1.00, is also unrealistic. The additional delay of individual cycle failures due to the randomness of arrivals is not reflected in this model.

In practical-terms, most studies confirm that the uniform delay model is a sufficient predictive tool (except for the issue of platooned arrivals) when the v/c ratio is 0. 85 or less. In this range, the true value of random delay is minuscule, and there is no overflow delay. Similarly, the simple theoretical overflow delay model (when added to uniform delay) is a reasonable predictor when v/c > 1. 15 or so. The problem is that the most interesting cases fall in the intermediate range (0.85 < v/c < 1.15), for which neither model is adequate. Much of the more recent work in delay modeling involves attempts to bridge this gap, creating a model that closely follows the " uniform delay model at low v/c ratios and approaches the theoretical overflow delay model at high v/c ratios (≥ 1.15), producing “reasonable” delay estimates in between. Figure 17 illustrates this as the dashed line.

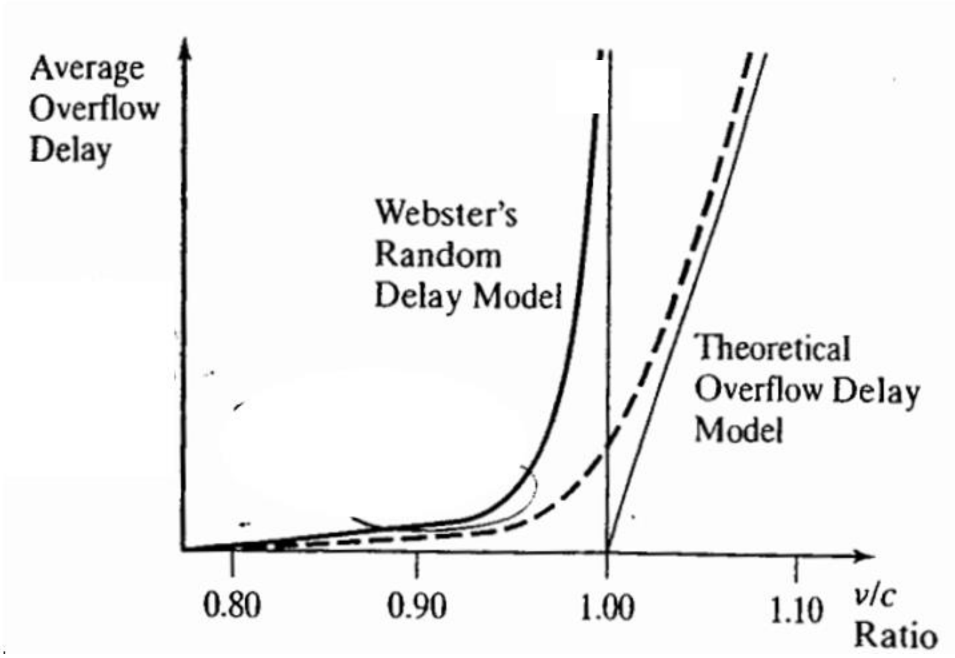


Figure 17: Random and Overflow Delay Models Compared.

The most commonly used model for bridging this gap was developed by Akceik for the Australian Road research Board's signalized intersection analysis procedure:

$$OD = \frac{cT}{4}\left[(X-1) + \sqrt{(X-1)^2 + \left(\frac{12(X-X_0)}{cT}\right)}\right] \tag{26}$$

$$X_0 = 0.67 + \left(\frac{sg}{600}\right)$$

<u>Where:</u>
T = analysis period, h
X = v/c ratio
c= capacity, veh/h
s= saturation flow rate, veh/sg (veh/s of green)
g = effective green time, s

The only relatively recent study resulting in large amounts of delay measurements in the field was conducted by Reilly et al. In the early 1980s to calibrate a model for use in the 1985 edition of the Highway Capacity Manual. The study concluded that Equation 17-26 substantially overestimated field measured values of delay and recommended that a factor of0.50 be included in the model to adjust for this. The version of the delay equation that was included in the 1985 Highway Capacity

Manual ultimately did not follow this recommendation and included other empirical adjustments to the theoretical equation.

Delay Models in the HCM

The delay model incorporated into the HCM 2000 includes the uniform delay model, a version of Akcelik's overflow delay model, and a term covering delay from an existing or residual queue at the beginning of the analysis period. The model is:

$$d = d_1 PF + d_2 + d_3 \qquad 27$$

Where:
d = control delay, s/veh
d_1 = uniform delay component, s/veh c
PF = progression adjustment factor di = overflow delay component, s/veh
d_3 = delay due to preexisting queue, s/veh

The progression factor was an empirically calibrated adjustment to uniform delay that accounts for the effect of platooned arrival patterns. This adjustment is discussed in greater detail in Chapter 24. The delay due to preexisting queues, d$_3$, is found using a relatively complex model (sec Chapter 24).

A significant revision has been included in the forthcoming HCM 2010. Traditional delay models have been replaced by Incremental Queue Analysis (IQA). Chapter 24 contains a more detailed discussion and presentation of this approach.

In the final analysis, all delay modeling is based on the determination of the area between an arrival curve and a departure curve on a plot of cumulative vehicles versus time. As the arrival and departure functions are permitted to become more complex and as rates are permitted to vary for various subparts of the signal cycle, the models become more complex as well.