Gap Acceptance and Theory

In traffic it happens rather frequently that a participant (a car driver, a pedestrian, a cyclist) has to 'use' a gap in another traffic stream to carry out a manoeuvre.

Examples are: crossing a street as a pedestrian; overtaking at a road with oncoming vehicles; entering a roundabout where the circulating vehicles have priority; entering a motorway from an on-ramp; a lane change on a motorway; etc.

Another important aspect of traffic flow is the interaction of vehicles as they join, leave, or cross a traffic stream. Examples of these include ramp vehicles merging onto an expressway stream, freeway vehicles leaving the freeway onto frontage roads, and the changing of lanes by vehicles on a multilane highway.

The most important factor a driver considers in making any one of these manoeuvres is the availability of a gap between two vehicles that, in the driver's judgment, is adequate for him or her to complete the manoeuvre. The evaluation of available gaps and the decision to carry out a specific manoeuvre within a particular gap are inherent in the concept of gap acceptance. Following are the important measures that involve the concept of gap acceptance:

1. Merging is the process by which a vehicle in one traffic stream joins another traffic stream moving in the same direction, such as a ramp vehicle joining a freeway stream.

2. Diverging is the process by which a vehicle in a traffic stream leaves that traffic stream, such as a vehicle leaving the outside lane of an expressway.

3. Weaving is the process by which a vehicle first merges into a stream of traffic, obliquely crosses that stream, and then merges into a second stream moving in the same direction; for example, the manoeuvre required for a ramp vehicle to join the far side stream of flow on an expressway.

4. Gap is the headway in a major stream, which is evaluated by a vehicle driver in a minor stream who wishes to merge into the major stream. It is expressed either in units of time (time gap) or in units of distance (space gap).

5. Time lag is the difference between the time a vehicle that merges into a main traffic stream reaches a point on the highway in the area of merge and the time a vehicle in the main stream reaches the same point.

6. Space lag is the difference, at an instant of time, between the distance a merging vehicle is away from a reference point in the area of merge and the distance a vehicle in the main stream is away from the same point.

Figure 1 depicts the time-distance relationships for a vehicle at a stop sign waiting to merge and for vehicles on the near lane of the main traffic stream.

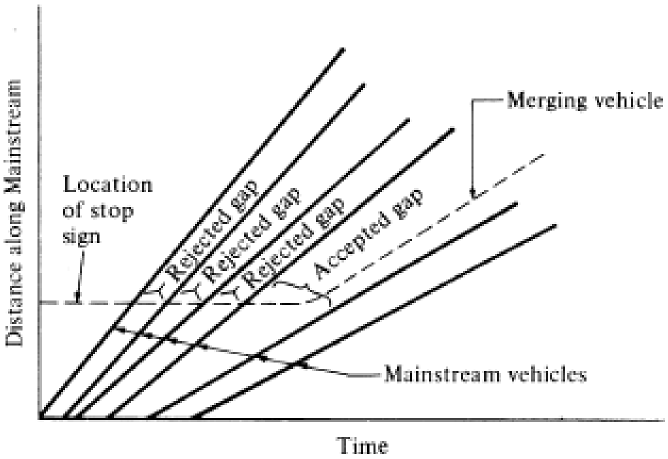

Figure 1 Time-Space Diagrams for Vehicles in the Vicinity of a Stop Sign.

A driver who intends to merge must first evaluate the gaps that become available to determine which gap (if any) is large enough to accept the vehicle, in his or her opinion. In accepting that gap, the driver feels that he or she will be able to complete the merging manoeuvre and safely join the main stream within the length of the gap.

This phenomenon is generally referred to as *gap acceptance*. It is of importance when engineers are considering the delay of vehicles on minor roads wishing to join a major road traffic stream at unsignalized intersections, and also the delay of ramp vehicles wishing to join expressways. It can also be used in timing the release of vehicles at an on ramp of an expressway, such that the probability of the released vehicle finding an acceptable gap in arriving at the freeway shoulder lane is maximum.

To use the phenomenon of gap acceptance in evaluating delays, waiting times, queue lengths, and so forth, at unsignalized intersections and at on-ramps, the average minimum gap length that will be accepted by drivers should be determined first. Several definitions have been given to this “critical” value. Greenshields referred to it as the “acceptable average minimum time gap” and defined it as the gap accepted by 50 percent of the drivers. The concept of “critical gap” was used by Raff, who defined it as the gap for which the number of accepted gaps shorter than it is equal to the number of rejected gaps longer than it. The data in Table 1 are used to demonstrate the determination of the critical gap using Raff’s definition. Either a graphical or an algebraic method can be used.

Table 1 Computation of Critical Gap (t_c)

(a) Gaps Accepted and Rejected

1 Length of Gap (t sec)	*2* Number of Accepted Gaps (less than t sec)	*3* Number of Rejected Gaps (greater than t sec)
0.0	0	116
1.0	2	103
2.0	12	66
3.0	$m = 32$	$r = 38$
4.0	$n = 57$	$p = 19$
5.0	84	6
6.0	116	0

(b) Difference in Gaps Accepted and Rejected

1 Consecutive Gap Lengths (t sec)	*2* Change in Number of Accepted Gaps (less than t sec)	*3* Change in Number of Rejected Gaps (greater than t sec)	*4* Difference Between Columns 2 and 3
0.0–1.0	2	13	11
1.0–2.0	10	37	27
2.0–3.0	20	28	8
3.0–4.0	25	19	6
4.0–5.0	27	13	14
5.0–6.0	32	6	26

In using the graphical method, two cumulative distribution curves are drawn as shown in Figure 2. One relates gap lengths *t* with the number of accepted gaps less than *t*, and the other relates *t* with the number of rejected gaps greater than *t*. The intersection of these two curves gives the value of *t* for the critical gap.

In using the algebraic method, it is necessary to first identify the gap lengths between where the critical gap lies. This is done by comparing the change in number of accepted gaps less than *t* sec (column 2 of Table1) for two consecutive gap lengths, with the change in number of rejected gaps greater than *t* sec (column 3 of Table 1) for the same two consecutive gap lengths. The critical gap length lies between the two consecutive gap lengths where the difference between the two changes is minimal. Table 1 shows the computation and indicates that the critical gap for this case lies between 3 and 4 seconds.

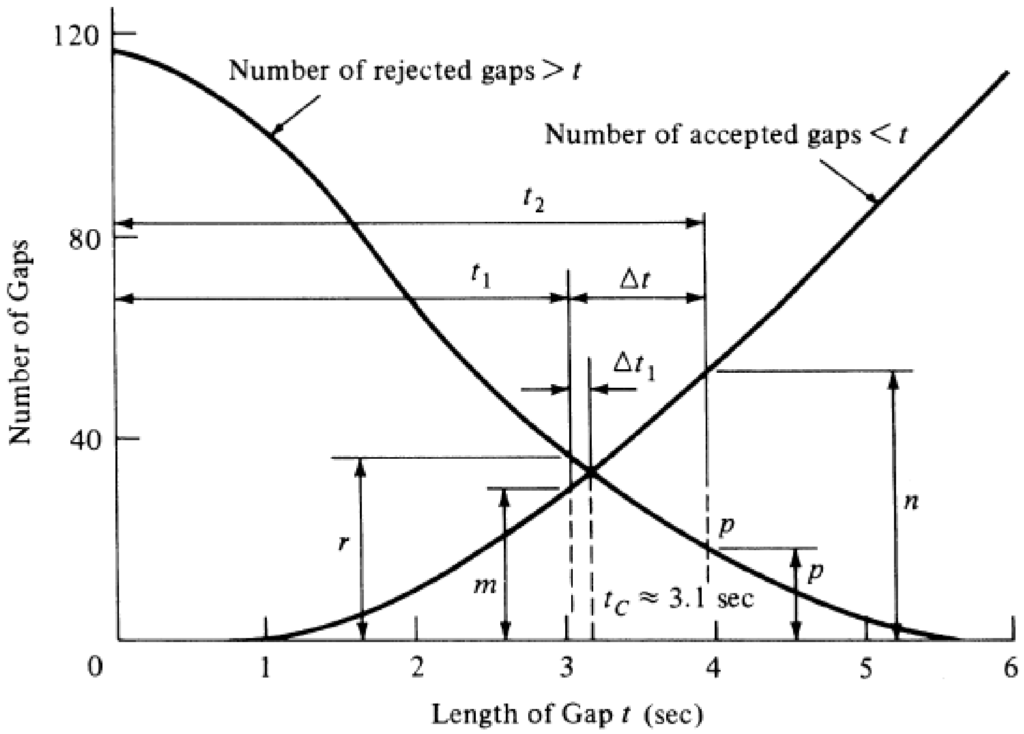

Figure 2 Cumulative Distribution Curves for Accepted and Rejected Gaps.

For example, in Figure 2, with Δt equal to the time increment used for gap analysis, the critical gap lies between t_1 and $t_2 = t_1 + \Delta t$

Where:

m = number of accepted gaps less than t_1
r =number of rejected gaps greater than t_1
n = number of accepted gaps less than t_2
p = number of rejected gaps greater than t_2

Assuming that the curves are linear between t_1 andt_2, the point of intersection of these two lines represents the critical gap. From Figure 2, the critical gap expression can be written as:

$t_c = t_1 + \Delta t_1$

Using the properties of similar triangles,

$$\frac{\Delta t_1}{r - m} = \frac{\Delta t - \Delta t_1}{n - p}$$

$$\Delta t_1 = \frac{\Delta t(r - m)}{(n - p) + (r - m)}$$

$$t_c = t_1 + \frac{\Delta t(r - m)}{(n - p) + (r - m)}$$

For the data given in Table 1, we thus have

$$t_c = 3 + \frac{1(38 - 32)}{(57 - 19) + (38 - 32)} = 3 + \frac{6}{38 + 6}$$
$$\approx 3.14 \text{ sec}$$

Stochastic Approach to Gap and Gap Acceptance Problems

The use of gap acceptance to determine the delay of vehicles in minor streams wishing to merge onto major streams requires a knowledge of the frequency of arrivals of gaps that are at least equal to the critical gap. This in turn depends on the distribution of arrivals of mainstream vehicles at the area of merge. It is generally accepted that for light to medium traffic flow on a highway, the arrival of vehicles is randomly distributed.
It is therefore important that the probabilistic approach to the subject be discussed. It is usually assumed that for light-to-medium traffic the distribution is Poisson, although assumptions of gamma and exponential distributions have also been made.
Assuming that the distribution of mainstream arrival is Poisson, then the probability of *x* arrivals in any interval of time *t* sec can be obtained from the expression

$$P(x) = \frac{\mu^x e^{-\mu}}{x!} \qquad (\text{for } x = 0, 1, 2 \ldots, \infty)$$

P(x2) = the probability of x vehicles arriving in time t sec.
μ = average number of vehicles arriving in time t.

If V represents the total number of vehicles arriving in time T sec, then the average number of vehicles arriving per second is

$$\lambda = \frac{V}{T} \qquad \mu = \lambda t$$

We can therefore write the Equation 1 below:

$$P(x) = \frac{\mu^x e^{-\mu}}{x!} \qquad (\text{for } x = 0, 1, 2 \ldots, \infty)$$

Now consider a vehicle at an unsignalized intersection or at a ramp waiting to merge into the mainstream flow, arrivals of which can be described by Eq. 1. The minor stream vehicle will merge only if there is a gap of *t* sec equal to or greater than its critical gap. This will occur when no vehicles arrive during a period *t* sec long.
The probability of this is the probability of zero cars arriving (that is, when x in Eq. 6.51 is zero). Substituting zero for x in Eq. 1 will therefore give a probability of a gap ($h \geq t$) occurring. Thus Equation 2 and 3 below,

$$P(0) = P(h \geq t) = e^{-\lambda t} \qquad \text{for } t \geq 0$$

$$P(h < t) = 1 - e^{-\lambda t} \qquad \text{for } t \geq 0$$

$$P(h < t) + P(h \geq t) = 1$$

It can be seen that t can take all values from 0 to∞, which therefore makes Eqs. 2 and 3 continuous functions. The probability function described by Eq. 2 is known as the exponential distribution.
Equation 2 can be used to determine the expected number of acceptable gaps that will occur at an unsignalized intersection or at the merging area of an expressway on ramp during a period T, if the Poisson distribution is assumed for the mainstream flow and the volume V is also known. Let us assume that T is equal to 1 hr and that V is the volume in veh/h on the mainstream flow. Since (V-1) gaps occur between V successive vehicles in a stream of vehicles, then the expected number of gaps greater or equal to t is given as in Equation 4 and 5 below:

$$\text{Frequency } (h \geq t) = (V - 1)e^{-\lambda t}$$

And the expected number of gaps less than t is given as:

$$\text{Frequency } (h < t) = (V - 1)(1 - e^{-\lambda t})$$

The basic assumption made in this analysis is that the arrival of mainstream vehicles can be described by a Poisson distribution. This assumption is reasonable for light-to-medium traffic but may not be acceptable for conditions of heavy traffic.
Analyses of the occurrence of different gap sizes when traffic volume is heavy have shown that the main discrepancies occur at gaps of short lengths (that is, less than 1 second). The reason for this is that although theoretically there are definite probabilities for the occurrence of gaps between 0 and 1 seconds, in reality these gaps very rarely occur, since a driver will tend to keep a safe distance between his or her vehicle and the vehicle immediately in front.
One alternative used to deal with this situation is to restrict the range of headways by introducing a minimum gap. Equations 4 and 5 can then be written as Equations 6 and 7 below:

$$P(h \geq t) = e^{-\lambda(t-\tau)} \qquad (\text{for } t \geq 0)$$

$$P(h < t) = 1 - e^{-\lambda(t-\tau)} \qquad (\text{for } t \leq 0)$$

Where:
τ is the minimum headway.

Queuing Theory

One of the greatest concerns of traffic engineers is the serious congestion that exists on urban highways, especially during peak hours. This congestion results in the formation of queues on expressway on ramps and off ramps, at signalized and unsignalized intersections, and on arterials, where moving queues may occur. An understanding of the processes that lead to the occurrence of queues and the subsequent delays on highways is essential for the proper analysis of the effects of queuing.
The theory of queuing therefore concerns the use of mathematical algorithms to describe the processes that result in the formation of queues, so that a detailed analysis of the effects of queues can be undertaken. The analysis of queues can be undertaken by assuming either deterministic or stochastic queue characteristics.

Deterministic Analysis of Queues

The deterministic analysis assumes that all the traffic characteristics of the queue are deterministic and demand volumes and capacities are known. There are two common traffic conditions for which the deterministic approach has been used. The first is when an incident occurs on a highway resulting in a significant reduction on the capacity of the highway. This can be described as a varying service rate and constant demand condition. The second is significant increase in demand flow exceeding the capacity of a section of highway which can be described as a varying demand and constant service rate condition.

1. Varying Service Rate and Constant Demand

Consider a section of three-lane (one-direction) highway with a capacity of c veh/h, i.e., it can serve a maximum volume of c veh/h. An incident occurs which resulted in the closure of one lane thereby reducing its capacity to c_R for a period of t hr, which is the time it takes to clear the incident. The demand volume continues to be V veh/h throughout the period of the incident as shown in Figure 3. The demand volume is less than the capacity of the highway section but greater than the reduced capacity. Before the incident, there is no queue as the demand volume is less than the capacity of the highway. However, during the incident the demand volume is higher than the reduced capacity resulting in the formation of a queue as shown in Figure 3b. Several important parameters can be determined to describe the effect of this reduction in the highway capacity. These include the maximum queue length, duration of the queue, average queue length, maximum individual delay, time a driver spends in the queue, average queue length while the queue exists, maximum individual delay, and the total delay.

The maximum queue length (q_{max}) is the excess demand rate multiplied by the duration of the incident and is given as:

$$q_{\max} = (v - c_R)t \text{ vehicles}$$

The time duration of the queue (t_q) is the queue length divided by the difference between the capacity and the demand rate and is given as

$$t_q = \frac{(c - c_R)t}{(c - v)} \text{ hr}$$

The average queue length is

$$q_{\text{av}} = \frac{(v - c_R)t}{2} \text{ veh}$$

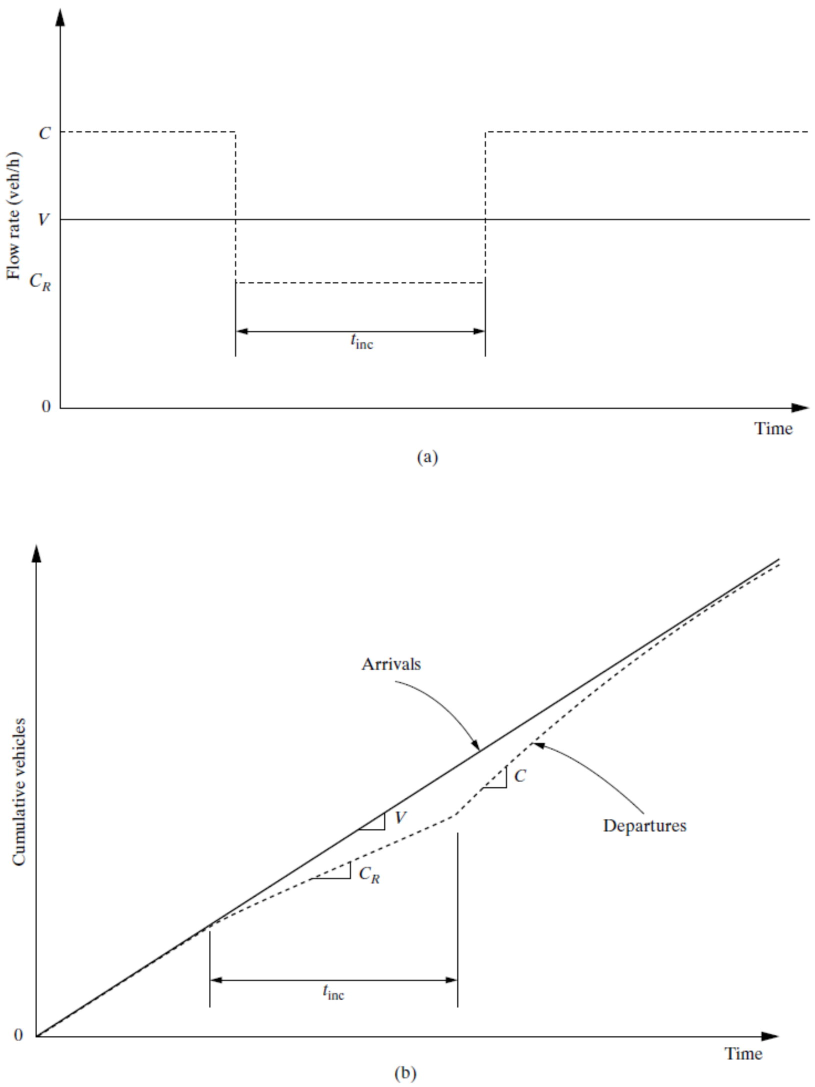

Figure 3 Queuing Diagram for Incident Situation.

The total delay (d_t) is the time duration of the queue multiplied by the average queue length and is given as:

$$d_T = \frac{(v - c_R)t}{2}\frac{(c - c_R)t}{(c - v)} = \frac{t^2(v - c_R)(c - c_R)}{2(c - v)} \text{ hr}$$

2. Varying Demand and Constant Service Rate

The procedure described in the previous section also can be used for varying demand and constant service rate, if it is assumed that the demand changes at specific times and not gradually increasing or decreasing.

Stochastic Analyses of Queues

Using a stochastic approach to analyze queues considers the fact that certain traffic characteristics such as arrival rates are not always deterministic. In fact, arrivals at an intersection for example are deterministic or regular only when approach volumes are high. Arrival rates tend to be random for light to medium traffic. The stochastic approach is used to determine the probability that an arrival will be delayed, the expected waiting time for all arrivals, the expected waiting time of an arrival that waits, and so forth.

Several models have been developed that can be applied to traffic situations such as the merging of ramp traffic to freeway traffic, interactions at pedestrian crossings, and sudden reduction of capacity on freeways. This section will give only the elementary queuing theory relationships for a specific type of queue; that is, the single channel queue. The theoretical development of these relationships is not included here. Interested readers may refer to any traffic flow theory book for a more detailed treatment of the topic.

A queue is formed when arrivals wait for a service or an opportunity, such as the arrival of an accepted gap in a main traffic stream, the collection of tolls at a tollbooth or of parking fees at a parking garage, and so forth. The service can be provided in a single channel or in several channels. Proper analysis of the effects of such a queue can be carried out only if the queue is fully specified. This requires that the following characteristics of the queue be given:

- The characteristic distribution of arrivals, such as uniform, Poisson, and so on;
- The method of service, such as first come–first served, random, and priority;
- The characteristic of the queue length, that is, whether it is finite or infinite;
- The distribution of service times; and
- The channel layout, that is, whether there are single or multiple channels and, in the case of multiple channels, whether they are in series or parallel. Several methods for the classification of queues based on the above characteristics have been used—some of which are discussed below.

Arrival Distribution. The arrivals can be described as either a deterministic distribution or a random distribution. Light-to-medium traffic is usually described by a Poisson distribution, and this is generally used in queuing theories related to traffic flow.

Service Method. Queues also can be classified by the method used in servicing the arrivals. These include first come–first served where units are served in order of their arrivals, and last in-first served, where the service is reversed to the order of arrival. The service method can also be based on priority, where arrivals are directed to specific queues of appropriate priority levels—for example, giving priority to buses. Queues are then serviced in order of their priority level.

Characteristics of the Queue Length. The maximum length of the queue, that is, the maximum number of units in the queue, is specified, in which case the queue is a finite or truncated queue, or else there may be no restriction on the length of the queue. Finite queues are sometimes necessary when the waiting area is limited.

- ***Service Distribution.*** The Poisson and negative exponential distributions have been used as the random distributions.

- ***Number of Channels.*** The number of channels usually corresponds to the number of waiting lines and is therefore used to classify queues, for example, as a single channel or multi-channel queue.

- ***Oversaturated and Undersaturated Queues.*** Oversaturated queues are those in which the arrival rate is greater than the service rate, and undersaturated queues are those in which the arrival rate is less than the service rate. The length of an undersaturated queue may vary but will reach a steady state with the arrival of units. The length of an oversaturated queue, however, will never reach a steady state but will continue to increase with the arrival of units.

- Single-Channel, Undersaturated, Infinite Queues

Figure 4 is a schematic of a single-channel queue in which the rate of arrival is q veh/h and the service rate is Q veh/h. For an undersaturated queue, $Q > q$, assuming that both the rate of arrivals and the rate of service are random, the following relationships can be developed:

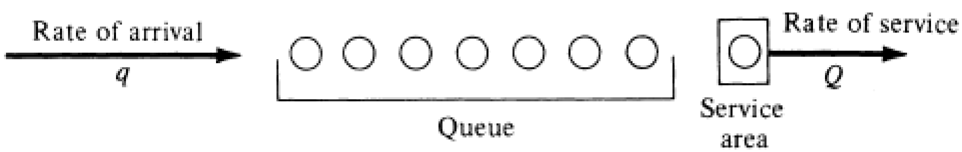

Figure 4 A Single-Channel Queue.

1. Probability of n units in the system, P (n):

$$P(n) = \left(\frac{q}{Q}\right)^n \left(1 - \frac{q}{Q}\right)$$

Where n is the number of units in the system, including the unit being serviced.

2. The expected number of units in the system, E (n):

$$E(n) = \frac{q}{Q - q}$$

3. The expected number of units waiting to be served (that is, the mean queue length) in the system, *E* (*m*):

$$E(m) = \frac{q^2}{Q(Q - q)}$$

Note that E (m) is not exactly equal to E (n) _ 1, the reason being that there is a definite probability of zero units being in the system, P (0).

4. Average waiting time in the queue, *E* (*w*):

$$E(w) = \frac{q}{Q(Q - q)}$$

5. Average waiting time of an arrival, including queue and service, E (v):

$$E(v) = \frac{1}{Q - q}$$

6. Probability of spending time *t* or less in the system:

$$P(v \leq t) = 1 - e^{-(1-\frac{q}{Q})qt}$$

7. Probability of waiting for time t or less in the queue:

$$P(w \leq t) = 1 - \frac{q}{Q}e^{-(1-\frac{q}{Q})qt}$$

8. Probability of more than N vehicles being in the system, that is, P (n > N):

$$P(n > N) = \left(\frac{q}{Q}\right)^{N+1}$$

Figure 4 is such a representation for different values of r. It should be noted that as this ratio tends to 1 (that is, approaching saturation), the expected number of vehicles in the system tends to infinity. This shows that q/Q, which is usually referred to as the traffic intensity, is an important factor in the queuing process. The figure also indicates that queuing is of no significance when r is less than 0.5, but at values of 0.75 and above, the average queue lengths tend to increase rapidly. Figure 6 is also a graph of the probability of n units being in the system versus q/Q.

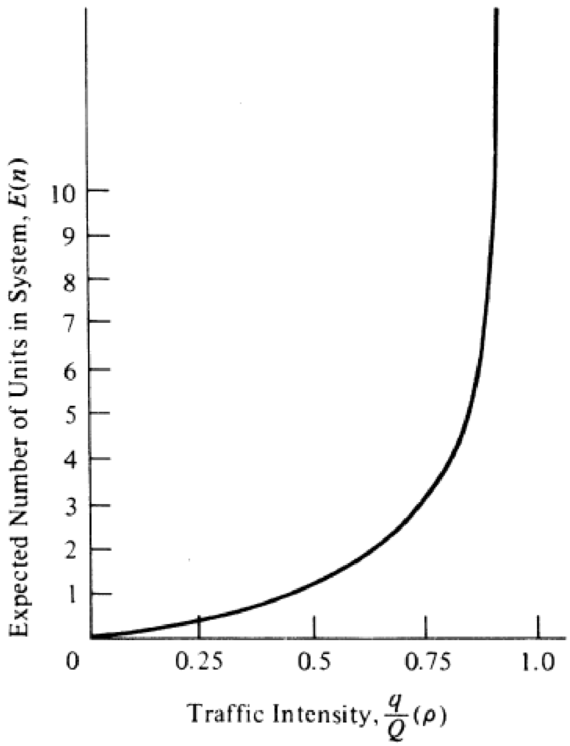

Figure 5 Expected Number of Vehicles in the System *E* (*n*) versus Traffic Intensity (r).

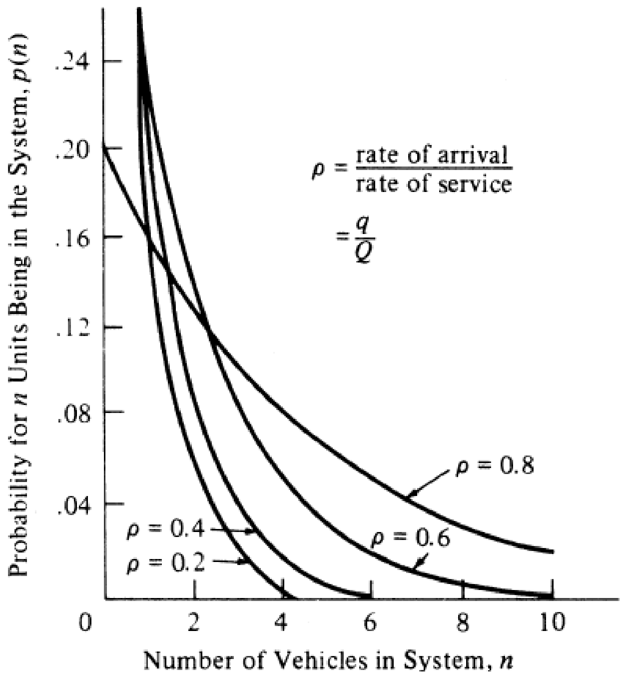

Figure 6 Probability of *n* Vehicles Being in the System for Different Traffic Intensities (r).

- Single-Channel, Undersaturated, Finite Queues

In the case of a finite queue, the maximum number of units in the system is specified. Let this number be *N*. Let the rate of arrival be *q* and the service rate be *Q*. If it is also assumed that both the rate of arrival and the rate of service are random, the following relationships can be developed for the finite queue.

1. Probability of *n* units in the system:

$$P(n) = \frac{1-\rho}{1-\rho^{N+1}}\rho^n$$

Where r = q/Q.

2. The expected number of units in the system:

$$E(n) = \frac{\rho}{1-\rho}\,\frac{1-(N+1)\rho^N + N\rho^{N+1}}{1-\rho^{N+1}}$$