Mathematical Study of Some Queueing Models

Manoj Kumar Jadoun, Associate Professor, D.A.V. (PG) College, Dehradun, India, jadounmanojkumar@gmail.com
Deepa Chauhan, Associate Professor, Axis Institute of Technology & Management, Kanpur, India, deepachauhan07@axiscolleges.in

Abstract
The queueing systems are prevalent throughout the society. The adequacy of those systems can have an important effect on quality of life and productivity. Since the advent of industrial revolution to computer-communication era, the world has observed remarkable growth in the size and complexity of organizations of today. An integral part of these revolutionary changes has resulted in a tremendous increase in the division of labour, resources, infrastructure or segmentation of management responsibilities. This leads in blocking up of resources with consequent losses not only in time but in money too. In this world with increasing population, the advances of service and technology have contributed a lot to the increasing needs and desire of the people to have more and more material, comfort and satisfaction in life. Philosophically, it is well said that, life is a big queue where people are constantly waiting the next happening.

Introduction

One important class of queueing system is commercial service system where customers receive service from commercial organization such as barbershop, bank taller, checkout counters, cafeteria, etc. For illustration, queueing system at bank is depicted in fig. 1(a-b). Another important class is transportation service system in which vehicles may wait at a tollbooth at traffic light, for loading and unloading, at parking lots etc. In recent years queueing theory has been applied to business-industrial service systems that include material and labour handling, maintenance systems, machine repair systems, manufacturing units, wage incentive plans or inspection systems etc. It has significant impact on the design of inventories and production control systems. Queueing theory is also applicable to social service systems, namely a judicial system, legislative system where court or cabinet tries to provide service to cases or bills respectively. Various health care systems are also queueing systems providing availability of emergency room, ambulance, X-ray machines, beds or doctors etc.

The computer communication field is rich with queueing problems that occur in wide area networks (WAN), packet radio network and local area network (LAN), etc. A significant component in many of their applications is that of multi-access to common resources. Now a days broadcast communication, multimedia services or on-line transaction processing systems
**Fig. 1 a:** Bank layout showing parallel serves fed by a single queue

**Fig. 1 b:** Queue diagram of Bank
have grown rapidly with ever-stringent performance requirements due to advent of VLSI technology, coupling with multi-microprocessors having considerable interest in geographically distributed databases with concurrency control. Future advances in networking coupled with the rapid advances in storage technologies will make it feasible to build multimedia-on-demand servers that provide service similar to those of neighborhood videotape rental stores on a metropolitan area network (MAN). A critical requirement in building of a multimedia server is the need for guaranteeing continuous playback of media streams.

The present age is of communication and information technology. The traffic modeling provides us the basis for the analysis and design of efficient communication and information systems. Much work has been done in this area using various types of mathematical modeling via queue theoretic approaches. The global change in computer communication system has made it easy to interact with any one throughout the world. A variety of the queueing models have received considerable attention from both the researchers, interested in theoretical developments and professionals/practitioners interested in practical applications, in order to cope up with the advancement of technology.

Common to all of these cases are the arrivals of people or jobs requiring service and delays when the service station is busy. The queueing models are basically relevant to service oriented organization and suggest ways and means to improve the efficiency of the service. An improvement in service level is always possible by increasing the number of servers, but immediate consequence is unutilized idle time that increases the cost of the service system. Queuing methodology indicates the optimal usages of existing resources and other available resources to improve the service. At a slow service system, queue builds up and the cost of queue system increases in terms of social cost, lost customers or so forth. Therefore the ultimate goal is to achieve an economic balance between the cost of service and the cost associated with the waiting for that service.

At present, there is no unified body of optimization theory for waiting line models. However, since it is often impossible to accurately predict when units will arrive to seek service and/or how much time will be required to provide that service, these decisions often are difficult ones. The queueing models can help us in making prediction about the behavior of the system in terms of its service, estimation of possible congestion and methods to eliminate them.

Performance is one of the fundamental factors in the design, development and configuration of service systems. Performance evaluation therefore has been and continues to be of great practical importance in the service industry. Of the two main approaches to performance evaluation, namely measurement and modeling, modeling has received the greatest research attention. Inspite, measurement is accurate and credible, it may not be feasible in the design and development stages of the system. But queueing modeling is used to aid decision making in an effective and systematic way when measurement is intractable.

In this thesis we study various queueing systems by formulating mathematical models of their operations and then using these models to derive various measures of the performance. Our study will provide vital information for effectively designing queueing systems that achieve an appropriate balance between the cost of providing services and the cost associated with waiting for those services.
Waiting Line Model

Queueing system can be described as composed of customers arriving for service, waiting for service if it is not provided immediate, and leaving the system after being served. The characterization of basic waiting line model can be done by considering the following factors:

Input Process

The usual description of the pattern of arrivals into the system is given by the probability distribution of the time between successive arrival events and the number of individuals or units that appear at each of those events. There may be single or bulk arrival from infinite or finite population.

In some situation, a customer seeing a long line or insufficient waiting space may balk, i.e., not joins the line. Some times, the customer often joining a queue may decide to leave the system due to impatience. This behaviour of customer is termed as reneging. The customers may jockey from one waiting line to another in the hope of better service. Thus input process may depend in part on the status of the queueing system.

Service Discipline

This characteristic describes the order in which customers entering the system are eventually served. For example it may be first come first served (FCFS), last come first served (LCFS), service in random order (SIRO), general service discipline (GD), processor sharing (PS), priority, etc.

Service Mechanism

A specification of the service mechanism includes description of time to complete a service, and of the number of individuals whose requirements are satisfied at each service event. The service mechanism also prescribes the number and configuration of servers or channels. e.g. in parallel or in series (tandem).

Performance Measures

Queueing theory tries to develop a mathematical representation of a system based on some assumptions. It is worth to have the knowledge of performance indices to measure the effectiveness of a service system. It may be user oriented or system oriented measure depending on standpoint of interest. From the user’s standpoint, performance is often judged in terms of how quickly the results are returned from servicing while user primarily focuses attention on the progress of individual units through the system. The system itself is more concerned with the collective behaviour and adopts a global view of the situation.

The amount of work done by a queueing system is often measured in units of time that the arrival of each unit brings into the system. Average quantity of work is equal to its mean service time. For prediction of system characterization, the following measures of effectiveness and measures of performance are used:
Traffic Intensity

The traffic intensity measures the total service demand on the system per unit time and equals the product of the average arrival rate of units and the mean service requirement.

Throughput

The throughput of a system refers to the average number of units served by system per unit time for a non-loss stable system. Throughput is inversely proportional to queue length either due to customer in attention or server fatigue.

Utilization

If we denote the mean arrival rate and service rate by $\lambda$ and $\mu$ respectively, then for m server systems, the server utilization $p$ is defined as the fraction of time that the server is busy. Explicitly it is defined as

$$ p = \frac{\lambda}{m\mu} $$

Response Time

The elapsed time between the moment of unit submission and the moment the results are delivered is said to be response time. The waiting time is defined to be the non-service component of the response time. The probability distribution of this continuous variable reflects the waiting cost to the customers and is important indicator of the performance.

Average number of units in the system

The number of units waiting in the queue including the one being served at a particular time is called average number of units in the system.

Expected busy period

This is defined as the amount of time that system has to work continuously commencing form the moment of unit arriving to an empty system and termination at the moment when the system next becomes empty, following the departure of the job.

Queueing Methodologies

Stochastic processes are related with the sequence of events governed by probabilistic laws which are commonly used for modeling queues that are formed in a number of real life situations namely in service system, machine interference problems, communication systems, computer systems etc. A brief account of stochastic processes is as follows:

Stochastic Processes

A family of random variables $\{X(t) : t \in T\}$, defined on a given probability space indexed by the parameter $t$, where $t$ varies over an indexed set is called Stochastic Process.
The values assumed by the random variable $X(t)$ are called states and a set of all possible values forms the state space of the process. If the state space of a stochastic process is discrete or countable then it is called a discrete state process. Alternatively, if the state space is continuous, then we have a continuous state process.

**Markov Process**

A stochastic process $\{X(t) : t \in T\}$ is called a Markov Process if for any set of time points $t_0 < t_1 < \ldots < t_n < t$, the conditional distribution of $X(t)$ for given values of $X(t_0), X(t_1), \ldots, X(t_n)$ depends only on $X(t_n)$, the immediate preceding value. Mathematically, we have

$$P[X(t) \leq x \mid X(t_n) = x_n, X(t_{n-1}) = x_{n-1}, \ldots, X(t_0) = x_0] = P[X(t) \leq x \mid X(t_n) = x_n]$$

A Markov process with discrete state space and discrete index set is said to be Markov Chain.

**Poisson Process**

It is the stochastic process $\{N(t): t > 0\}$ with mean $\lambda t$ where $N(t)$ denotes the number of successive events which are independent or identically distributed in time interval $(0, t]$ according to exponential distribution

$$F(x) = 1 - e^{-\lambda x}$$

The Poisson process is a continuous parameter discrete state process that is most often used for determining the number of units in the service system.

**Birth Death Process**

A continuous parameter homogeneous Markov Chain $\{X(t), t \geq 0\}$ with the state space $\{0, 1, 2, \ldots\}$ is known as a birth death process if there exist constants $\lambda_i$ (i = 0, 1, 2, ...) and $\mu_i$ (i = 0, 1, 2, ...) such that the transition from state $q_k$ to $q_{k+1}$ signifies birth with rate $2\lambda_k$ and from state $q_k$ to $q_{k-1}$ signifies death with rate $\mu_k$. Birth death process plays an important role in studying many queueing situations.

Let $P_k(t)$ be the probability that the population size is $k$ at time $t$. The following differential difference equations are called Chapman Kolmogorov equations that represent the dynamics of the system

$$\frac{dP_k(t)}{dt} = -\left(\lambda_k + \mu_k\right)P_{k-1}(t) + \mu_{k+1}P_{k+1}(t), \quad k \geq 1 \quad \text{..... (1)}$$

$$\frac{dP_0(t)}{dt} = -\lambda_0 P_0(t) + \mu_1 P_1(t) \quad \text{..... (2)}$$

**Renewal Process**

A process $\{X(t); t > 0, 1, 2, \ldots\}$ whose state space belongs to a denumerable set $\{0, 1, 2, \ldots\}$ and for which $X_0, X_1, X_2, \ldots$ are non-negative independent identically distributed random variables, is said to be renewal process.
Stationary Process

A stochastic process $X(t)$ is said to be stationary if $P\{X(t); t \in T\}$ is invariant to shifts in all time for all values of its arguments; i.e., given any constant $t$, the following must hold

$$P\{X(t + \tau)\} \neq P\{X(t)\}$$

Independent Process

A stochastic process $\{X(t); t \in T\}$ is said to be an independent process provided its nth order joint distribution satisfies the condition

$$F(x,t) = \prod_{i=1}^{n} F(X_i,t_i) = \prod_{i=1}^{n} P\{X(t_i) \leq X_i\}$$

The System States

The fundamental concept in the analysis of a waiting line system is that of a state of the system. It involves study of a system’s behaviour over time. It is classified as follows:

Transient States

A system is said to be in transient state when it’s operating characteristics depend upon time. This state occurs at the beginning of the operation of the system.

- **Steady State**

A system is said to be in steady state when its operating characteristics become independent of time. This state occurs in the long run of the system.

- **Explosive State**

If the arrival rate of the system is more than its servicing rate, the length of the queue will go on increasing with the time and will tend to infinity as $t \to \infty$, this state of the system is said to be explosive state.

Some Queueing Techniques

An analytical approach is used to obtain the explicit expression, but, numerical methods provide an alternative approach using inherent structure to analyze the queueing system. Model is first formulated mathematically as a continuous time markov process with discrete states resulting the birth-death Chapman-Kolomogorov differential equations. The model with stationary distributed Markov process can also be represented as stochastic matrix given by

$$\pi Q = 0^T$$

where $\pi$ represents probability vector or in terms of characteristics value and characteristics vector
\[ \pi^T (1 + Q_t) = \pi^T \quad \cdots (5) \]

where \( \pi_t \) can be considered as the left eigen vector corresponding to the unit eigen value of \( P = 1 - Q_t \), where \( P \) will be stochastic matrix if ~ is chosen such that \( t \pi \max_1 \{q_{ii} \} \) and have all rows sum equal to 1. Subsequently its transient or steady state probability vector is calculated using well-known equation solving techniques. Some of them are briefly discussed here.

**Product Form Solution**

The solution of the birth-death equations (1.1) - (1.2) in steady state can be given in product form.

Rearranging the equation (1.1), we get

\[
\lambda_k P_k - \mu_{k+1} P_{k+1} = \lambda_{k-1} P_{k-1} - \mu_k P_k = \ldots \lambda_0 P_0 - \mu_1 P_1
\]

Thus,

\[
P_k = \frac{\lambda_{k-1}}{\mu_k} P_{k+1}
\]

so that

\[
P_k = P_0 \sum_{i=0}^{k-1} \frac{\lambda_i}{\mu_i + 1} \quad k \leq 1 \quad \cdots (6)
\]

where \( P_0 \) can be determined using normalizing condition

\[
\sum_{k=0}^{\infty} P_k = 1
\]

**Generating Function Method**

Generating function is a tool that derives the set of probability values for discrete distribution. If the values \( P_0 \), \( P_i \) form a probability distribution then the probability generating function of that distribution is given by

\[
G(z) = E [z^i] = \sum_{k=0}^{\infty} z^k P_k \quad \cdots (7)
\]

The convergence of the series is guaranteed for all \( |z| \leq 1 \). The generating function contains all the information about probabilities. We can extract the value of any probability by using repeated differentiation, as

\[
P_j = \frac{1}{j!} \frac{d^j G(z)}{dz^j} \bigg|_{z=0} \quad \cdots (8)
\]
The probability generating function is also known as z transform.

**Direct Method**

This is a numerical method, which yields the exact results in a finite number of arithmetic operations. But in practice, it produces approximate results due to round off errors. Gauss elimination method is the examples of direct method.

**Iterative Methods**

The iterative method for the solution of a system of linear equations governing steady state queueing models is frequently used. Iterative method starts with an initial approximation and proceeds with appropriate algorithm to obtain successively better approximation. By using this method we preserve the sparsely parameter matrix. Successive convergence to the desired solution is also provided by an iterative algorithm. A good initial estimate can speed up the computation considerably. The iterations can be terminated by imposing a pre-specified tolerance level. Round-off errors are also avoided by this method because the parameter matrix is not altered in this method.

The power method, Gauss Seidel and Successive Over Relaxation (SOR) techniques, are some iterative methods for solving the system of linear equations describing the queueing problems.

**Matrix-Geometric Method**

If the state space can be described by a pair of integers (i, j) with only i being unbounded, then, it is possible that a matrix geometric solution may exist. This procedure can be used when the transition rate matrix has a particular block (lower or upper) Hessenberg structure that occurs if rate matrix has block tri-diagonal structure of block sub-matrices. With the help these submatrixes, the model is structured as square matrix of infinite dimension that converges to finite dimension matrix using the minimal matrix to get recursive relation of probability vectors.

**Diffusion Approximation**

When random variables defining the system change continuously instead of discretely, the system is characterized by a probability density function that satisfies a second order partial differential equation called Focker-Planck equation or diffusion equation. For the diffusion approximation, we propose that the arrival process and the departure process are both to be approximated by continuous random processes that at time t are normally distributed.

**Conclusion**

The useful papers explored more general distribution of arrivals and service, various priority rules for service, open and closed queueing network, blocking of queueing systems and other practical queues or queue like situations. Closer to the present time, we note that theoretical developments were driven by technological changes. Computer, computer communication, automation and robotics, automatic storage and retrieval devices, all provided challenges and opportunities. The theoretical studies of the fifties and sixties provided a sound bases for meeting the challenges. Infact, queues appear in all phases of life i.e. waiting lines are omnipresent. To the contrary, queueing systems are surprisingly prevalent in a wide
variety of contexts. To broaden our horizons on the applicability of queueing theory we shall briefly mention various examples of real life queueing systems.

References