CHAPTER 1

Introduction

1.1 Nature of subject

It is rather difficult to define precisely what 'queueing theory' includes. The classic prototype problem is the following: customers (for example, people) arrive at certain time instants at a service point (at a bank counter, an air ticket counter, a highway intersection, etc.). The service facility requires a certain time to serve each customer but is capable of serving only finitely many customers at a time (for example, one). If customers arrive faster than the facility can serve them, customers must wait in a queue. Typically, both the customer arrivals and the service times are specified to have some given probability distributions. One wishes to relate the delays in queue, queue lengths, etc., to the given properties of the arrivals and service. In practical applications one frequently wishes further to compare the operation of several possible modes of operation with regard to its type of service, cost, etc. Perhaps a service facility performs more than one function, a bank clerk cashes checks and also sells government bonds, an airline clerk sells tickets and checks passengers, etc. Should one have separate facilities for separate functions? If one must have several servers, how many should there be?

Much of the terminology of queueing theory is motivated by the type of applications described above but there are other possible labels. In a production process, articles enter some shop to have a bolt tightened, for example. Articles are stored until the facility is ready to perform its task. The mathematical problem is clearly the same as above. The 'customers' are not people but objects and the 'queue' is a storage. In some cases, the term 'service' also seems inappropriate. Cars queue to cross an intersection because the intersection will pass cars only at a restricted rate. The intersection is 'serving' the cars. The same is true of the landing strip of an airport. Sometimes the 'queue' has no obvious physical location. In telephone service, customers wish to use one of only finitely many trunk lines. Sometimes the system reaches capacity and calls are rejected. Whether customers keep trying or go away, there is no obvious physical queue anywhere. Riordan's book is titled 'Stochastic Service Systems' (he worked for the telephone company) but this book covers essentially the same topics as do books on queueing theory.

In 'queueing theory,' it seems that the usual motive is to keep queues short. The mathematical theory of queues, however, differs only slightly from the 'theory of inventory' and the 'theory of dams.' In inventory theory, there is a source of supply (which, however, one can usually control) and there is an outflow. To add confusion, a 'customer' in inventory theory is usually the consumer of the inventory – the output. Regardless of what one calls these things, however, one has a reservoir, a supply, a queue, an inventory or whatever one wishes to call it. There is an input, and there is an output, and there is a conservation principle – what goes in must come out. In inventory theory the objective is usually to have a non-zero supply (a non-zero queue). The theory of dams and the theory of inventory differ only in that for the former one usually imagines that one can control the output but not the input whereas for the latter the reverse is true. Mathematically, however, there is little distinction even between output and input. If one has a waiting room full of customers, one can think of a service as taking a customer out of the waiting room through the service, or one can imagine a 'hole' or 'anti-customer' as going through the service backwards into the waiting room.

Much of the theory of queues deals with stochastic models but in some of the simple approximations it is convenient to disregard the discrete nature of customers (if they are in fact discrete) and treat them as a continuous fluid. Perhaps the only thing that all the above physical problems have in common is that we have a reservoir, an input, and an output, and a conservation principle. Each physical application, however, has its own peculiarities. It is not surprising then that the literature on queues, dams, inventories, etc., runs into the thousands of papers.

1.2 Mathematical representation

We consider first the most typical class of queueing situations. A customer first arrives and joins a queue (if there is one); he later leaves the queue and enters service and finally leaves the service. Suppose we start our observation of this system at some time $t = 0$ chosen as some time when both the queue and the service are empty. Let

$$0 < t_1 < t_2 \dots$$

be the arrival times of customers to the system. Suppose also we number the customers in order of their arrival.

Whenever there is a queue of more than one customer, the order in which customers enter service need not be the same as the order in which they arrive. Rules relating to order of service are called 'queue disciplines.' The discipline by which customers are served in order of their arrival is usually called 'first in first out' or 'FIFO' or 'first come first served.' This is the simplest mathematically because a labeling of departure times from the queue in the order

$$0 < t_1' < t_2' < \dots$$

is the same as defining t_j' as the departure time for the customer numbered j as above. If the service is not FIFO then, depending upon mathematical convenience, we may wish to identify the times t_j' as departure times of the j^{th} customer and thereby sacrifice the property $t_j' < t_{j+1}'$, or we may wish to label the departure time in increasing order with some subsidiary rule of identification which gives the customer number of the customer that leaves at time t_j', or its inverse, the departure number of the j^{th} customer.

Some common types of discipline other than FIFO are:

(a) First come last served (or is it last come first served?) Suppose a clerk brings letters one at a time and piles them on your desk. The latest letter being placed on top. The letters are processed by taking the one on top. The letters are the 'customers,' the pile is the 'queue' and the service is last come first served. The letter on the bottom is not taken until everything on top is gone.

(b) Priority service. Each customer is assigned a priority class when he arrives. Important customers have high priority and are served ahead of lower priority customers. There are a wide variety of special versions of this depending on whether one interrupts service or not.

(c) Random order. Queues to board a train are perhaps this way. People do not line up as they arrive. If a queue forms when the gate opens, the order in the queue has little relation to the time people may have arrived originally.

Regardless of whether t'_j is the j^{th} ordered departure time or the departure time of the j^{th} ordered arrival,

$$t_j \leq t'_j$$

because a total of j customers cannot have left the queue before j customers have arrived, nor can the j^{th} customer itself depart from the queue before it has arrived. It is also customary to assume that when the system is empty (both the service and the queue) $t_j = t'_j$; i.e., customers enter service immediately if the service is idle.

If the service handles only one customer at a time, then customers must leave the service in the same order in which they enter. If they enter at times $0 < t'_1 < t'_2 \ldots$, then the times of departure t''_j must be ordered so that

$$0 < t'_1 < t''_1 \leq t'_2 < t''_2 < \ldots ;$$

the $j+1^{th}$ customer cannot enter until the j^{th} customer has left. In some practical situations one may wish to attach some special significance to both the t''_j and the t'_{j+1}. In most mathematical analyses of queues it is usually assumed, however, that t''_j and t'_{j+1} are equivalent if by time t''_j the $j+1^{th}$ customer has arrived. One could, for example, redefine the behavior of the service and imagine that as soon as customer j leaves, customer $j+1$ is immediately considered to have entered some hypothetical service. Or one could also redefine the completion times and arbitrarily say that customer j has not really left until the service is ready to serve customer $j+1$.

If a service can handle more than one customer at a time, then it is no longer true that the departure times t''_j are ordered in the same way as the entrance times t'_j. A typical example of this is a check-out counter at a grocery store with several cashiers. A customer which arrives at one cashier may arrive before another customer at another cashier, but if the former has a long service time he may leave the service later than the second customer. A service facility which can serve m customers simultaneously is called in the queue literature an 'm-channel' facility.

Another example of a multiple-channel server is a parking lot with m stalls. The identification is less obvious here, perhaps, because one does not ordinarily think of a parking stall as performing a 'service.' Customers (cars) arrive at the lot and perhaps wait for an empty spot. Once they have entered the service (parked), they remain there for some length of time which probably has little connection with the properties of the lot. When they leave, the channel is available for another customer. There is clearly no reason why customers should leave in the same order as they arrive.

A mathematical description of a queueing facility consists of any set of rules whereby the times t'_j and t''_j, and perhaps also the customer identities, can be evaluated from given values of the t_j. These rules could conceivably be quite complicated and involve complex interrelations between service times, arrival times, etc., restricted only by the universal principle that a customer cannot leave before he has arrived, which in turn guarantees that the number of customers in queues or in service is non-negative.

Most systems which are analysed in the queueing theory literature have very simple rules for the evolution of the queue. The mathematical complications are not directly associated with the queue mechanism, but with the stochastic analysis. Even though the postulated relations between arrival times and departure times appear quite simple, they lead to fairly complex relations between probability distributions for arrivals, departures, queue lengths, waits, etc.

In a typical single channel queueing problem, one is given the arrival time of customers t_j or their differences

$$\tau_j = t_j - t_{j-1}, \tag{2.1}$$

the interarrival times (actually one is given the probability distributions for the τ_j), and one is given the service times

$$s_j = t''_j - t'_j \tag{2.2}$$

(or their probability distributions). One also specifies a queue discipline and some times δ_j (usually zero) between the completion of the j^{th} service t''_j and the time the service is ready to accept customer $j+1$ (if he has arrived). If the $j+1^{th}$ customer has not arrived by time $t''_j+\delta_j$, he is assumed to enter service immediately upon his arrival at time t_{j+1}. Thus

$$t'_{j+1} = \max\,(t_{j+1}, t''_j+\delta_j) = \max\,(t_{j+1}, t'_j+s_j+\delta_j). \qquad (2.3)$$

Equations (2.1), (2.2), and (2.3), along with a specification of the arrival time t_1 of the first customer and that the system was empty then ($t'_1 = t_1$), represent a formal description of the system. From (2.1) one can sequentially evaluate $t_2, t_3, \ldots$ from given values of the τ_j and t'_1. From (2.3), one can sequentially evaluate $t'_2, t'_3, \ldots$ from given values of the $s_j+\delta_j$ and t'_1, and the previously evaluated t_j. The t''_j can then be evaluated from the s_j and t'_j.

One can also evaluate t_j explicitly as

$$t_j = t_1 + \sum_{k=2}^{j} \tau_k.$$

If it weren't for the fact that a customer cannot be served until he arrives, there would be no coupling between the arrivals and departures, and one could also evaluate the t'_j in the same way

$$t'_j = t'_1 + \sum_{k=1}^{j-1} (s_k+\delta_k), \text{ if } t_{k+1} < t'_k + s_k + \delta_k.$$

So far, the coupling appears to be only a minor nuisance to the computations, which involve only trivial additions and comparisons anyway. We shall see later, however, that it is this coupling which causes most of the mathematical problems in the stochastic analysis of queues in which the τ_j and s_j are treated as random variables.

1.3 Graphical representation

A graph of a function always seems to convey information more rapidly than a formula, and graphical constructions of solutions to mathematical problems, when they exist, frequently can replace very complicated analytical procedures. This is certainly the case in the analysis of queues. We shall see, though, that some things which are analytically simple are not graphically simple, and vice versa. The quickest way to find solutions to mathematical problems is to exploit each technique in its proper place. There are several methods of graphical representation of the evolution of a queueing system, a few of which will be described here.

The most obvious way of identifying graphically a set of arrival times t_j is to mark points on a straight line at distances t_j from some origin. It turns out, however, that it is more convenient here to draw a graph of a function $A(t)$ which, for each t, represents the cumulative number of arrivals to time t

$$A(t) = \text{number of } t_j \text{ with } t_j \leq t. \tag{3.1}$$

This is a step function which increases by one at each time t_j as shown in Figure 1.1.

One immediate advantage of this representation is that we will also have occasion to analyse inflow and outflow of quantities other than *numbers* of customers, for example value of products, amounts of work to be done, etc. If the quantity in question is something that is conserved in the sense that what comes in must go out, then it is useful to generalize (3.1) to

$$A(t) = \text{cumulative quantity (or number) to arrive by time } t. \tag{3.1a}$$

This is also a monotone non-decreasing function of t, but is not necessarily integer valued. It may or may not be a step function depending upon whether the arrivals are discrete or not.

On the same graph as $A(t)$, one can draw another curve for

$$D(t) = \text{number of } t_j' \text{ with } t_j' \leq t \tag{3.2}$$

or

$$D(t) = \begin{array}{l}\text{cumulative quantity (or number) to enter}\\ \text{service by time } t,\end{array} \tag{3.2a}$$

and

$$D^*(t) = \text{number of } t_j'' \text{ with } t_j'' < t \tag{3.3}$$

or

$$D^*(t) = \begin{array}{l}\text{cumulative quantity (or number) to have left}\\ \text{the service by time } t.\end{array} \tag{3.3a}$$

On this graph one can easily identify most of the quantities of interest in queueing theory (particularly for the queue with first in first out discipline and one channel). At any time t, $A(t) - D(t)$ represents the number of customers (or quantity) which has arrived since time 0 but has not yet left the queue. Thus

$$\text{quantity in queue or queue length} = Q(t) = A(t) - D(t) \tag{3.4a}$$

is the vertical distance at time t between the two curves. Similarly,

$$\text{quantity or number of customers in system} \\ = Q^*(t) = A(t) - D^*(t) \tag{3.4b}$$

is the vertical distance between $A(t)$ and $D^*(t)$ at time t, and

$$\text{quantity or number of customers in service} \\ = D(t) - D^*(t). \tag{3.4c}$$

These three quantities are all required to be non-negative, thus

$$A(t) \geq D(t) \geq D^*(t) \text{ for all } t.$$

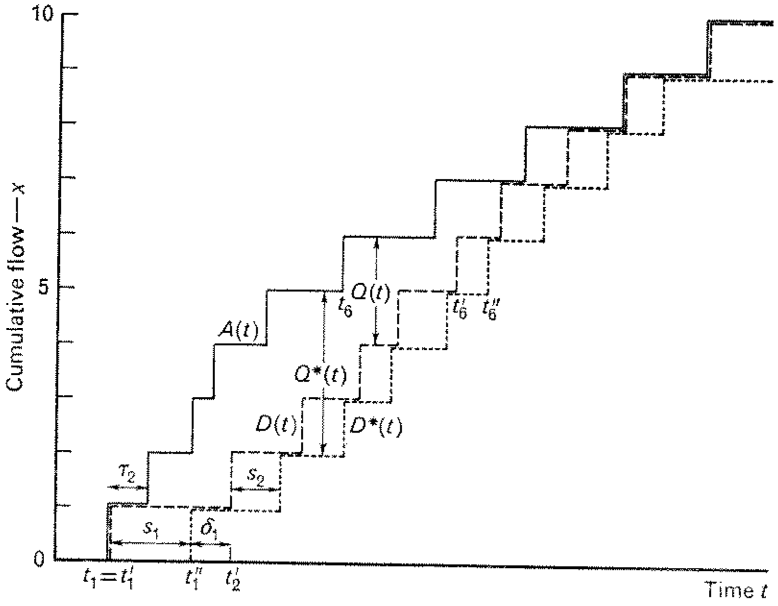

Figure 1.1

In the case of FIFO discipline, the arrival time of the j^{th} customer is the time when $A(t)$ jumps from $j-1$ to j and his departure time from the queue is the time when $D(t)$ jumps from $j-1$ to j. The time spent in queue is the difference between these times and is identified geometrically as the horizontal distance between $A(t)$ and $D(t)$ at the height between $j-1$ and j. This time in queue is also the area of a horizontal strip between $A(t)$ and $D(t)$ between heights $j-1$ and j.

For a single channel server (or actually any server for which the departures are in the same order as the entrances) the horizontal distance between $D(t)$ and $D^*(t)$ represents a service time and the horizontal distance from $A(t)$ to $D^*(t)$ is (for FIFO) the total time in the system, the time in queue plus time in service.

If $A(t)$ represents a non-integer quantity, but service is FIFO in the sense that any indivisible units of that quantity leave in the same order as they arrive, then times spent in the queue and/or service can still be identified geometrically as horizontal distance, in the same sense.

Geometrically, a horizontal distance is no more complicated to visualize than a vertical distance, but in the customary mathematical notation, the former is perhaps more awkward. If we had chosen to draw a graph of arrival times *v.* cumulative number instead of the reverse, the horizontal axis would be the vertical axis and vice versa; in fact, the new graph would be just a reflection of the old graph across a 45° line through the origin. The customary notation for this is that if x and t represent the co-ordinates of the graph, and we draw the graph $x = A(t)$ as in Figure 1.1, the reflected graph is the 'inverse', $t = A^{-1}(x)$. Similarly for $D(t)$. The horizontal distance between $A(t)$ and $D(t)$ at height x would be identified as $D^{-1}(x) - A^{-1}(x)$. The horizontal distance associated with a customer who arrives at time t would be $D^{-1}(A(t)) - t$, a rather awkward symbol for something that is geometrically so simple.

Although Figure 1.1 furnishes a convenient graphical representation of the count or quantity of arrivals, departures, and queue lengths (regardless of the queue discipline or multiplicity of servers), it does not always furnish a convenient way of identifying queue disciplines or delays if the order of service is not FIFO. If the service is not FIFO, the delay to a customer who arrives at time t is identified as a horizontal distance between time t and *some* point at which $D(t)$ increases, but it is no longer the length of a horizontal line between $A(t)$ and $D(t)$. To identify delays to individual customers, one must superimpose upon this graph some scheme of identifying which departure is associated with which arrival. There is no convenient way of doing this (except perhaps for a few other special types of discipline).

If the queue discipline is not FIFO and/or there is a multiple-channel server, there is an alternative way of representing the evolution of the system. We draw $A(t)$ as before except that we interpret it as a graph of the inverse, i.e., for each cumulative number of arrivals x, $A^{-1}(x)$ is the arrival time of the x^{th} arrival. This x is considered as a continuous variable, any fractional part of the customer being considered to arrive at the time of the whole. For $j-1<x<j$, $A^{-1}(x)$ is the arrival time of the j^{th} customer, identified in section 1.2 by t_j, i.e., the function $A^{-1}(j)$ is essentially equivalent to the function t_j of j.

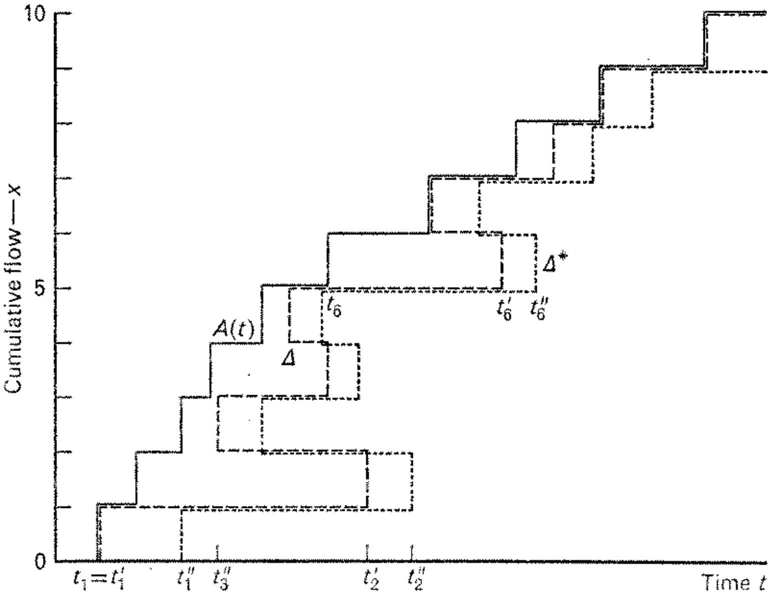

Figure 1.2

The function $D^{-1}(x)$ represents the time of the x^{th} cumulative departure from the queue. Instead of drawing the curve $D^{-1}(x)$, however, we now draw a curve for the time of departure from the queue of the x^{th} cumulative *arrival*, i.e., of the customer originally labeled with x. For $j-1<x<j$, this is the quantity t_j' of Section 1.2 indexed with the arrival numbers. This graph is labeled in Figure 1.2 as Δ and is drawn with the 'independent variable' x along the vertical axis.

If service were FIFO, the curve Δ and $D(t)$ would be the same. Otherwise the curve Δ represents a single-valued function of x, but it is not monotone in x nor does it have a unique inverse (it does not define a single-valued function of t). Since $t'_j \geq t_j$, the curve Δ must never lie to the left of $A(t)$.

The advantage of this diagram as compared with Figure 1.1 is that it contains all relevant information about the system including the discipline (one can see from the diagram the order in which various customers entered service). Also, the delay in queue to any j^{th} customer is still the horizontal distance $t'_j - t_j$ between $A(t)$ and Δ at height j. The disadvantage of the figure is that the geometric interpretation of quantity or number in the queue is less convenient. Since Δ and $A(t)$ never cross, they will enclose an area in the $x-t$ plane, the locus of all horizontal lines from $A(t)$ to Δ. If one draws a vertical line at time t, it will slice this area in such a way that any vertical segment will be identified with a set of heights x for customers who have arrived but have not left the queue. The total length of all segments will be the quantity or number in queue. In contrast with Figure 1.1, this is not generally just the length of a single segment.

For a multiple-channel server, one can also draw a curve Δ^* which describes the times at which customers leave the service, in the same way as Δ describes the times at which they leave the queue.

1.4 Averages

The above representations of queues involve only elementary mathematics. It is, in a sense, the approximate theory of queues which is complicated. In analyzing the behavior of queues, one does not want to be forced to observe the arrival and service times of every customer. This, on the one hand, involves too much data and secondly it is not very interesting data because it is not the same in two different experiments. One would prefer to specify only a few characteristics, some average arrival rates, service rates, etc. Furthermore, even if one could describe in detail exactly how large the queue would be at each instant of time, this is not what one wishes to determine. One would prefer some more or less qualitative description, some 'figures of merit' which measure the typical behavior. This is essentially the reason why one treats the queueing phenomena as a stochastic process. One only wishes to consider the average behavior of the system over a wide range of typical patterns.

Whether one treats the problem stochastically or deterministically, however, there are certain gross properties one may wish to calculate; for example, the average wait in queue for a collection of n customers or the time average queue length over some period of time.

For FIFO queue discipline, the total queueing time to customers $j+1$ to $j+n$ inclusive is equal to the area of the region in Figure 1.1 bounded by $A(t)$, $D(t)$ and two horizontal lines at height j and $j+n$. The average queueing time to these customers is

$$\text{average queueing time} \equiv \frac{1}{n}\sum_{j+1}^{j+n}(t'_k - t_k) \tag{4.1}$$

$$= \frac{1}{n} \times \text{total queueing time,}$$

equivalently the average horizontal distance between $A(t)$ and $D(t)$ over the vertical range j to $j+n$. Similar interpretations of the total time in the system involve the areas between A and D^*.

The total queueing time spent by customers during a time interval t to $t+dt$ (during which $Q(t)$ is constant) is $Q(t)dt$, the number of customers in queue during $(t, t+dt)$ multiplied by the time dt each customer is delayed. The total queueing time during a time interval (a, b) must be

$$\int_a^b Q(t)dt = \int_a^b [A(t)-D(t)]dt \tag{4.2}$$

which is also interpreted geometrically as an area, the area bounded by $A(t)$ and $D(t)$ and two vertical lines at times a and b. The average queueing time per unit time is the average queue length, the average of the vertical distance between $A(t)$ and $D(t)$.

$$\text{average queue length} \equiv \frac{1}{(b-a)}\int_a^b Q(t)dt \tag{4.3}$$

$$= \frac{1}{(b-a)} \times \text{total queueing time.}$$

If in (4.1) and (4.3) the times a and b are chosen as any times for which $Q(a) = Q(b) = 0$, then the curves $A(t)$ and $D(t)$ themselves enclose an area. If $A(a) = j$ and $A(b) = j+n$, then the total queueing time during the time (a, b) in (4.3) is the same as the total queueing time to customers $j+1$ to $j+n$ in (4.1), i.e., the areas in question can be computed either by addition of horizontal strips or integration of vertical strips. Also by definition of the averages as above

$$\begin{aligned}(\text{time interval } b-a) &\times \text{average queue length}\\ &= (\text{number of customers } n)\\ &\times (\text{average queue time per customer})\\ &= \text{total queueing time.}\end{aligned} \tag{4.4}$$

If the queue behavior is such that the queue vanishes repeatedly, every day at midnight or at other perhaps irregular (maybe stochastic) times with a finite spacing, the above relation holds for any or all choices of times a and b when the queue vanishes (not necessarily consecutive times). If $(b-a)$ is sufficiently large and the queue vanishes many times between a and b, more or less evenly distributed over (a, b), then it should not be important whether the queue vanishes at a and b or not. Any contribution to the total delay coming from the end conditions, between a and b and neighboring times when the queue does vanish, will be relatively unimportant anyway in computing averages.

Over a long time one might also expect that the number of arrivals n during (a, b) divided by $b-a$, the average queue time per customer, and the average queue length would all have limiting values (be essentially independent of $b-a$). If so, then these averages must be related by

$$\begin{aligned}\text{average queue length} &= (\text{average number of arrivals}\\ &\qquad \text{per unit time})\\ &\times (\text{average queue time per customer}).\end{aligned} \tag{4.5}$$

The 'average number of arrivals per unit time' is by definition the long time average. If the quantities in (4.5) are meaningful, (4.5) is true by virtue of the definition of the averages.

Much of the literature on queueing theory deals with what is known as 'stationary arrivals' and this relation (4.5) is one of the basic equations, also one of the few relations that is valid for a wide class of stochastic systems. Theorems relating to (4.5) can be quite involved, but the difficulties, in effect, center around the verification of mathematical conditions under which these averages are well-defined, conditions which would be more difficult to verify experimentally than (4.5) itself. For proof of this, see J. D. C. Little, 'A proof for the queueing formula $L = \lambda W$,' *Operations Research* **9** (1961), 383-387, and W. S. Jewell, 'A simple proof of: $L = \lambda W$,' *Operations Research* **15** (1967), 1109-1116.

There are also companion relations to (4.1) to (4.5) involving arrivals and departures from the system (queue plus service). If a and b are chosen at times when $Q^*(t) = 0$, the system is empty, then for a single server queue the counterpart of (4.4) is

$$\begin{aligned}\text{(time interval } b-a) &\times \text{(average number in system)}\\ &= \text{(number of customers } n)\\ &\times \text{(average time in system per customer)}\\ &= \text{total time in system,}\end{aligned} \tag{4.6}$$

or if long time averages are well-defined

$$\begin{aligned}\text{average number in system} &= \text{(arrivals per unit time)}\\ &\times \text{(average time in system}\\ &\ \text{per customer)}\end{aligned} \tag{4.7}$$

By the same type argument, one can obtain corresponding relations for the service system alone. Although the curves $D(t)$ and $D^*(t)$ may be strongly dependent upon each other, there is nothing in the derivation of (4.4) or (4.6) that specifies how these curves are obtained. Thus if averages are well-defined, it must also be true that

$$\begin{aligned}\text{(average number in service)} &= \text{(arrivals per unit time)}\\ &\times \text{(average time of a customer}\\ &\ \text{in service)}\end{aligned} \tag{4.8}$$

Note that if long time averages are well-defined, the arrivals per unit time must be equal to the number entering service per unit time or the number leaving service per unit time.

Although the above relations (4.4) to (4.8) were derived under the assumption that the queue discipline was FIFO and the queue was a single channel server (customers entered and left the service in the same order in which they arrived), these relations are also true for *any* queue discipline or any number of channels (the ordering of customers is irrelevant). Also generalizations of these equations can be obtained for any conserved quantity.

In a figure similar to Figure 1.2, suppose times a and b are two times when the queue vanishes. Consider the region R enclosed by $A(t)$ and Δ between the times a and b, or between heights $A(a)$ and $A(b)$. The area of this region can be obtained by integration of vertical slices through R of width dt or by integration (or summation) of horizontal slices. If we now use $Q(t)$ to denote the quantity in queue at time t, the length of the intersection of a vertical line at t with R, then

$$\text{area of } R = \int_a^b Q(t)dt. \tag{4.9}$$

Also if we had drawn the curves $A(t)$ and $D(t)$ for cumulative arrivals and departures, this $Q(t)$ would be $A(t)-D(t)$ as in (4.2). The analogue of (4.3) is

$$\text{average quantity in queue} = \frac{1}{b-a}\int_a^b Q(t)dt = \frac{\text{area of } R}{b-a}. \tag{4.10}$$

If $A(t)$ is composed of steps of height a_k, i.e., the k^{th} arrival brings to the queue a quantity a_k, and the heights $A(a)$ to $A(b)$ enclose customers $j+1$ to $j+n$ inclusive, then

$$\text{area of } R = \sum_{k=j+1}^{j+n} a_k(t'_k - t_k). \tag{4.11}$$

We can interpret this in (at least) two ways. If we define

$$\frac{1}{n}\sum_{k=k+1}^{j+n} a_k(t'_k - t_k) \equiv \text{average area per customer} \tag{4.12}$$

then

$$\text{area of } R = n \times \text{average area per customer.}$$

This interpretation is of interest particularly if the quantity in question, the a_k for customer k, happens to be the cost per unit of delay for customer k (the value of his time), or if a_k is the value of the k^{th} arrival (objects in a production line, for example) times the interest rate of money so that a_k is the cost per unit time of storage in the queue. The average area per customer is then interpreted as the average cost per customer.

A second interpretation is to let

$$\frac{1}{\sum_{j+1}^{j+n} a_k} \sum_{j+1}^{j+n} a_k(t'_k - t_k) \equiv \text{average delay per unit quantity,} \qquad (4.13)$$

so that

$$\text{area of } R = [\text{quantity to arrive during } (a, b)] \times [\text{average delay per unit quantity}].$$

If now, we equate the expressions for area of R, we have

$$\text{average quantity in queue} = \left[\frac{\text{number of arrivals in } (a, b)}{(b-a)}\right] \times (\text{average area per customer}) \qquad (4.14a)$$

$$= \left[\frac{\text{quantity to arrive in } (a, b)}{(b-a)}\right] \times (\text{average delay per quantity}) \qquad (4.14b)$$

If these expressions have limits for large $(b-a)$, the quantities in square brackets are interpreted as the long time arrival rates of customers or quantity in (4.14a) and (4.14b) respectively.

In the special case in which the quantity is the number of customers, (4.14a) and (4.14b), both reduce to (4.5) except that here the queue discipline was unspecified. Analogously (4.7) and (4.8) do not depend upon the queue discipline or the number of channels in the server.

One should not necessarily infer from the above argument, however, that the average delay per customer for FIFO is the same as for any other queue discipline. We have only shown that (4.5) to (4.8) are true for any queue discipline or equivalently that the average delay per customer for any queue discipline is the same as for a (perhaps hypothetical) system with FIFO queue discipline, and the *same* departure curve $D(t)$. It may be that this hypothetical

FIFO queue discipline cannot, in fact, be realized or if it could, it would yield a *different* departure curve than the given one. There are, of course, situations in which the departure times do not depend upon the identity or order of the customers (all customers are equivalent), in which case it is true that the average delay per customer is independent of the queue discipline.

In this latter case, the advantage of FIFO over other types of queue discipline is not related to the *average* delay but with the variations in delay about the average. Obviously last come first served discipline gives a high proportion of very short delays, also some very long ones, but the same average as FIFO.

1.5 Applications of elementary relationships

In most typical queueing problems, one specifies the arrival rate of customers and the service times and one wishes to determine (among other things) the average queue length and/or the average delay per customer. Equation (4.5) relates these two unknowns in a simple way, so it suffices to evaluate either one or the other.

Equation (4.8), however, has some more direct applications. Suppose one has a service facility of so many channels that a queue never forms, each customer enters service immediately. The arrival rate of customers is specified (this is the long time average arrival rate T^{-1} × number of arrivals during time T, for large T) and so is the average service time per customer (the usual arithmetic average over many customers each with equal weight). The question is: what is the time average of the number of servers that are busy? This would give at least a preliminary (low) estimate of how many servers one needs. Equation (4.8) gives this directly because the average time of a customer in service is the average service time and the number of busy servers is equal to the number of customers in service.

This is a rather special but important type of 'queueing' situation. A telephone company, for example, knows the frequency of calls (arrival rate) between two cities and the average duration of a call (service time) and wishes to know how many telephone lines it should provide, or at least how many would be used on the average (number in service). An airport manager knows how many aircraft movements there will be each day, or each week (arrival rate) and the average 'turn-around time,' or the average time spent at a gate (service time). He wishes to know how many gates will be occupied, on the average (average number in service). In both cases, one would actually like to know also something about the peaks in demand, but these simple formulas at least give some useful information with a minimum of effort.

For most multiple-channel queueing systems in which each channel serves just one customer at a time and all channels are similar, the service time of a customer does not depend upon the *number* of channels. For example, for the telephone trunk line, the length of a call does not depend upon the number of trunk lines, and for the airport, the turn-around time does not depend upon the number of gates. In this case, the right hand side of (4.8) does not depend upon the number of servers, provided it is sufficient to guarantee that queues do not grow indefinitely, i.e., the long time average arrivals per unit time into the service is the same as the arrivals to the system. Thus the average number of busy servers is independent of the number of channels.

The difference between a service with many channels and one with only a few channels is that the former can serve customers with less delay in queue. During temporary surges in the arrivals, many channels are in use; but during lulls, very few are in use. In the latter case, a queue forms during the surges and it is served during the lull; the servers are kept busy most of the time. The time average number of busy servers is the same in both cases, but the former has larger fluctuations.

Equation (4.8) also gives some interesting information for a single channel server. Again one typically knows the arrival rate and average service time. Equation (4.8) determines the time average of customers in service. In the case of a single channel server, the number in service at any time is either 0 or 1, and the average number in service is the same as the fraction of time one server is busy. The left hand side of (4.8) is, in this case, called the 'traffic intensity' denoted usually by ϱ. The quantity $1-\varrho$ is therefore the fraction of time the server is idle. It is also interpreted as the probability that a customer arriving 'at a random time' will enter service without delay.

PROBLEMS

1.1 From a table of random numbers, choose a set of 50 consecutive digits (random numbers from 0, 1, . . . , 9), and designate these as times τ_j, $j = 2, \dots, 51$. From a different set of numbers, choose the first 50 digits other than 9's (random numbers from 0, 1, . . . , 8) and designate these as times s_j, $j = 1, \dots, 50$. Suppose these are the interarrival times and service times, respectively, of a single-channel service with the first arrival at time 0. Draw a graph of $A(t)$ and $D(t)$ (scaled so as to fit on a standard size sheet of graph paper). Also draw a graph of $Q(t)$.

1.2. Show how on a curve of $A(t)$ and $D(t)$ one would identify the delay to customer j if the queue discipline were last come first served.

1.3. (a) Let $0 = t_1 < t_2 \dots t_n$ be ordered arrival times and $0 < t_1' < t_2' \dots < t_n' = t_n$ ordered departures from a queue which vanishes at time 0 and t_n. If t_{n_j}' is the departure time of customer j, show that the sum of the squares of the delays

$$\sum_{j=1}^{n} (t_{n_j}' - t_j)^2$$

is least if $n_j = j$, i.e., for FIFO service. (The t_j' are assumed to be independent of the order of service).

(b) As a generalization of (a), suppose the cost of delay to the j^{th} customer is a function $C(w_j)$ of the delay $w_j = t_{nj}' - t_j$ with the function $C(x)$ the same for all customers. The total cost of delay to all customers is

$$\sum_{j=1}^{n} C(w_j).$$

If the marginal cost per unit delay $c(x)$,

$$c(x) = dC(x)/dx,\ C(x) = \int_0^x c(x')dx',$$

is a monotone increasing function of x, show that the total cost of delay is least for FIFO.