IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 12, December 2015

Some Basic Concepts in Queuing Theory

Binay Kumar[1], Ashu Vij[2], Pankaj Kumar[3]

Department of Mathematics, M. L. U. D.A.V. College, Phagwara[1]

Department of Mathematics, D.A.V. College, Amritsar[2]

Department of Physics, M. L. U. D.A.V. College, Phagwara[3]

Abstract: In this paper we study some basic concept of queuing theory and provide brief overview of queuing theory. We analyze the basic component of queuing theory and different type of distribution that are used to analyze a queuing model. The importance and requirement of state dependent queuing model also explained. Finally some basic model of queuing theory, performance measures is discussed and methodologies used to analyze such model are explained.

Keywords: State dependent model, Vacation model, Breakdown model, Maximum entropy principle.

1. INTRODUCTION

A queuing system can be described as arrival of unit for some service at a service channel (or counter), forming or joining the queue, if service is not immediately available, and leaving the system after being served. A queue is generally formed when the available resources are not sufficient to satisfy the demands of units and hence enforce them to wait for service. Formation of queue is common phenomenon whether it is visible (in case of human being) or invisible (in case of inanimate object), but waiting in queue is one of the most unpleasant and undesirable activity. Society as well as service provider can get optimum benefit if the queue is managed in such a way that waiting unit as well as service provider gets the most benefit.

Queuing theory is branch of operational research, the objective of queuing theory is to understand queuing system behavior in order to predict its performance, control and optimize its performance. This analysis of queue system is based upon by the mode by which a unit join a queue, the rule by which they join the service, and the time it take to serve the unit. Queuing theory is a very critical aspect of successful management of queuing system.

Hence, the importance of queuing theory cannot be underestimated. Queuing theory has wide variety of application in business, manufacturing and production, industry, public services and in daily life. Most frequent application can be observed in customer service situation, transport and telecommunication. Queuing theory is directly applicable to automobile traffic system, call centre, network telecommunication, server queuing and mainframe computer queuing of telecommunication.

One of critical aspect of queuing theory is performance evaluation of a system. When the system is in working condition, the studies of performance evaluation can be used in enhancing the performance of the system. For designing and development of a suitable queuing system, the studies of performance of the system are carried out in advance before actual developing it by use of stochastic model. The information obtained from pre analysis of system can be used for developing and increasing the efficiency of the system.

2. BASIC CHARACTERISTICS OF QUEUING PROCESSES

The basic elements that characterize the queuing processes are as follow:

2.1. Arrival pattern

The arrival of unit (or customer) in queuing system are not predetermined, they are random in nature. Thus, the best way to describe the arrival process is random variable, and hence we must have to know the probability distribution that's describing the inter-arrival time or number of units arriving during a time interval. Further a unit may join the queue singly or in batch. If the units arrive in batches, then we need random variable to describe their size.

The one of the factor is customer reaction, when it enters into the system. Customer reaction can be characterize by balking (if customer decide not join the queue), reneging (the customer join the queue, but after some time lose patience and leave the queue) and jockey (customer may switch from one waiting line to another waiting line). The final factor is stationary (if the arrival of units is time independent) or non stationary behavior of the input process.

2.2. Service patterns

As like inter-arrival time of unit, the service time of the units are also uncertain, thus there is need to use appropriate probability distribution to specify the service time. Further the service of the units may also be single or group. If the units served in-group, then their size can also be a random variable.

2.3 Queue discipline

Queue discipline indicates the manner in which units are selected for service from a queue. The most common queue discipline is first come first served (FCFS) or first in first out (FIFO), however some queue last come first served (LCFS) or selection in random order (RSS) are also adopted. In many congestion situation units in some classes gets priority in service over others. Generally, there are two priority disciplines, first one is preemptive and the second one is non-preemptive. In preemptive discipline customer with highest priority is allowed to join the service immediately by stopping the service of lower priority customer, while in non preemptive discipline

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 12, December 2015

service of the served unit is not stopped, only customer with highest priority is allowed to go to head of the queue.

2.4. System capacity

The space or room where units wait is known as capacity of the system. System capacity is one of the significant factors for consideration. If there is finite limit to maximum system size, then we can say capacity queuing system is finite otherwise infinite.

2.5. Number of service channel

A queuing system having more than one parallel service channel is known as multi-server queuing system. In design of multi server queuing system, there may slight variation by feeding the multiserver queuing system by single queue or having separate queue for each channel.

3. POISSON DISTRIBUTION

A random variable X is said to have Poisson distribution with parameters λ , then the probability density function $f_X(x)$ and distribution function $F_X(x)$ of the variable X is given by

$$P(X = x) = f_X(x) = \frac{\exp(-\lambda)\lambda^x}{x!}, x = 0,1,2,...$$

$$F_X(x) = P(X \le x) = \sum_{i=0}^{x} \frac{e^{-\lambda}\lambda^i}{i!}$$

The mean and variance of the Poisson distribution is

$$E(X) = \lambda = Var(X)$$

During the analysis of queuing systems Poisson distribution has important place. The distribution of arrival of units in most of queuing system is represented by Poisson distribution. If the arrival of units occurs according to Poisson distribution then the distribution of inter arrival time of units follows exponential distribution.

3.2. Exponential distribution

The exponential distribution is used to specify the distribution of continuous random variable. A random variable X is said to have exponential distribution with parameters μ , then the probability density function $f_X(x)$ of the variable is given by

$$P(X = t) = f_X(t) = \begin{cases} \mu \exp(-\mu t) & \text{for } 0 < t < \infty \\ 0 & \text{otherwise} \end{cases}$$

The mean and variance of the exponential distribution is

$$E(X) = \frac{1}{\mu}, Var(X) = \frac{1}{\mu^2}$$

One of the basic property of the exponential distribution is memory less property due to which it is generally used to specify the service time unit in the queuing system.

3.3. Erlangian distribution

A random variable X is said to have Erlangian distribution with parameters μ and k , then the probability density function $f_X(x)$ of the variable is given by

$$f(x; k, \mu) = \frac{(\mu k)^k x^{k-1} \exp(-k\mu x)}{(k-1)!} \quad \text{for } x, \mu \ge 0$$

The parameter k is called shape parameter, which is a non negative integer while λ is parameter of the distribution. The mean and variance of erlangian distribution is given by

$$E(X) = \frac{1}{\mu}, Var(X) = \frac{1}{k\mu^2}$$

For any particular value of k , the resulting Erlang is referred to as an Erlang of type k or E_k distribution. Erlang family of distribution provides much more modeling flexibility than the exponential distribution. If we take $k = 1$,then the Erlang distribution reduces into exponential distribution and if we take $k \to \infty$ then Erlang becomes deterministic with value $\frac{1}{\mu}$

4. STATE DEPENDENT QUEUING SYSTEMS

In general, during analysis of queuing system, it is assume that arrival rate of the customer is constant. However, this assumption of constant arrival rate may not valid in all congestion problems. To understand we site the banking, communication networks , manufacturing and production system, healthcare system etc, where the arrival rate of units depends upon the length of queue, server status (i.e. whether server is working, break down or on vacation state) and on rate with sever provides the service . In addition to this, in some queuing problems service rate of the server may also influenced with the, amount workload present, level of balking (reneging) and priority level of the customer.

In our day-to-day life, we encounter many queue situations where state depended queuing models play vital role in resolving the congestion problem. State dependent queuing models represent the congestion problems more closely and accurately. One can easily observe utility of state dependent modeling in almost all part of real life queue as well as industrial queue. Thus, state dependent modeling creates an important realm of the queuing theory. Many real life practical scenarios inspire and motivate us to study the state dependent queuing models. To illustrate we site the queuing system where server is human being, the amount of workload present directly influenced the human's productivity; in queuing system with finite capacity (such as lift, communication and transmission etc) the arrival rate immediately reduced to zero when number of customer exceed the capacity of the system.

One of the basic tasks that have to accomplish by a growing business or industry is to understand the customer need and to adopt the policy according to the customer requirement. State dependent arrival and service are two key features of almost all real life queues, manufacturing systems as well as telecommunication systems. Due to wide applicability, state dependent systems receive significant amount of attention by many researcher working in area of queuing theory. Many prominent

researcher study the state dependent modeling to design and upgrading the industrial queuing problems. In state dependent queuing system either arrival rate or service rate or both depends upon the system parameters such as queue length, workload, server states, impatient behavior of customer and many more parameters of the system.

5. METHODOLOGY

Once the mathematical model is formulated, it is vital important to analyze the problem by using appropriate technique to obtain performance measures. Some commonly used techniques are as follow:

5.1. Supplementary variable technique

The queuing process where the inter arrival time and service time of a unit follow exponential distribution, then the queue size $N(t)$ can be easily modeled by using markov process as the both distribution posses memory less property, which is the primary requirement for markav process. However, if either of distribution follow general distribution then we cannot model the queue size $N(t)$ using markav process as the new distribution may not have memory less property. Such process in which either of distribution does not follow exponential distribution is called non markav process.

Supplementary variable technique is one of powerful and elegant technique that is used for non markav process. The supplementary variable technique is generally used for non markovian process in continuous time. However, in many queuing problems, this technique can be used in steady state for more readily treatment. In this technique one or more supplementary variable introduced to convert the non markovian process into markovian process. In supplementary variable technique approach, the supplementary variables are introduced corresponding to either elapsed time or remaining time of the random variable. When the service time follow general distribution then SVT introduces corresponds to service time and when arrival follow general distribution then SVT introduce for inter arrival time. This technique first time used by Cox (1955) to study the $M/G/1$ queuing model. We explain this technique by assuming that service distribution follows the general probability law. Then queue size $N(t)$ becomes non Markovian as the service distribution does not possess memory less property. We introduce a new random variable $X(t)$, denotes the elapsed service of the customer that is in the service at time t. Then it can easily shown that the joint stochastic process $(N(t), X(t))$ convert in markovian process.

5.2. Probability generating function method

Consider a discrete random variable X assuming non negative values say $q_n = P(X = n),\ n = 0,1,2\ldots$,then the probability generating function of random variable X can be defined as

$$Q(z) = \sum_{n=0}^{\infty} q_n z^n = \sum_{n=0}^{\infty} P(X = n) z^n \text{ , } |z| \le 1 \text{ with assumption}$$

that $Q(1) = 1$ and $Q(0) = q_0$

As probability generating function are uniquely defined by their discrete sequence of probabilities. So probability generating functions can be used in mathematics to compact representation of the sequence of probabilities which can further employed to getting more generic solution by using well developed theory of power series or with some other technique.

Probability generating function technique is a powerful technique that is widely used in queuing system to provide the solution of differential difference equations. In this technique firstly a closed expression for probability generating function is evaluated by using differential difference equations for different probability state, then power series expansion or laplace –Steiltjes technique are used to find different probability state. In some models, it's quite difficult to obtain the series expansion from closed form of probability generating, however still it's provide valuable information of the system.

5.3. Maximum entropy principle

The maximum entropy principle is elegant technique, which is used to estimate the unknown probability distribution, when the partial information of the distribution available. Maximum entropy approach was first used by Shannon (1948) in information theory. The maximum entropy principle provides a means to obtain least-biased statistical inference when insufficient information is available. The maximum entropy principle is widely used in queuing theory to select a appropriate probability distribution of a queuing situation. In queuing theory, generally, we first select the probability distribution to specify the arrival distribution of units, service time of the units and batch size etc and then using the structure of death birth process we derive the performance measures of the queuing system. Some time these distribution not known completely, but only partial information in terms of moments is available. In such situation, we can obtain the unbiased distribution of concerned queuing system by maximizing the entropy function in terms of known performance indices. Many researcher used maximum entropy principle to analyze complex queuing model, Some important works in this regard were done by Ferdinand (1970), Shore (1982), El-Affendi and Kouvatsos (1983), Arizono et al. (1991), Wang et al. (2002), Wang et al. (2005) etc.

5.4. Runge-Kutta method

Due to the complex nature of the differential equations governing the queuing models, most often it is not easy to find the analytical solution in particular when the equations having the transient probabilities of the system states. In such cases, the numerical techniques can be used to find the solution of the set of differential equations. The method developed by Runge and Kutta (cf. Iwaarden, 1985) is one of the techniques to provide the numerical solution of a set of differential equations. The fourth order Runge-Kutta method is the most popular one and many researchers have used it to obtain the solution of a set of differential equations governing the queuing model.

Various computer softwares namely MATLAB, Maple, Mathmatika, etc. are available for computational purposes to facilitate the numerical solutions of differential

equations. For the modelling of some queuing systems, MATLAB with routines ode45 based on 4^{th} and 5^{th} order Runge-Kutta method is commonly used to find the numerical solution of the system of differential equations (cf. Ingolfsson et al., 2007).

6. SOME BASIC MODEL

6.1. Vacation queuing model

The queuing models in which servers are unavailable from service for random period of time is known as 'vacation queuing model'. During vacation time, server may perform supplementary jobs, being checked for repair, or simply take a break. Many real world queuing systems can be modeled as vacation model with different policies. The queuing model with server vacation has wide applicability in many areas of industrial problems including computer and communication, production and manufacturing systems, tele communication, etc. Thus, the server vacation models represent the more realistic and flexible picture of real life queues. The classification of vacation model can be done on the basis of following three aspects.

- Vacation startup rule determines when the server starts vacation. Start up rule can be divided into exhaustive and non-exhaustive services. In exhaustive type server cannot take a vacation until system becomes empty, while in non-exhaustive type server can take vacation even system is not empty.
- Based on termination of vacation, the vacation model can be classified into single vacation policy, multiple vacation policy, second optional vacation policy and threshold policy. In single vacation policy server takes only one vacation after each busy period while in multiple vacation policy server takes vacations until at least one customer waiting in the system at the vacation completion instant. In second optional vacation policy, server may take another optional vacation with some probability p, after completing the first phase of regular vacation. In threshold policy server keeps vacation until N (threshold level) customer accumulate in the system. In addition to this in multi server queuing system if all server takes vacation together then it is called synchronous vacation policy and if individual server take vacation independently then it is called asynchronous.
- The duration of the server vacations may be deterministic or independently identically distributed random variable. More vacation policy can be obtained by possible combination of above vacation policy.

Some of basic vacation model are investigated by Baba (1986), Doshi (1986), Takagi (1991), Zhang and Vickson (1993), Lee et al. (1995), Li and Zhu (1996) , Borthakur and Choudhury (1997) and Chao and Zhao (1998).

6.2. Break down model

In classical queuing models, it is a common assumption that the service station is not subject to failure, but in much real life congestion situations the role of server is performed by automatic devices, which are subject to failures and require repair or replacement of the server for the smooth functioning of the system. Break down of the server making negative impact on the performance measures of any queuing system; hence it is necessary to consider the breakdown of server an important factor. Depending upon failure conditions break down can classified into active and passive break down. If the server fails during its busy period then it is said that active break down, however if break down occurs during idle state it is said to be passive break down. Queuing model with server break down and repair have received significant attention by many researcher working in area of queuing theory and reliability theory. Some of pioneer work in this direction are done by Wartenhorst (1995), Li et al. (1997), Tang (1997), Gray et al. (2000) etc.

6.3. Two phase service

Queuing models in which customer require additional second stage service after completion of main service is referred as two-phase service queuing model. Phase modeling has its own importance as in many real life systems service completed in number of phases. The second stage service may be essential or optional. The utility of queuing model with essential second stage service can be observed in many real time system namely in manufacturing system where in machine producing certain items require two phase of service in succession. In some queuing model second stage service required optionally, for example at the teller counter of the bank, customer visits for their cash transaction, however some of them may ask for update of their passbook apart from their essential credit /deposit of the cash.

7. PERFORMANCE MEASURES

Performance measures of the queuing system are very critical and important aspect. The mathematical analysis of queuing system done only to obtain the various performance measures which can be use to determine the measure of effectiveness of the given process. These performance measures used by queuing analyst to design an 'optimal' system and to determine the values of appropriate measure of effectiveness for real world congestion situation. Performance measures obtain by analyzing the queuing system can be further use to upgrading the system by giving proper direction to ensure that proper level of service is provided by service facility. The commonly used performance measures during the analysis of state dependent queuing model are as follows.

- The state probability P_n is described as probability of n customers residing into the system either being served or waiting. Thus
  $P_n = \Pr\{\ n\ customer \quad in\ system\ \}$
- The traffic intensity ρ of the system is given by the ration of arrival rate and service rate. Thus $\rho = \frac{\lambda}{\mu}$
- The server utilization u may defined as the proportion of time, a server or a group of server may busy and is given by $u = \frac{\lambda}{m\,\mu}$ Where λ, μ and m denotes arrival rate, service rate and number of server

respectively. $u < 1$ is the stability condition of a queuing system.

- Throughput σ of a system is defined as the average number of customer leaving the system. throughput rate is defined as

$$\sigma = \sum_{n=1}^{\infty} \mu_n p_n$$

- Reliability of a system is defined as the probability that the system will perform efficiently throughout the interval $(0, t)$ under operating conditions. The availability of the system can be defined as time for which system is able to perform its function under the assumption that it is operated and maintained as per prescribed condition.
- The average number units waiting in the queue for getting service is termed as mean queue length and is denoted by Lq while the total number of units waiting in the queue plus number of units being served is termed as average queue length in the system L_s. Depending upon system being study both measure have their own importance. First one is require in order to determine to design for waiting space while later one is require to know how many server may unavailable for use. It is also an important performance measure from customer point of view as everybody avoids joining the long queue.
- The average time duration spent by a customer in queue for getting his turn for service is referred as average waiting time in queue W_q while total time in queue for wait and service time is referred as average waiting time in the system W_s. Depending upon system being study both measure have their own importance. For example if we studying a model of amusement park then waiting time is only the time spent in queue that is W_q, while if dealing with repair of our machine then the waiting time is total down time that is W_s. It is also an important performance measure from customer point of view.

REFERENCES

1. Arizono, I., Cui, Y. and Ohta, H. (1991). An analysis of M/M/s queuing systems based on the maximum entropy principle. Journal of Operational Research Society **42(1)**: 69-73.
2. Baba, Y. (1986). On the $M^x / G / 1$ queue with vacation time. Operations Research Letters **5(2)**: 93-98.
3. Borthakur, A. and Choudhury, G. (1997). On a batch arrival poisson queue with generalized vacation. Sankhya: The Indian Journal of Statistics **59(3)**: 369-383.
4. Chao, X. and Zhao, Y.Q. (1998). Analysis of multi-server queues with station and server vacations. European Journal of Operational Research **110(2)**: 392-406.
5. Cox,D.R. (1955). The analysis of non markovian stochastic processes by the inclusion of supplementary variables. Proc.Cambridge Philos. Soc **51**:433-441.
6. Doshi, B.T. (1986). Queuing systems with vacations: A survey. Queuing Systems **1(1)**: 29-66.
7. El-Affendi, M.A. and Kouvatsos, D.D. (1983). A maximum entropy analysis of the M/G/1 and G/M/1 queuing systems at equilibrium. Acta Information **19(4)**: 339-355.
8. Ferdinand, A.E. (1970). A statistical mechanical approach to system analysis. IBM Journal of Research Development **14(5)**: 539-547.
9. Gray, W.J., Wang, P.P. and Scott, M. (2000). A vacation queuing model with service breakdowns. Applied Mathematical Modelling **24(5-6)**: 391-400.
10. Ingolfsson, A., Akhmetshina, E., Budge, S., Li, Y. and Wu, X. (2007). A survey and
11. experimental comparison of service-level-approximation methods for non stationary
12. $M(t) / M / s(t)$ queuing systems with exhaustive discipline. INFORMS Journal on
13. Computing **19(2)**: 201-214.
14. Lee, S.S., Lee, H.W., Yoon, S.H. and Chae, K.C.(1995). Batch arrival queue with N -policy and single vacation. Computers & Operations Research **22(2)**: 173-189.
15. Li, H. and Zhu, Y. (1996). Analysis of $M / G / 1$ queues with delayed vacations and exhaustive service discipline. European Journal of Operational Research **92(1)**: 125-134.
16. Li, Q.L., Xu, D.J. and Cao, J. (1997). Reliability approximation of a markov queuing system with server breakdown and repair. Microelectronics Reliability **37(8)**: 1203-1212.
17. Shannon, C.E. (1948). A mathematical theory of communication. Bell System Technical Journal **27**:379–423.
18. Shore, J.E. (1982). Information theoretic approximations for $M / G / 1$ and $G / G / 1$ queuing systems. Acta Information **17(1)**: 43-61.
19. Takagi, H.(1991). Queuing analysis: a foundation of performance evaluation. Vacation and priority systems.**1(1)**, North Holland, Amsterdam.
20. Tang, Y.H. (1997). A single server $M / G / 1$ queuing system subject to breakdowns- some reliability and queuing problems. Microelectronics Reliability **37(2)**: 315-321.
21. Wang, K.H., Chuang, S.L. and Pearn, W.L. (2002). Maximum entropy analysis to the N -policy $M / G / 1$ queuing system with a removable server. Applied Mathematical Modelling **26(12)**: 1151-1162.
22. Wang, K.H., Wang, T.Y. and Pearn, W.L. (2005). Maximum entropy analysis to the N-policy $M / G / 1$ queuing system with server breakdowns and general startup times. Applied Mathematics and Computation **165(1)**: 45-61.
23. Zhang, Z. and Vickson, R.G. (1993). A simple approximation for mean waiting time in $M / G / 1$ queue with vacations and limited service discipline. Operations Research Letters **13(1)**: 21-26.
24. Wartenhorst, P. (1995). N Parallel queuing systems with server breakdown and repair. European Journal of Operational Research **82(2)**: 302-322.