ARAB REPUBLIC OF EGYPT

THE INSTITUTE OF NATIONAL PLANNING

Memo No (390)

Lectures On Queueing Theory

By

T.L Saaty

Jan 12 1964

جمهورية مصر العربية – طريق صلاح سالم – مدينة نصر – القاهرة – مكتبب ريد رقم ١١٧٦٥

A.R.E Salah Salem St. Nasr City , Cairo P.O.Box : 11765

Lectures On Queueing Theory

By

T.L. Saaty.

With the ever increasing world population is associated a host of problems, a pressing one among which is the correspondingly increasing demand on existing service facilities.

Queuing theory provides stochastic models for the analysis of congestion problems in order to predict demands and sizes of facilities to cope with these demands etc. The operation of a modern airport with large numbers of arriving and departing aircraft requires substantial order to maintain schedules and avoid disasters. A telephone system is another example of a queue. A.K. Erlang began the first formal considerations of congestion problems at the telephone company in Copenhagen early this century.

Today throughout the world there are many individuals both in Universities and outside who are interested in the stochastic properties of queues. We shall again give a condensed series of lectures on the subject stopping to point out possible applications of the subject for planning purposes.

T. L. Saaty

T.L. Saaty

Cairo

Jan 12, 1964

CHAPTER 1

A DESCRIPTION OF QUEUES

1-1. Introduction

In this book we present the existing theory of queues (waiting lines) within a structural franmework and emphasize different mathematical models and useful methods of solution with applications of various ideas to several activities such as telephone traffic, inventory, machine repairs, dam operation, aircraft operation, and others.

In this chapter, in a painless descriptive manner, a structure for queues is presented. Chapter 2 is an introductory, technical chapter, with a variety of ideas for the interested reader. The following three chapters give ideas of the theory (mostly probability) and the mechanies (mostly mathematical analysis) involved in solving queueing problems. In fact, Chaps. 4 and 5 specialize in Poisson queues. Other parts of the structure are then studied in Chaps. 6 to 13, followed by Chap. 14 on applications. Chapter 15 is concerned with renewal theory. A comprehensive bibliography is then given.

Because of the limitation of available techniques, it is natural that the material in Chap. 1 should be more extensive in ideas than the representation provided in the rest of the book. However, progress in queueing research is rapidly providing many of the missing links toward a coherent representation of the structure outlined and even beyond.

A queue, or a waiting line, involves arriving items that wait to be served at the facility which provides the service they seek. Suppose that the facility is a check-out counter at a grocery store. If the line is long, customers may become impatient and leave, thus causing a loss in profit. The owner of the facility may decide that investment in another checkout counter is worthwhile because its cost is offset by the profits taken from the impatient customers, more of whom now remain to be served. In this manner, one introduces optimization into queueing theory. In general, unlike optimization theory in which the main concern is to maximize or minimize an objective function subject to constraints, queueing theory is mostly a mathematically descriptive theory. It attempts to formulate, interpret, and predict for the purpose of better understanding of queues and for the sake of introducing remedies.

The reader might reflect on how often and how long he is made to wait or is delayed in queues in his day-to-day activities. He waits for transportation, for food, for the doctor, in an airplane waiting to land, at the theater, etc. Then, from a practical point of view, he might roughly compute how long, on an average, he waits per week in all the congestions he meets. Extended to a total for a lifetime, the amount of time wasted in queues is appalling. If one is not properly occupied, many queueing activites amount to an appreciable depreciation of the enjoyment of life. But by anticipation and good planning the agony and waste of queues could be minimized.

That queues are inevitable and are here to stay is a corollary of the increase in population. The more individuals that there are demanding a certain type of service (whether in the theater or in city traffic), the longer the waiting line. This demand is rising with increasing population; therefore the total amount spent in queues in a lifetime is an increasing function of time. Of course, one might proportionately increase service, but the costs are frequently prohibitive, e.g., widening city streets to allow for increased traffic. Projected planning could anticipate some of these difficulties by allowing for an increase in population.

Most queueing systems involve human beings. Studying causes and remedies of queueing problems cannot be completely divorced from consideration of human factors and their influence on the problem. Nor can remedies be applied without regard to the fact that it is the people using the facilities who matter in the final analysis.

While built-in queues, or planned queues, are a well-known feature of city life, they are also indispensable in intorducing order in the various phases of mass production. The presence of a queue can be a good or a bad recommendation for a public facility. For example, at times it is commonly understood that regular queues before a theater or a cinema are one sign that the show is worth seeing and that a feature which is easily accessible is not. Thus, to find a good show, look for the larger queues are bus queues. They are sometimes but a little removed from the chaos which they are set up to avoid. Many queueing places are served by more than a single bus; consequently, when two or more buses arrive simultaneously

there is a scramble and a general disintegration of the queue. If the buses become full, the rejected rider has great difficulty finding his rightful place in the queue and frequently must queue up all over again.

Waiting for the doctor on the basis of a schedule designed to accommodate both doctor and patient would be very desirable. Frequently, the patient's loss of time comprises a greater social and economic loss than that of the doctor. If it is observed that the schedule of appointments is carefully developed after some study, it could add to the prestige of the doctor. It would, at least, point to the high regard that he has for his patients' time.

There is need for both the "queuee" joining the waiting line and the "queuer" who organizes it to understand and anticipate queues and their subtleties. Often the queuee would benefit by knowing that the line in which he waits has been optimally designed to minimize his wait individually or as a member of a group, and the queuer would know better how to plan for the queue associated with his activity. The most significant queue of all is the life queue, i.e., birth-life-death cycle with life as the waiting aspect. In a sense, then, queue consciousness becomes a consciousness for planning toward better living.

Some general types of queueing phenomena to which the ideas of this book have applicability are in the fields of communications traffic (telephone, telegraph, post office), transportation traffic (air, land, sea), queueing for service (theaters, restaurants, buses, hospitals, and clinics), inventories and industrial processes (maintenance, assembly lines, machine interference), physical processes (operations of a set of dams, particles moving through a hole), epidemic processes in biology, population growth, even refereeing papers for publication, psychological flow of nervous impulses, etc.

1-2 An Illustrative Example and a Numerical Illustration

Many of the ideas arising in queueing theory can be illustrated in one important example: take-off and landing of aircraft at a metropolitan airport-an operation of interest to a large number of people who use the

facilities. For emphasis, the queueing terms illustrated by this example are italicized in the next paragraphs.

The airport is assumed to have several runways (parallel channels) used for take-off and for landing. These runways lead to a smaller or larger number of paths ending at the terminals (channels in series or queues in tandem). After an aircraft, which arrives according to a certain arrival distribution, lands, it joins the queue of aircraft awaiting service (movement) on a path to the debrkation point. Thus, the output of one queue becomes the input to another. The waiting line itself is both on the ground (take-off of aircraft) and stacked in the air (landing aircraft). Both these queues have input distributions. Landing aircraft may arrive in batches where the members of each batch must be spaced for circling over the airport and landing in order. (In case a runway is very wide it is not difficult to conceive of aircraft landing in batches.) The duration of the service operation (landing or take-off time on a runway) is about a minute. In any case, there is a service distribution, and if different types of aircraft are allowed to use different runways, which may be larger for jet aircraft, for example the service distribution may vary from one runway to another.

It is essential that an appropriate measure of effectiveness be chosen for the selection for landing. For example, if it is desired to minimize the total waiting time of individuals, it may be more desirable first to land those aircraft with the greater number of passengers.

An informal type of priority system is often used whereby a circling aircraft is allowed to land prior to take-off of waiting aircraft. This priority system is further extended to emergency cases where a late-arriving aircraft is allowed to land first for urgent reasons. Frequently jet aircraft, because of limited fuel capacity, are given landing priority. Sometimes, by the nature of the holding pattern, an arriving aircraft, having joined the queue of stacked aircraft waiting to land, is chosen at random for landing (a form of service priority). Thus the aircraft which is nearest to a point where it

can leave the pattern will be given instructions to land. Between being given the priority for landing and a "clear to land" instruction, an aircraft moves from the stack to the "landing aid." The time used is known as the "approach time." The "landing time" is spent in the landing operation until the aircraft turns off the runway.

A circling aircraft may be in a quasi emergency with others that are in actual emergency and decide to join a shorter queue at a nearby airport and land there. An arriving aircraft may decide not to stack and goes (balks) to another airport. This aircraft is then said to be "lost" to airport-as distinguished from being delayed; or, after joining the queue and waiting longer than desirable, it leaves (reneges) for the neighboring airport. A landing aircraft may be considered to cycle when it joins the queue of aircraftwaiting to take off, again becoming an input item into the system. If a landing aircraft has information on the size of the stacked queue in a neighboring airport, it might join that queue, if it has information on yet another airport, it might go there (a rare situation). This moving back and forth when there are several lines is called jockeying (queue-selection rule).

An airport may be temporarily shut down and arriving aircraft diverted to another airport if the number of stacked aircraft reaches a prescribed size. The service operation may be speeded up by building "turn offs" which make it possible for an aircraft to turn off the main runway at high speed.

A fundamental problem in airport operations is that of communications. When the input both on land and in the air is large, the airport control must communicate rapidly with the aircraft and obtain a response. How many operators and communication channels there must be to handle various congestion situations which might arise is an important communications problem. Here one must decide on an optimum number of channels to service items arriving by a given distribution. One may even compare the cost of an additional channels with the cost of increased sercice at the existing channels.

The problem of having adequate waiting space for the queue is important. For example, an essential aspect in airport design must be adequate ground taxi-way for the aircraft ready for take-off.

For many queueing problems it is enough to know the input distribution, the queue discipline (e.g., random, ordered, or priority selection for service), and the service-time distribution to determine the desired measures of effectiveness. In other queueing problems one must have additional information. For example, in the case of balking (reneging) one must determine the probability that an arriving unit would balk on (after) arrival and hence abandon the queue before (after) joining it.

From a theoretical standpoint, queues may be regarded in terms of flow through a network connecting service points in series and in parallel, as the situation may be. The flow is influenced by various phenomena which can delay it, cause it to overflow, etc.

Consider the following simplified situation of a single queue as shown in Table 1-1. For customer A the first row gives his arrival time and the second row the elapsed time between the arrival of the preceding customer and customer A. The third row gives his service time, and the fourth row is the sum of the service time and waiting time of the previous customer, minus the interarrival time of the customer whose waiting time is being computed. Thus customer A would experience a wait equal to the waiting time plus service time of his predecessor minus the arrivaltime interval of A. If the result is zero or negative the waiting time is zero.

TABLE 1-1. SAMPLE QUEUEING DATA

Chronological time	0	2	6	11	12	19	22	26	36	38	45	47	49	52	61
Between-arrival times	0	2	4	5	1	7	3	4	10	2	7	2	2	3	9
Service time	5	7	1	9	2	4	4	3	1	2	5	4	1	2	1
Waiting time.....	0	3	6	2	10	5	6	6	0	0	0	3	5	3	0

A considerably larger sample would be required for a statistically valid study of an actual operation, but some important queue data can now be computed. In the above example, of the total number of customers, 10 have waited. The average waiting time of those who waited is 49/10, whereas the average waiting time for everyone is 49/15.

The total idle time of the channel may also be computed. The channel is idle and waiting for customer A if the interval between the arrival of A and his predecessor exceeds the total wait in queue and in service of the predecessor. Thus the total idle time is equal to the sum of the differences between the arrival interval of the present unit and the waiting time plus service time of the previous unit whenever the difference is positive. The fraction of time during which the channel is idle is the ratio of the previous quantity and the total time of operation.

If the sample were large enough, the frequency of a single unit waiting, that of two, etc., could be computed by counting the frequency of occurrence of single-unit waits, groups of two, etc. This gives the probability of occurrence of these groups. A question which we shall ask the reader to answer is: What use can one make of this probability?

Another useful quantity is the probability of a given number waiting at any time. Note, for example, that the fourth arrival waited with the third arrival for one time unit and was then left to wait with the fifth arrival for one time unit. This gives two groups of two units waiting in line.

The reader is urged to construct a diagram of horizontal parallel lines, each corresponding to a customer and with lengths corresponding to the duration of wait of each customer, over a base line which is divided according to chronological time. Each line must begin with the arrival time of the customer and terminate with the time in which he enters service. In this manner the number waiting and

the duration of wait of this number can be measured. For example, the line corresponding to the second customer extends from the second to the fifth time unit, and that of the third customer extends from the sixth to the twelfth time unit. There is no overlap between the two the two lines. However, the line corresponding to the fourth customer extends from the eleventh to the thirteenth time unit and overlaps by one time unit the line corresponding to the previous customer. It overlaps the line corresponding to the subsequent customer by an equal amount, etc. Thus, to obtain the frequency of waiting of a group of two at any time, one takes the ratio of the total number of time units in which a group of two waited and divides by the total time.

Repetition of the above experiment with new data and table gives rise to new situations. With sufficient repetitions for a practical case in which measurements are taken, one can compute the probability of a given number waiting at a given time; this is different from the previous probability, which is given at any time. It is obtained by counting for the many runs of the experiment the frequency of occurrence of a single wait, groups of two, etc., at the prescribed time.

Thus we have three types of quantities to compute for the number waiting. They are (1) the frequency of occurrence of a given group of items waiting together (i.e., how often they occur), (2) the frequency of occurrence of a group at any time, and (3) the frequency of occurrence of a group at a prescribed time.

In passing, passing, we note that a queueing situation must be studied over the period in which meaningful action is required with regard to congestion. For example, at a restaurant congestions usually occur at noon and in the evening. Sometimes it is of little use to study the two together because of the different intensities in traffic and because of the difference in its fluctuation in the two periods. If the congestion were independent of the time of day (i.e., homogeneous in time), matters would be

simpler. However, caution is required in properly examining and sorting the periods in which congestion occurs in a given operation.

As a final remark, note that one can also obtain the arrival and service rates from the above data. If the data were more extensive, the distribution of arrival times and service times could be found.

1-3. Varieties of Queues

We divide a queueing operation into four parts: the input; the waiting line, the service facility, and the output. With each of these is associated a set of alternative assumptions concerning the queueing process, some of which have been the object of research in the field, as indicated in the historical background. Other assumptions lead to as yet unsolved queueing problems which require investigation. We give a general description of various queueing possibilites with some repetion of ideas given in the previous section.

1. Types of Arrival and Service-time Distributions

Arrivals into a queue (which may have an initial number waiting before the operation starts) occur by assumption according to a certain frequency distribution, and so do the intervals between arrivals. These intervals may be independently distributed for many application purposes or may be dependent, as, for example, in the case of flow leaving a traffic light. The same remarks apply to times of entry to service and to service times.

2. Initial-input Variations

The initial number in a system when an operation begins may be given by a distribution because it is different for each complete run of the operation (e.g., from day to day). The input to a queue may be from a limited or an unlimited population which may also

consist of several categories (populations) of customers, each of which may arrive by a different distribution, singly or in batches, and may queue in a prescribed order. The input distribution may depend on the output distribution, as in a hospital where patients are admitted if there are vacant beds.

3. Customer Behavior

a. Balking. Customer behavior can vary. Arriving customers may balk (i.e., not join the queue) because of the length of the existing queue, or simply because they have to wait at all, and are consequently lost. Sometimes they are lost because they have no opportunity to wait, as is the case with a busy telephone signal (lost calls) although they can reinitiate a call. It is also possible to hold such a call, delaying it until a trunk becomes free. There may be a single line in which to wait before going into service, or there may be several lines, as in banks or in supermarkets. An item may join the nearest line independently of its length. If arrivals are designed to occur at constant intervals, they can, for example, still occur sooner or later by a distribution about the arrival point as a mean.

b. Influence of Incomplete Information. For many problems a decision may be required as to which line of a multiple-queue operation to join when information about only a few is immediately available a case of incomplete information. In congested traffic the lack of knowledge as to which is the best route to follow without trying them all is also "incomplete information".

c. Customer Adapation to Queue Conditions. On the basis of experience, passengers may learn to travel earlier or later to avoid intolerable queueing, and such measures, when adequately studied, may even relieve congestion. For example, ships approaching the Suez Canal

can be notified to slow down until the queue at Port Said is reduced to an acceptable size. An item may join a large waiting line at closing hours for fear that a short line which it encounters may be closed suddenly - a familiar experience. But there are situations in which an item which arrives before another must go into service before the following one. There are cases in which each service facility has its own specialty and consequently its own queue, as a stamp-sale counter and money-order or special registry counter at the Post Office.

d. Collusion, Jockeying, and Reneging. Several customers may be in collusion whereby only one person waits in line while the rest are then free to attend to other things. Some may even arrange to take turns waiting. Units may jockey from one line to another, as in a bank. A customer may lose patience and leave the line, i.e., renege.

4. Queue and Channel Variations

a. Full or Limited Availability. Service channels may be available to any unit waiting in a system (full availability) or may be available only to some waiting units. Other units are blocked and must wait until a channel that can provide the required service becomes available. In telephone link systems, whether one obtains a free connection depends on whether a free inlet to the next waiting line can be combined with a free outlet. The idea itself shows the need for economy in setting up possible combinations. This is particularly true for some long-distance calls which may have to pass through more than one switching center.

b. Service Procedures or Discipline. While in line customers may be chosen for service by allocation to the channels in an ordered first-come-first-served manner or at random, they may be assigned priorities with errors committed when initially it is not clear which priority to