Sindh. Univ. Res. Jour. (Sci. Sr) Vol: 29 (2) 45-49 (1997)

MANAGEMENT OF QUEUEING SYSTEMS

Noor Ahmed Shaikh, Asif Ali G. Kazi, and Sirajuddin Memon
Institute of mathematics and Computer Science, University of Sindh, Jamshoro, 76080, Sindh, Pakistan

Abstract

The purpose of this study is to provide a survey of operation research results of management of queueing systems. The emphasis is on the results themselves and how these results may be useful for problems of practical nature, rather than theoritical derivations, Two main issues on the subject areas namely, static design and dynamic control issues are discussed. Finally a brief look at the management of MRP system is given.

Introduction

The queueing systems are frequently encountered in diverse application areas such as telecommunication, production, computer, transportation, health care, machine interfrence (or machine repair) and other services systems of immense interest. The interest of queueing systems management has grown substantially inthe past three decades.

The object area is divided into two parts: Static optimal design issues versus dynamic, optimal control issues. In the first part various system configurations are described and the analysis of the resulting equilibrium, steady state behaviour allows one to determine the best system for optimizing some long run criteria such as average annual cost or profit. The term "Static" is used because once the system configuration is chosen, then it does not change over time. For example, a system may be designed to have five parallel servers, and then once this system is developed,it continues with five servers indefinitely.

The second part of the subject area, the study of dynamic, optimal control issues, pertains to situations where characteristics of the system are allowed to change over time. The term "dynamics" refers to the fact that the optimal choice of some performance measurers (such as the number of servers) may be allowed to vary according to some state of the system (such as the number of customers) which, in trun, would normally rather than steady-state distributions, tend to be used to determine the optimal decisions for these dynamic situations.

While the majority of relevant references are of a specific nature, a few of general interest should be mentioned; Crabil *el al* (1977) compiled a bibliography of research on the optimal design and control of queues. Two

CamScanner

survey papers on dynamic controll issues were presented at the same conference in 1973; Prabhu and Stidham (1974) surveys the use of the decision theory methodology for queueing control, and Sobel, (1974) focuses on the structure of optimal polices for the service rate. These three source list about 200 different references. More recently, Heyman and Soble, (1982), Shaikh and S.M. Aqil Burney (1974), Shaikh (1996; 1997) and Noor A. Shaikh (1996; 1997) have discussed queueing system management. The balance of this study is devoted to the basic principles of queueing system management and is divided into two sections. First section deals with static design and second one with dynamic control issues.

In static optimal design queueing problems, generally the approach is to formulate a fairaly standard optimization problem, with objective function and constraints being expressions that involve steady-state. equilibrium quantities such as the arrival rate y, the service rate u, and the number of servers s are the decision variables. Since steady-state, equilibrium quantities are involved, this approach is most accurate for situations where the system charactersitic are constant over time. However this approach can also be used for situations where the system's characteristics vary over time: one assumes that during short periods the system's behaviour can be approximated by steady-state expressions.

Fior static, optimal design problem the objective function is often cost or profit, one example is to choose the number of servers s so as to minimize the cost C_w L+C_SS, where C_w is the waiting cost per unit time, C_S is the operating cost per server per unit time, and the steady-sate quantity L is the average number of customers in the system. Note that L is actually the function of s.

A slightly different situations is to choose the service rate u for a single server system so as to minimize cost C_W L+C (u). Where C (u) is the operating cost per unit time as a function of u. Now L depends on u, with problems like this, one might assume that the decision parametter must take one of a finite number specified values, in which case one would simply compare a finite number of objective function value to determine the best design.

For example suppose one can choose any u > y for M/M/1 system with c (u) = c u. Since L = y / (u-y), the objective function is

$$C_W \, y/(u-y) x \, c \, u.$$

Differentiating this with respect to u shows this is minimized by the value

$$u = y + \sqrt{C_w \, y / c.}$$

A rather than different kind of design problem is to organize the customers into priority categories so as to reduce the average number of customers in the system. This is done by giving higher priority to customers who can be reorganized in advance as having shorter processing or service times. Just as with the shortest processing time discipline for job shop scheduling

CamScanner

(Conway *et al*, 1967) by promptly serving those customers who can be quickly dispatched. The average time for all the customers can often be reduced. Of course, this is a deviation from the first-come-first-served discipline, and the average waiting times for the more time consuming customers may be greater than under the conventional FCFS discipline.

It is necessary to estimate a customer's service time in advance to implement shorter processing time idea. In many cases this requirement is oftenly met. The examples of such situations include airline check-in and hospital emergency rooms; when higher priority is given to those who can be treated more quickly.

Perhaps the best references on priority queues is the book of Jaiswal (1968). Unfortunately, formulas for the average waiting time for each priority class of customers are only known for a small number of simple situations. It is therefore often necessary to rely on computer simulation. An application of such computer simulation to queueing model, may be seen in Shaikh (1997). In particular, one would first specify a finite number of priority organization schemes, then estimates the average waiting times for each scheme with computer simulation and finally select the best scheme.

Another kind of design problem concerns with the trade-off for multi-server systems of having a single waiting line versus individual waiting lines for each server. The choice is often available for banks and Government offices. Having individual line often requires less space and is a system that is easier to implement, but the average waiting time may be greater. Further more, with a single line the customers are more satisified knowing, the FCFS discipline is adhered to

Dynamic Optimal Control

It is often appropriate to change a queueing systme's characteristics in a dynamic manner, depending on how various parameters might be varying with time.

It is convenient to divide, optimal control of the arrival process and control of the service process. For arrival process control there are two main possibilities; one is to simply deny admission to entering customers under certain circusmtances. In other words, this is binary situation where sometime you "shut the door". Another possibility is to physically control the arrival rate. Such as rejecting every second arrival when the system becomes crowded.

The other kind of dynamic control problem an values controlling the service process. Here there are also two main possiblities. The first to add servers when the system becomes crowded. The second is to keep the same numbers of servers but vary the service rate for each server Shaikh and Burney, (1994).

There are two main approaches for finding the best control function. The first, called the direct approach, is to specify some control functions and then analyse each with standard queueing theory methods.

CamScanner

The second approach is to assume little about the optimal control function but to use dynamic programming methods to compute the optimal control function. The advantage of this dynamic programming approach is that a larger class of control functions will automatically be considered.

We conclude this study of how to manage queueing system with a brief look at the queueing systems formated as machine repair problem (MRP) which finds its applications in Industries. The term machine repair is generally reserved for the loss of production due to a machine having a wait for repair. Two quantities of fundamental importance in dealing with management are the operative utilization (the proportion of time that the operative is working on repairs, and the machine effeciency (ratio of the total production achieved to that which would have occured had no stoppages taken place). The latter quantity is often expressed as a percentage. Thus suppose the operative looks after 10 machines and that during 8 hours shift he spends 7 hours on repairs and that a total of 60 machine hours of production are obtained. The operative utilization would be

$$\cup = 7/8 \text{ or } 87.5\%$$

The machine efficiency would be

$$E = \frac{60}{8 \times 10} = \frac{3}{4} \text{ or } 75\%$$

An alternative to the efficiency is to consider the average number of machines running, i.e. the number of machines running for the whole time would have achieved the production actually obtained. For the example just cited this would be 7.5 (out of 10).

In general, if the operative looks after N machines and we consider the system over a long time T hours, during which time the operative actually works on repairs for T_O hours, and total of T_r running hours are achieved, then

Operative utilisation = $\cup = T_O/T$,

Efficiency = $E = T_r/NT$

There is a simple relationship between two quantities. If the mean number of stops per unit running time per machine is λ, and the mean repair time is C, then in time T, there will be a total Tr stops (asymptotically) and it will take a time T_r c to rapier these stops.

$$T_O = \lambda T_r = \rho T_r$$

where $\rho = \lambda C$

$\therefore$ operative utilisation = $\cup = N \rho E$.

$$\therefore E = \frac{U}{N\rho}$$

CamScanner

If N=1 the operative looks after just one machine the operative utilisation will be $\rho/(1+\rho)$, since each hour of running will require ρ hours to repairs. The effieciency is just $1/(1+\rho)$.

It is also possible to deal with the case in which $N \to \infty$, i.e. the operative has charge of a very large number of machines. In this situation the operative utilisation U will be 1. He will be working for the whole time. Thus from the last equation cited above.

$$E \to \frac{1}{N\rho}$$

It is perhaps more informative and useful for management to look at the average number of machines running in this carse.

This is given by

$$NE \to \frac{1}{\rho}$$

Thus the management must note carefully at the time of assigning machines to the operative that there is an upper limit of machines running which approaches $1/\rho$, no matter how many machines are assigned to the operative.

For other values of N, apart from the extremes of 1 and $N \to 00$, one has to make a more detailed study of the nature of the breakdowns, and distribution of repair times.

The application of multi-user queueing system formated as M/M/m machine interference (repair) problem which also incorporated profit model can be seen in Burney *et al,* (1994).

References

1. Burney, S.M. Shaikh, Noor Ahmed and Kazi Asif Ali., (1994). Optimal profit Analysis for M/M/m Machine Interference model J.Sc. Res, **Vol. XXIII**, 182, 125, 142.
2. Conway, R.W., Maxwell, W.L and Miller, L.W., (1967). Theory of Scheduling. Addison-Wesley, Reading Massachusters.
3. Crabil, T.B., Gross, D. and Magazine, M.J. (1979). Design and control of Queues. Opns. Res. 25,219-232.
4. Heyman, D.P and Sobel, M.J., (1982). Stochastic Models in Operations Research. **Vol. 1**, and **2**. Mc Graw-Hill, New York.
5. Jaiswal, N.K., {1968}. Priority Queues. Academic press, New York.
6. Shaikh Noor Ahmed, (1997). The maintenance of bidirectronally patrolled machines: A Simulation approach, Mehran University Research Journal of Engineering and Technology, 40-44.
7. Shaikh Noor Ahmed, (1996). Management of machine Interference problem in textile knitting process, Journal of social sciences, Sindh University, Jamshoro, Vol. pp. 61-67.
8. Shaikh, Noor Ahmed and Bureny, S.M., (1994). Empirical queuing model: A study st C.D.F. Hospital Maternity patients, Journal of social sciences, Sindh University, Jamshoro, Vol. 1, 165-74.
9. Sobel, M.J., (1974). Optimal operations of Queues. Lecture Notices in Economics and Mathematicsl Systems. No. 98, springer-verlage Berlin, 23, 251.

CamScanner