Chapter 1

Introduction

1.1 Basic System Elements

Queues (or waiting lines) help facilities or businesses provide service in an orderly fashion. Forming a queue being a social phenomenon, it is beneficial to the society if it can be managed so that both the unit that waits and the one that serves get the most benefit. For instance, there was a time when in airline terminals passengers formed separate queues in front of check-in counters. But now we see invariably only one line feeding into several counters. This is the result of the realization that a single line policy serves better for the passengers as well as the airline management. Such a conclusion has come from analyzing the mode by which a queue is formed and the service is provided. The analysis is based on building a mathematical model representing the process of arrival of passengers who join the queue, the rules by which they are allowed into service, and the time it takes to serve the passengers. Queueing theory embodies the full gamut of such models covering all perceivable systems which incorporate characteristics of a queue.

We identify the unit demanding service, whether it is human or otherwise, as *customer*. The unit providing service is known as the *server*. This terminology of customers and servers is used in a generic sense regardless of the nature of the physical context. Some examples are given below:

(a) In communication systems, voice or data traffic queue up for lines for transmission. A simple example is the telephone exchange.

(b) In a manufacturing system with several work stations, units completing work in one station wait for access to the next.

(c) Vehicles requiring service wait for their turn in a garage.

(d) Patients arrive at a doctor's clinic for treatment.

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1_1

Numerous examples of this type are of everyday occurrence. While analyzing them we can identify some basic elements of the systems.

Input Process If the occurrence of arrivals and the offer of service are strictly according to schedule, a queue can be avoided. But in practice this does not happen. In most cases, the arrivals are the product of external factors. Therefore, the best one can do is to describe the input process in terms of random variables that represent either the number arriving during a time interval or the time interval between successive arrivals. If customers arrive in groups, their size can be a random variable as well.

Service Mechanism The uncertainties involved in the service mechanism are the number of servers, the number of customers getting served at any time, and the duration and mode of service. Networks of queues consist of more than one server arranged in series and/or parallel. Random variables are used to represent service times, and the number of servers, when appropriate. If service is provided for customers in groups, their size can also be a random variable.

System Capacity The number of customers that can wait at a time in a queueing system is a significant factor for consideration. If the waiting room is large, one can assume that for all practical purposes, it is infinite. But our everyday experience with the telephone systems tells us that the size of the buffer that accommodates our call while waiting to get a free line is important as well.

Queue Discipline All other factors regarding the rules of conduct of the queue can be pooled under this heading. One of these is the rule followed by the server in accepting customers for service. In this context, the rules such as "first-come, first-served" (FCFS), "last-come, first-served" (LCFS), and "random selection for service" (RS) are self-explanatory. Others such as "round robin" and "shortest processing time" may need some elaboration, which is provided in later chapters. In many situations, customers in some classes get priority for service over others. There are many other queue disciplines which have been introduced for the efficient operation of computers and communication systems. Also, there are other factors of customer behavior such as balking, reneging, and jockeying, that require consideration as well.

The identification of these elements provides a taxonomy for symbolically representing queueing systems with a variety of system elements. The basic representation widely used in queueing theory is due to D. G. Kendall (1953) and made up of symbols representing three elements: input, service, and number of servers. For instance, using M for Poisson or exponential, D for deterministic (constant), E_k for the Erlang distribution with scale parameter k, and G for general (also GI, for general independent) we write:

$M/G/1$: Poisson arrivals, general service, single server
$E_k/M/1$: Erlangian arrival, exponential service, single server
$M/D/s$: Poisson arrival, constant service, s servers.

These symbolic representations are modified when other factors are involved.

1.2 Problems in a Queueing System

The ultimate objective of the analysis of queueing systems is to understand the behavior of their underlying processes so that informed and intelligent decisions can be made in their management. Three types of problems can be identified in this process.

Behavioral Problems The study of behavioral problems of queueing systems is intended to understand how they behave under various conditions. The bulk of the results in queueing theory is based on research on behavioral problems. Mathematical models for the probability relationships among the various elements of the underlying process are used in the analysis. To make the ideas concrete let us define a few terms that are defined formally later. A collection or a sequence of random variables that are indexed by a parameter such as time is known as a *stochastic process*; e.g., an hourly record of the number of accidents occurring in a city. In the context of a queueing system, the number of customers with time as the parameter is a stochastic process. Let $Q(t)$ be the number of customers in the system at time t. This number is the difference between the number of arrivals and departures during $(0, t)$. Let $A(t)$ and $D(t)$, respectively, be these numbers. A simple relationship would then be $Q(t) = A(t) - D(t)$. In order to manage the system efficiently, one has to understand how the process $Q(t)$ behaves over time. Since the process $Q(t)$ is dependent on $A(t)$ and $D(t)$, both of which are also stochastic processes, their properties and dependence characteristics between the two should also be understood. All these are idealized models to varied degrees of realism. As done in many other branches of science, they are studied analytically with the hope that the information obtained from such study will be useful in the decision-making process.

In addition to the number of customers in the system, which we call the *queue length*, the time a new arrival has to wait till its service begins (*waiting time*) and the length of time the server is continuously busy (*busy period*) or continuously idle (*idle period*) are major characteristics of interest. It should be noted that the queue length and the waiting time are stochastic processes and the busy period is a random variable. Distribution characteristics of the stochastic processes and random variables are needed to understand their behavior. Since time is a factor, the analysis has to make a distinction between the *time-dependent*, also known as *transient*, and the *limiting*, also known as the *long-term*, behavior. Under certain conditions a stochastic process may settle down to what is commonly called a *steady state* or a state of *equilibrium*, in which its distribution properties are independent of time.

Statistical Problems Under statistical problems we include the analysis of empirical data in order to identify the correct mathematical model, and validation methods to determine whether the proposed model is appropriate. Chronologically, the statistical study precedes the behavioral study as could be seen from the early papers by A. K. Erlang (as reported in Brockmeyer et al. (1960)) and others. For an insight into the selection of the correct mathematical model, which could be used to derive its properties, a statistical study is fundamental.

In the course of modeling we make several assumptions regarding the basic elements of the model. Naturally, there should be a mechanism by which these assumptions could be verified. Starting with testing the goodness of fit for the arrival and service distributions, one would need to estimate the parameters of the model and/or test hypotheses concerning the parameters or behavior of the system. Other important questions where statistical procedures play a part are in the determination of the inherent dependencies among elements, and dependence of the system on time.

Decision Problems Under this heading we include all problems that are inherent in the operation of queueing systems. Some such problems are statistical in nature. Others are related to the design, control, and the measurement of effectiveness of the systems.

1.3 A Historical Perspective

The history of queueing theory goes back more than 100 years. Johannsen's "Waiting Times and Number of Calls" (an article published in 1907 and reprinted in *Post Office Electrical Engineers Journal*, London, October, 1910) seems to be the first paper on the subject. But the method used in this paper was not mathematically exact and therefore, from the point of view of exact treatment, the paper that has historic importance is A. K. Erlang's, "The Theory of Probabilities and Telephone Conversations" (*Nyt tidsskrift for Matematik, B, 20* (1909), p. 33). In this paper he lays the foundation for the place of Poisson (and hence, exponential) distribution in queueing theory. His papers written in the next 20 years contain some of the most important concepts and techniques; the notion of statistical equilibrium and the method of writing down state balance equations are two such examples. Special mention should be made of his paper "On the Rational Determination of the Number of Circuits" (see Brockmeyer et al. (1960)), in which an optimization problem in queueing theory was tackled for the first time.

It should be noted that in Erlang's work, as well as the work done by others in the twenties and thirties, the motivation has been the practical problem of congestion. See for instance, Molina (1927) and Fry (1928). During the next two decades, several theoreticians became interested in these problems and developed general models which could be used in more complex situations. Some of the authors with important contributions are Crommelin, Feller, Jensen, Khintchine, Kolmogorov, Palm, and Pollaczek. A detailed account of the investigations made by these authors may be found in books by Syski (1960) and Saaty (1961). Kolmogorov's and Feller's study of purely discontinuous processes laid the foundation for the theory of Markov processes as it developed in later years.

Noting the inadequacy of the equilibrium theory in many queue situations, Pollaczek (1934) began investigations of the behavior of the system during a finite time interval. Since then and throughout his career, he did considerable work in the analytical behavioral study of queueing systems; see Pollaczek (1965). The trend toward the analytical study of the basic stochastic processes of the system continued, and queueing theory proved to be a fertile field for researchers who wanted to do fundamental research on stochastic processes involving mathematical models.

A concept that plays a significant role in the analysis of stochastic systems is *statistical equilibrium*. This is a state of the stochastic process which signifies that its behavior is independent of time and the initial state. Suppose we define

$$P_{ij}(s,t) = P[Q(t) = j|Q(s) = i] \qquad s < t$$

as the *transition probability* of the process $\{Q(t),\ t \geq 0\}$, which is a statement of the probability distribution of the state of the process at time t, conditional on its state at time s, $s < t$. The statement that the process attains statistical equilibrium implies that

$$\lim_{t\to\infty} P_{ij}(s,t) = p_j$$

which does not depend on time t and the initial state i.

Even though Erlang did not explicity state his results in these terms, he used this basic concept in his results. To this day a large majority of queueing theory results used in practice are those derived under the assumption of statistical equilibrium. Nevertheless, to understand the underlying processes fully, a time-dependent analysis is essential. But the processes involved are not simple and for such an analysis sophisticated mathematical procedures become necessary. Thus, the growth of queueing theory can be traced on two parallel tracks:

(i) Using existing mathematical techniques or developing new ones for the analysis of the underlying processes

(ii) Incorporating various system characteristics to make the model closely represent the real-world phenomenon

Queueing theory as an identifiable body of literature was essentially defined by the foundational research of the 1950s and 1960s. For a complete bibliography of research in this period, see Syski (1960), Saaty (1961, 1966), and Bhat (1969). Here we mention only a few papers and books that, in the opinion of this author, have made a profound impact in the direction of research in queueing theory.

The queue $M/M/1$ (Poisson arrival, exponential service, single server) is one of the earliest systems to be analyzed. Under statistical equilibrium, the state balance equations are simple and the limiting distribution of the queue size is obtained by recursive arguments. But for a time-dependent solution, more advanced mathematical techniques become necessary. The first such solution was given by Bailey (1954) using generating functions for the differential equations governing the underlying process, while Lederman and Reuter (1956) used spectral theory in their solution. Laplace transforms were used later for the same problem, and their use together with generating functions has been one of the standard and popular procedures in the anlaysis of queueing systems ever since.

A probabilistic approach to the analysis was initiated by Kendall (1951, 1953) when he demonstrated that imbedded Markov chains can be identified in the queue length process in systems $M/G/1$ and $GI/M/s$. Lindley (1952) derived integral equations for waiting time distributions defined at imbedded Markov points in the general queue $GI/G/1$. These investigations led to the use of renewal theory in queueing systems analysis in the 1960s. Identification of the imbedded Markov chains also facilitated the use of combinatorial methods by considering the queue length at Markov points as a random walk. See Prabhu and Bhat (1963) and Takàcs (1967).

Mathematical modeling of a random phenomenon is a process of approximation. A probabilistic model brings it a little bit closer to reality; nevertheless it cannot completely represent the real-world phenomenon because of involved uncertainties. Therefore, it is a matter of convenience where one can draw the line between the simplicity of the model and the closeness of the representation. In the 1960s several authors initiated studies on the role of approximations in the analysis of queueing systems. Because of the need for useable results in applications, various types of approximations have appeared in the literature. For an extensive bibliography, see Bhat et al. (1979). To mention a few, one approach to approximation is the analysis under heavy traffic (when the traffic intensity, the ratio of the rates of input to output, approaches 1) and investigations under this topic were initiated by Kingman (for an extensive bibliography, see Kingman (1965)) with the objective of deriving a simpler expression for the final result. The heavy traffic assumption also led to diffusion approximation as well as weak convergence results by researchers such as Iglehart (see Iglehart and Whitt (1970a, b)). Also see Whitt (2000) with an extensive bibliography. Gaver's analysis (1968) of the virtual waiting time of an $M/G/1$ queue is one of the initial efforts using diffusion approximation for a queueing system. Fluid approximation, as suggested by Newell (1968, 1971) considers the arrival and departure processes in the system as a fluid flowing in and out of a reservoir, and their properties are derived using applied mathematical techniques. For a recent survey of some fluid models see Kulkarni (1997).

By the end of 1960s most of the basic queueing systems that could be considered as reasonable models of real-world phenomena had been analyzed and the papers coming out dealt with only minor variations of the systems without contributing much to methodology. There were even statements made to the effect that queueing theory was at the last stages of its life. But such predictions were made without knowing what advances in computer technology would mean to queueing theory. Advances inspired or assisted by computer technology have come in two dimensions: methodology and applications. Given below are some of the prominent topics explored in such advances. Since in applied probability, methodology, and applications contribute to the growth of the subject in a symbiotic manner they are listed below without being categorized.

(i) The Matrix-Analytic Method

Starting with the introduction of phase type probability distributions, Marcel Neuts (1975) has developed an analysis technique that extends and modifies the earlier transform method to multivariables and makes it amenable for an algorithmic solution. See Neuts (1978, 1981), Sengupta (1989), and Ramaswami (1990, 2001). The use of phase type distributions in the representation of system elements and the matrix-analytic method in their analysis has significantly expanded the scope of queueing systems for which useable results can be derived. See, Chapter 8 for details.

(ii) Transform Inversion

The traditional method of analysis of queueing systems depends on inverting generating functions and/or Laplace transforms to derive useable results. The complexities of transform inversion has spurred more research on it and beginning with Abate and Dubner (1968), Dubner and Abate (1968), and Abate et al. (1968) many papers have been published on the subject. For a comprehensive survey of the state of the art of the Fourier series method of inversion see Abate and Whitt (1992).

In the inversion of Laplace transforms and probability generating functions, finding roots of characteristic equations is a key step. The celebrated Rouché's theorem only establishes the existence of the roots, not their magnitude. Pioneering and painstaking work in adapting various root finding algorithms for use in inverting transforms and generating functions is due to Professor M. L. Chaudhry (1992). Starting from the 1970s, along with his associates, he has put together a significant amount of research on various queueing systems of interest (see, Chaudhry and Templeton (1983)). For instance Chaudhry et al. (1992) provides a good illustration.

(iii) Queueing Networks

The first article on queueing networks is by J. Jackson (1957). Mathematical foundations for the analysis of queueing networks are due to Whittle (1967, 1968) and Kingman (1969), who treated them in the terminology of population processes. Complex queueing network problems have been investigated extensively since the beginning of the 1970s.

Two key concepts that advanced investigations into the properties of queueing networks are: the Poisson nature of the departure process from an M/M/s type queue (Burke 1956) and the local balance in state transitions (Whittle 1967, 1968). The $M \to M$ property, as the Poisson property has been called in computer network lierature, is a necessary condition for the limiting distribution to be in the product form. Going beyond the simple Jackson network, Baskett et al. (1975) show that the product from solutions are valid for networks more general than those with simple M/M/s type nodes, such as, with state-dependent service; heterogeneous service times; Coxian service time distributions; processor sharing discipline; and last-come, first-served discipline.

Since the publication of Baskett et al., a large body of literature has grown in the performance modeling of queueing networks. Courtois (1977), Kelley (1979), Sauer and Chandy (1981), Lavenberg (1983), Disney and Kiessler (1987), Malloy (1989), Perros (1994), Gelenbe and Pujolle (1999) and Giambene (2005) are some of the significant books that have come out on this subject.

(iv) Computer and Communication Systems

The need to analyze traffic processes in the rapidly growing computer and communication industry is the primary reason for the resurgence of queueing theory after the 1960s. Research on queueing networks (see references cited earlier) and books such as Coffman and Denning (1973) and Kleinrock (1975, 1976) laid the foundation for a vigorous growth in the application of queueing theory in computer and communication system operation.

In tracking this growth, we may cite the following survey type articles from the journal *Queueing Systems*: Denning and Buzen (1978) on the operational analysis of queueing network models; Coffman and Hoffri (1986), describing important computer devices and the queueing models used in analyzing their performance; Yashkov (1987) on analytical time-sharing models, complementary to McKinney (1969) on the same topic; three special issues of the journal edited by Mitra and Mitrani (1991), Doshi and Yao (1995), and Konstantopolous (1998); and a paper by Mitra et al. (1991) on communication systems. Research on queueing applications can also be found in various computer journals. Several books have appeared and continue to appear on the subject as well. Some of the more recent developments are discussed in Chapter 13.

(v) Manufactruring Systems

The machine interference problem analyzed by Palm (1947) and Benson and Cox (1951, 1952) was the first problem in manufacturing systems in which queueing theory methodology was used. The classical Jackson network (1957) originated out of the manufacturing setting since a job-shop is a network of machines. (Also, see Jackson (1963)). Simulation studies reported in Conway et al. (1967) provide excellent examples of the incorporation of queueing models with job-shop scheduling. Since the 1970s, with the advent of new processes in manufacturing incorporating computers at various stages, the application of queueing theory results as well as the development of new techniques have occurred at a phenomenal rate. Three articles in Buzacott and Shanthikumar (1992) and the book Buzacott and Shanthikumar (1993) bring together most of the important developments in the application of queueing theory in manufacturing systems up to that time.

As described by Buzacott and Shanthikumar (1993) the "product-to-order" and "product-to-stock" models make direct use of queueing theory results. With demand as a customer and the manufacturing process as a server, the first model is a direct application of queueing models, while the second incorporates production–inventory system concepts, with the production system substituting for multiple or infinite number of servers. Other applications include job flow lines as tandem queues, and job-shops and flexible manufacturing systems as queueing networks. Some of the more recent applications are discussed in Chapter 12. For recent articles on the applications of queueing theory in manufacturing system modeling readers may also refer to various journals such as *Management Science, European Journal of Operational Research, IIE Transactions, Computers and Industrial Engineering*, and journals on production and manufacturing research.

(vi) Specialized Models

Specialized queueing models of the 1950s and 1960s have found broader applicability in the context of computer and communication systems. We mention below three such models that have attracted considerable attention.

Polling Models These models represent systems in which one or more servers provide service to several queues in a cyclical manner (Koenigsberg (1958)). Based on variations on the system structure and queue discipline a large number of models emerge. For research on polling models see a special issue of *Queueing Systems* edited by Boxma and Takagi (1992), as well as Takagi (1997) and Hirayama et al. (2004), all of which provide excellent bibliography on the subject.

Vacation Models Queueing systems with service breaks are not uncommon. Machine breakdowns, service disruption due to maintenance operations, cyclic server queues, and scheduled job streams are some of the examples. A key feature of the results is the ability to decompose them into results corresponding to systems without vacations and results depending on the distributions related to the vacation sequence. For bibliographies on this topic, see Doshi (1986) and Alfa (2003).

Retrial Queues In finite capacity systems, customers, denied entry to the system, trying to enter again, is quite common. Since they have already tried to get service once, they belong to a different population of customers than the original one. Problems related to this phenomenon have been extensively explored in the literature. The following papers and more recent ones appearing in journals provide bibliographies for further study: Yang and Templeton (1987), Falin (1990), and Kulkarni and Liang (1997).

(vii) *Statistical Inference*

In any theory of stochastic modeling statistical problems naturally arise in the applications of the models. Identification of the appropraite model, estimation of parameters from empirical data, and drawing inferences regarding future operations involve statistical procedures. These were recognized even in earlier investigations in the studies by Erlang; see Brockmeyer et al. (1960), Molina (1927), and Fry (1928).

Since elements contributing to the underlying processes in queueing systems can be modeled as random variables and their distributions, it is reasonable to assume that inference problems in queueing are not any different from such problems in statistics in general. However, often in real-world systems, sampling plans appropriate for data collection to estimate parameters of the constituent elements, may not be possible to implement. Consequently, modifications of the standard statistical procedures become necessary.

The first theoretical treatment of the estimation problem was given by Clarke (1957) who derived maximum likelihood estimates of arrival and service rates in an $M/M/1$ queueing system. Billingsley's (1961) treatment of inference in Markov processes in general and Wolff's (1965) derivation of likelihood ratio tests and maximum likelihood estimates for queues that can be modeled as birth and death processes are other significant advances that have occurred in this area. Also see Cox (1965) for a comprehensive survey of statistical problems as related to queues. Cox also provides a broad guideline for inference investigations in non-Markovian queues.

The first paper on estimating parameters in a non-Markovian system is by Goyal and Harris (1972), who used the transition probabilities of the imbedded Markov chain to set up the likelihood function. Since then, significant progress has occurred in adapting statistical procedures to various systems. Some of the examples are: Basawa and Prabhu (1981, 1988) and Acharya (1999) considered the problem of estimation of parameters in the queue $GI/G/1$; Rao et al. (1984) used a sequential probability ratio technique for the control of parameters in $M/E_k/1$ and $E_k/M/1$; Armero (1994) and Armero and Conesa (2000) used Bayesian techniques for inference in Markovian queues; Thiruvaiyaru et al. (1991) and Thiruvaiyaru and Basawa (1994) extended the maximum likelihood estimation to include Jackson networks; Pitts (1994) considered the queue as a functional that maps the service and inter-arrival time distribution functions on to the stationary waiting time distribution function to determine its confidence bound. For a comprehensive survey of inference problems in queues see Bhat et al. (1997). More recent investigations are by Bhat and Basawa (2002) who use queue length as well as waiting time data in estimating parameters in queueing systems. A recent paper (Basawa et al. 2008) uses waiting time or system sojourn time, adjusted for idle times when necessary, to estimate parameters of inter-arrival and service times in *GI/G/*1 queues.

(viii) Design and Control

The study of real systems is motivated by the objectives of improving their design, control and effectiveness. Until the 1960s when operations researchers trained in mathematical optimization techniques got interested in queueing problems, operational problems were being handled using primarily behavioral results. It should be noted that Erlang's interest in the subject was for building better telephone systems for the company for which he was working. His paper "On the rational determination of the number of circuits" (Brockmeyer et al. (1960)) deals with the determination of the optimum number of channels so as to reduce the probability of loss in the system.

Until computers made them obsolete, graphs and tables, prepared using analytical results of measures of effectiveness, assisted the designers of communication systems such as telephones. Other examples are the papers by Bailey (1952) which looked into the appointment system in hospitals, and Edie (1956) that analyzed the traffic delays at tollbooths. From the perspective of applications of queueing results to realistic problems Morse's (1958) book has been held in high regard. This is because he presented the theoretical results available at that time in a manner appealing to the applied researchers and gave procedures for improving system design.

Hillier's (1963) paper on economic models for industrial waiting line problems is, perhaps, the first paper to introduce standard optimization techniques to queueing problems. While Hillier considered an *M/M/*1 queue, Heyman (1968) derived an optimal policy for turning the server on and off in an *M/G/*1 queue, depending on the state of the system.

Since then, operations researchers trained in mathematical optimization techniques have explored their use in much greater complexity to a large number of queueing systems. For an excellent overview, a valuable reference is a special issue of the journal *Queueing Systems* edited by Stidham (1995), which includes several review-type articles on special topics. Also see Bäuerle (2002) who considers an optimal control problem in a queueing network.

(ix) Other Topics

Even though there were a few papers on discrete time queues before the 1970s, since then, these systems have taken a larger significance because of the discreteness of time, however short the interval maybe, in computer and communication systems. It is not hard to imagine that a large portion of the results for discrete time queues are in fact derived in the same way as for continuous time queues with obvious modifications in methodology.

There have also been theoretical advances in stochastic processes with the introduction of modified processes such as Markov modulated processes, marked point processes and batch Markovian processes. These processes are used to represent various patterns such as burstiness and heterogeneity in traffic.

In the preceding paragraphs, we have outlined the growth of queueing theory identifying major developments and directions. For details of any of the facets, readers are referred to the articles and books cited above. Also see Prabhu (1987) who gives a bibliography of books and survey papers in various categories and subtopics, Adan et al. (2001) who give a broad treatment of queues with multiple waiting lines, and Dshalalow (1997) who considers systems with state-dependent parameters. The last two articles also provide extensive bibliographies. It is hoped that with the help of these references and modern Internet tools, applied researchers will be able to build on the systems covered in this text so as to establish an appropriate model to represent the system of their interest.

1.4 Modeling Exercises

These exercises are given as an introduction to modeling a random phenomenon as a queueing system. In addition to answering the questions posed in the exercises, the reader is required only to identify (i) model elements, (ii) system structure, and (iii) the assumptions one has to make in setting up the model.

1. A city bus company wants to establish a schedule for its bus fleet. In order to do this in a scientific manner, the company entrusts this job to an operations research specialist with sufficient data processing support. Describe the queueing systems involved in this process and the types of data that need to be collected in order to come up with the schedule. Identify the measures of performance for the bus system and the factors that affect these measures when the system is in operation.

2. A newly established business would like to decide on the number of telephone lines it has to install in a cost-effective manner. Identify the elements of the underlying process of the telephone answering system and indicate the specific data that need to be collected to establish the parameters of the system. Also identify the performance measures of interest.

3. In a manufacturing system, a product undergoes several stages (e.g., an automobile assembly line) and within each stage there may be several substages, including testing of components. How can such a system be modeled as a queueing system (including queueing systems for stages and substages) in order to improve the performance of the manufacturing process?

4. An airline offers three types of check-in service for the passengers: (1) First class and business class check-in, (2) regular check-in, and (3) self check-in. Describe the structure of the queueing system that can represent the check-in system and identify the data elements that need to be known to measure its performance. Also indicate the complexities that may result in improving the system by incorporating flexibilities in the system operation.

5. Several terminals used for data entry to a computer share a communication line. Terminals use the line on a first-come, first-served basis and wait in a queue when the line is busy.

   Describe the elements of this queueing system and identify the assumptions that need to be made to analyze system characteristics. (Allen (1990)).

6. In store-and-forward communication networks messages for transmission are stored in buffers of fixed size. Each message may use one or more buffers. The message is transmitted through several identical channels. Knowing the characteristics of the arrival process, transmission rate, and the message length, we are interested in the storage requirements of a network node.

   Describe the general characteristics of the approach in order to estimate the long run storage requirements for this type of a system.

7. In a warehouse, items are stacked in such a way that the most recently stacked item gets removed first. In order to use a queueing model to determine the amount of time the item is stored in the warehouse, describe the elements of such a system and say how we may characterize the time interval of interest.

8. In order to reduce the waiting time of short jobs, a round-robin (RR) service discipline is used. Under an RR queue discipline, each job gets a fixed amount of service, known as a quantum, when it is admitted to the central processing unit (CPU). If the service requirement of the job is more than the quantum, it is sent back to the end of the queue of waiting jobs. This process continues until the CPU can provide the required number of quanta of service to the job.

   Describe how the total service time of the job can be characterized in order to determine the mean amount of time the job spends in the system. (This is known as the *mean response time.*) (See Coffman and Kleinrock (1968) and Coffman and Denning (1973)).

9. A uniprogramming computer system consists of a CPU and a disk drive. After one pass at the CPU a job may need the services of the disk I/O with a certain probability, say p, and the job is complete with the probability $1-p$. There are three independent phases to disk service time: (1) seek time; (2) latency time; and (3) transfer time, each with a specified distribution. After disk service the job goes back to CPU for completing the execution. (Note that a uniprogramming system cannot start another job until the service on the one in the system is complete.)

   We are interested in determining the average response time (waiting time + service time). What type of a model is appropriate for this problem? If a queueing model is appropriate, describe the elements of the system (Trivedi (2002)).

10. In a drum storage unit a shortest-latency-time-first (SLTF) file drum is used to read or write records on files while the drum is rotating. Once a decision is made to process a particular record, the time spent waiting for the record to come under the read/write heads which are fixed is called the latency. The records are not constrained to be of any particular strength. Also, no restrictions are placed on the starting position of the records. Assume that the circumference of the drum is the unit of length and the drum rotates at a constant angular velocity, with period τ (Fuller (1980)).

    Suppose a queueing model is to be used to analyze the performance of the drum-storage unit described above. Describe the elements of such a system and the characteristics to be considered for its performance evaluation.