Science Bulletin of NCTU, Vol. VII, No. 1

SOME FUNDAMENTAL PROBLEMS IN THE THEORY OF QUEUES

WAH-CHUN CHAN

College of Engineering, National Chiao Tung University And The University of Calgary, Calgary, Alberta, Canada

(Received 18 December 1973)

Abstract—*The purpose of this article is an attempt to introduce some fundamental ideas and problems to the young engineers and scientists and to draw their attention to the queuing problems which they may encounter in every day life. It is with the hope that the references may provide a minimum guide for one to get into the field of queuing theory. When the population of a city is getting larger, it is natural to expect that the congestion phenomena in many aspects would become more severe. For the study of various types of queuing problems it is the theory of queues to serve as the most valuable means for such investigations.*

1. INTRODUCTION

Queuing theory is concerned with the study of congestion phenomena in queuing (stochastic service) systems. The first systematic development of the queuing theory was given by the famous Danish scientist A. K. Erlang (1878-1939) of the Copenhagen Telephone Company. Erlang's basic research in queuing theory dates from the years 1908-1932. From then on, interest in the problems formulated hy Erlang increased rapidly. More and more mathematicians, engineers and operations research workers became interested in similar problems and developed new techniques accordingly. It turs out that problems arising in telephone traffic are also relevant for problems in many other fields of research, such as air traffic control, inventory control, machine servicing and maintenance, dam operations, reliability, production line, hospital operations, library operations and postal operations. Moreover, queuing theory has served to present, in clear and challenging form, one of the fundamental problems in the control of operational systems and to determine the indices of performance of queuing systems.

A queuing problem may arise from a conflicting situation in that on the one hand a customer may have to wait too long before obtaining service and on the other hand a facility may remain excessively idle for a long time. Queuing theory provides a means for finding a compromise solution to the conflicting situation.

2. DESCRIPTION OF QUEUING SYSTEM

In order to study the queuing phenomenon in a queuing systen it is necessary to specify the system sufficiently fully. A queuing system is basically characterized

by a flow customers arriving randomly at some service facilities. These customers upon arrival at the service facility may obtain service immediately or may have to wait until the facility is available. The service time of each customer may be regular (fixed) or random depending on the type of service. To specify a queuing system the following characteristics are essential.

(a) Input process. In practice the arrival of customers in a queuing system is controlled by external factors which in general undergo occasional fluctuation. The best that one can do is to represent the input process in terms of randam variables. This means that both the average arrival rate of customers and the probability distribution of the arrival of customers must be known. The simplest input process and the most commonly used one in applications is the Poisson input process.

(b) Service mechanism. This consists of specifying the number of servers (service facilities) in the system, how many customers can be served at a time and how long service takes. Usually the service time of each customer is specified by a probability distribution. Mathematically the simplest service time distribution in the negative exponential distribution which has the property of memorylessness and facilitates the mathematical treatment in many situations.

(c) Queue discipline This is concerned with the method how a customer waiting in the queue is selected for service and how the customer behaves when waiting. The most natural way for selecting a customer for service is "first-come, first-served" or "service in order of arrival". The other possibilities are "last-come, first-served" and "randam selection for service". In many situations it may be necessary to introduce "priority" discipline in order to improve the performance of the system.

The specification must also include the interaction (if any) among various parts of the svstem. When the system has been described sufficiently fully, it is possible to predict the performance of the system by a mathematical model like the one in Fig. 1.

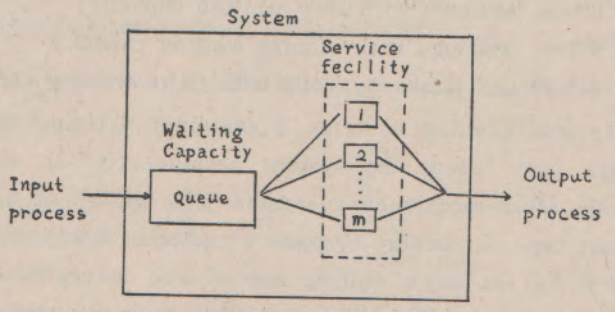

Fig. 1. A schematic representation of a queuing system with m servers

3. SOME PROBLEMS IN THE THEORY OF QUEUES

The problems arising in queuing theory can be classified into three categories.

(a) Behavioral problems. The study of behavioral problems of queuing systems deals with the investigation of a particular system as thoroughly as possible by using mathematical models. These models are studied analytically with the hope that the information obtained from the study will be useful in the design of a queuing system.

Major quantities required to understand the behavior of queuing system are the queue length (number of customers waiting in queue) at a given instant, the waiting time of a customer (the time a new arrival has to wait until his service commences), the length of busy period (the length of time when a server is busv continuously) and the probability of waiting. In this category the study may be further divided into time dependent (transient) and steady state behaviors. In many practical situations, after the system has heen in operation for a sufficiently long time, it may settle down to a behavior which is independent of time. If the major interest of the study of the system is the steady state behavior, then the conditions for the existence of the steady state must be specified. Most of the works in queuing theory was in the study of steady steat behavior of the system. The transient behavior is much more difficult to investigate and has not yet been fully studied.

(b) Statistical problems. The statistical problem of queuing theory is concerned with the collection and analysis of empirical data, estimation of system parameters and test of hypotheses regarding queuing systems. The validity of a mathematical model is justified only when the assumptions made in the model can be verified by empirical data. In order to apply the mathematical model to a practical queuing system the parameters in the model must be determined empirically.

(c) Operational problems. The operational problems of queuing theory consist of all problems concerning the operation of a practical queuing system. Some of these problems are statistical in nature. Others are mainly related to the design, control, improvement and the measurement of effectiveness of queuing systems.

4. THE FUNDAMENTAL QUEUING SYSTEMS

Depending on whether the waiting capacity of a queuing sysfem is zero, finite or infinite, queuing systems are classified into three different types;

(i) systems with losses (systems with zero waiting capacity)

(ii) systems with delays (systems with infinite waiting capacity)

(iii) systems with delays and losses (systems with finite waiting capacity)

In the first type of queuing systems a customer arriving at the system and finding all servers busy leaves the system permanently. In the second type of queuing systems the afore-mentioned customer joins the queue and waits until his service. In the last type of queuing systems a customor arriving at the systems and finding the queue is full (no more waiting spaces) also leaves the system. However, if a waiting space is available, he will wait in the queue for service. Obviously, the last type of queuing systens is a combination of the first and second (delay and loss) types of queuing systems. In fact it includes the first and second systems as particular cases. However, the second type of queuing systems is relatively easier to tackle than the last one.

The first two types of systems were investigated by Erlang extensively[7]. It should be pointed out that the systems just described differ not only in their structure in the waiting capacity but also in the nature of performance.

In Erlang's investigation the input process of the system is a Possion process. Mathematically it is represented by the probability distribution

$$p\{N(t)=k\}=\frac{e^{-\lambda t}(\lambda t)^k}{k!},\ k=0,1,\ldots\ldots \tag{1}$$

where $N(t)$ is a randon variahle representing the number of customers arriving at the system in a time interval of length t, λ is the mean arrival rate and $P\{N(t)=k\}$ is the probability of the event $\{N(t)=k\}$. The input process described by (1) is also known as random arrivals.

The service time of customers is assumed to be exponentially distributed

$$P\{S\leq t\}=1-e^{-\mu t} \tag{2}$$

where S is a randon variable repressenting the length of the service of each customer, μ is the mean service rate and $P\{S\leq t\}$ is the probability of the event $\{S\leq t\}$. Furthermore, the number of servers in the system is m, a positive integer.

The queue discipline of Erlang's system is the first-come, first-served rule or that the customers are served in order of arrivals.

Since the combined delay and loss-queuing system includes the other two as particular cases, only the last type of queuing systems is discussed here.

Under the above specifications the system is described by the set of state equatious[13]

$$\begin{aligned}
&P'_o(t)=-\lambda P_o(t)+\mu P_1(t)\\
&P'_k(t)=\lambda P_{k-1}(t)-(\lambda+k\mu)P_k(t)+(k+1)\mu P_{k+1}(t),\ 1\leq k\leq m-1\\
&P'_k(t)=\lambda P_{k-1}(t)-(\lambda+m\mu)P_k(t)+m\mu P_{k+1}(t),\ m\leq k\leq n-1\\
&P'_k(t)=\lambda P_{n-1}(t)-m\mu P_n(t)
\end{aligned} \tag{3}$$

where m is the number of servers in the system and n is the waiting capacity of the system. The prime denotes derivative with respect to t.

Suppose that $\frac{\lambda}{m\mu}<1$ and that the systen is in the steady state. The system of state equations (3) now reduces to

$$\begin{aligned}
&-\lambda P_o+\mu P_1=0\\
&\lambda P_{k-1}-(\lambda+k\mu)P_1+(k+1)\mu P_{k+1}=0,\ 1\leq k\leq m-1\\
&\lambda P_{k-1}-(\lambda+m\mu)P_k+m\mu P_{k+1}=0,\quad m\leq k\leq n-1\\
&\lambda P_{n-1}-m\mu P_n=0
\end{aligned} \tag{4}$$

The solution of the set of difference equations (4) is given by[13]

$$P_k=\frac{\rho^k}{k!}P_o,\ 0\leq k\leq n,$$

$$P_k=\frac{\rho^k}{k!}\ \frac{P_O}{m^{k-m}},\ m\leq k\leq n, \tag{5}$$

where

$$\frac{1}{P_o}=\sum_{k=0}^{m-1}\frac{\rho^k}{k!}+\frac{\rho^m}{m!}\cdot\frac{1-(\rho/m)^{n-m+1}}{1-(\rho/m)} \quad \text{and} \quad \rho=\frac{\lambda}{\mu}$$

Case 1. If $n=m$, the system becomes Erlang's pure loss system and

$$P_k=\frac{\rho^k}{k!}P_o,\ 0\leq k\leq m \tag{6}$$

where

$$\frac{1}{P_o}=\sum_{k=0}^{m}\frac{\rho^k}{k!}$$

Case 2. If $n=\infty$, the system become Erlang's pure delay system and

$$P_k=\frac{\rho^k}{k!}P_o,\ o\leq k\leq m$$

$$P_k=\frac{\rho^k}{n!}\cdot\frac{R_o}{m^{n-m}},\ k\geq m \tag{7}$$

where

$$\frac{1}{P_o}=\sum_{k=0}^{m-1}\frac{\rho^k}{k!}+\frac{\rho^m}{m!}\cdot\frac{1}{1-(\rho/m)}$$

5. THE COMPUTER-CONTROLLED QUEUING SYSTEM

The advent of electronic digital computers has given rise to many new problems in science and technology. The use of electronic computers has also broaden the the scope of investigation in queuing theory. Computers may be used as a tool for simulation studies or as an information processor for control purposes. When a computer is employed for control and processing information in a queuing system, new techniques for investigating the performance of the system is required. A schematic representation of a computer controlled queuing system is shown in Fig. 2.

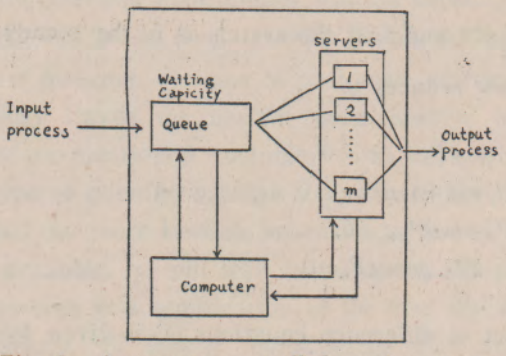

Fig. 2. A computer-controlled queuing system

In order to specify a computer-controlled queuing system in addition to the three characteristics mentioned in Section II it requires the information how the computer affects the operation of the queuing system. In many situations a computer is operated on a time-sharing basis, so that it only processes the customers in the

queue periodically. In this case the customers arriving at the system first join the queue and wait for the computer to process. This means that in a computer-controlled queuing system customers in the queue are served only at certain epochs in a periodic manner. Note that in this situation the tine axis is divided into equal lengths shown in Fig. 3. The period τ is called the access cycle of the queuing system. Since in computer-controlled queuing systems time is no longer a continuous parameter the state equations are difference equations. However, using the method of generating functions the steady state probability distribution for the system can be calculated and is given by the generating function[14~15]

$$\Phi(z, X)=\frac{e^{-\lambda\tau(1-z)}\sum_{k=0}^{m-1}[z^m\{\theta+z(1-\theta)\}^k-z^k\{\theta+z(1-\theta)\}^m]P_k}{z^m-\{\theta+z(1-\theta)\}^m e^{-\lambda\tau(1-z)}}$$

where $\theta=1-e^{-\mu\tau}$ and X is a random variable representing the number of customers (waiting and being served) in the system at the access point. For obtaining the numerical solution of the state probability distribution one may refer to Reference 16.

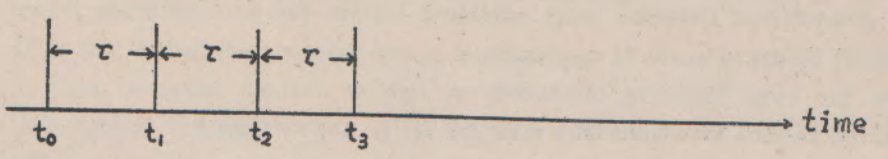

Fig. 3.

6. CONCLUDING REMARKS

After more than half a centry of its development the theory of queues has reached the state of maturity. Since queuing situations exist in every day life, it is often desirable to have information concerning the effectiveness of a queuing situation based on some cost analysis. An increase in the number of servers in the system would eventually decrease the expected waiting time of the customer and and hence improve the quality of service, but it would also increase the cost of service. If the expected waiting time can be controlled, it is possible to select the optimal number of servers in the sense of minimizing the costs of service and and waiting time. In most practical situations, these costs are subtle that a meaningful estimate may be difficult. To overcome this difficulty much further research and analysis are required in this direction.

REFERENCES

1. Kendall, D.G.: Some Problems in the Theory of Queues, J. Roy. Statist. Soc., Ser. B, Vol. 13, No. 2, pp. 151-185, 1951.
2. Kendall, D.G.: Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of the Imbedded Markov Chain, Ann. Math. Statist., Vol. 24, pp. 338-354, 1953.

3. Cox, D.R. and Smith, W.L.: Queues, Methuen & Co., London, 1961.
4. Cox, D.R.: Renewal Theory, Methuen & Co., London, 1962.
5. Morse, P.M.: Queues, Inventories and Maintenance, John Wiley & Son, New York, 1958.
6. Saaty, T.L.: Elements of Queueing Theory, McGraw Hill, New York, 1961.
7. Khintchine, A.Y.: Mathematical Methods in the Theory of Queueing, Hafner, New York, 1960.
8. Riordan, J.: Stochastic Service Systems, John Willey & Son, New York, 1962.
9. Syski, R.: Introduction to Congestion Theory in Telephone Systems, Oliver and Boyd, London, 1960.
10. Gnedenko, B.V. and Kovalenko, I.N.: Introduction to Queueing Theory, Isael Program for Scientific Translations, 1968.
11. Bhat, U.N.: Sixty Years of Queueing Theory, Management Sci., Vol. 15, No. 6, pp. B-280-B-294, 1969.
12. Lee, A.M.: Applied Queueing Theory, St. Martin's Press, 1966.
13. Chan, W.C.: Combined Delay-and Loss-Queuing System, Proc. IEE, Vol. 117, No. 11, pp. 2073-2076, 1970.
14. Chan, W.C. and Chung, W.K.: Computer-Controlled Queuing Systems with Feedback, Proc. IEE, Vol. 118, No. 10, pp. 1373-1377, 1971.
15. Chung, W.K. and Chan, W.C.: Waiting-Time Distribution in Computer-Controlled Queuing System, Proc. IEE, Vol. 118, No. 10, pp. 1378-1382, 1971.
16. Chan, W,C. and Omar, A.: Waiting Time Distribution in a Computer Processing Queueing System, Inform and Control, Vol. 23, pp. 152-164, 1970.