

Modelling Finite and Infinite Queueing Processes

¹Mustapha, Adeniyi Mudashiru, ²Oyeyemi, G.M.

¹Dept. of Business Administration, University of Ilorin, Ilorin, Nigeria

²Dept. of Statistics, University of Ilorin, Ilorin, Nigeria

Abstract

This paper considers an infinite queue where the arrival of future customers is not affected by the numbers of customers already on the queue and a finite queue where customers attempting to enter when people are already present are denied entry, or blocked. In particular, we analyze the problem based on the assumption that one or more elements of the queueing system can be expressed only in probabilistic terms. To study the effect of blocking arriving customers, we first establish some stochastic comparison of the performance measure formulas that permit us to compare the different operating policies. These comparisons show that in M/M/1 queueing model, there is an infinite room to hold arrivals waiting for service. M/M/K/F queueing model has a limited capacity.

Keywords

Infinite Queue, Finite Queue, Poisson Distribution, Stochastic Modeling

1. Introduction

The most important elements of any queue model are the arrival rate and service rate. A queue model is based on certain assumptions. Any attempt to relax these assumptions would lead to a general model. The vast majority of the queueing models are, however, based on the assumption that one or more elements of the queueing system can be expressed only in probabilistic terms. Hence, nearly all the queueing models are of probabilistic or stochastic type.

A model is stochastic model if it has random arrival and service rates. Stochastic modeling according to Vohra (2007) assumes that;

- Arrival rate has Poisson distributions
- Service rate follows-exponential distribution;
- Service patterns is first-come-first=serve;
- The system is a steady state (i.e. at equilibrium);
- Calling population is infinite;
- An infinite queues exists.

Arrivals might follow any pattern. The frequently employed assumptions, which adequately support many real world situations, is that the arrivals are Poisson distributed. Lawrence and Pasternack (2002) gave the following conditions for a Poisson arrival process.

Orderliness – In anytime instant, at most one customer will arrive at the service facility.

Stationarity – For a given time frame the probability that a customer will arrive within a certain time arrival is the same for all time intervals of equal length.

Independence – Customers arrive independently of one another; that is, an arrival during a given time interval does not affect the probability of an arrival during other time intervals.

Queueing theory has been used extensively in the banking industry to increase businesses while at the same time alleviating both the actual and perceived amount of time a customer spends waiting in line (Wayne and DiSotu 1994). Queueing theory has been applied to computer simulation models to help with business decisions and problems (Proctor 1994). Daellenbach and George (1978) opined that the number of potential uses of waiting line or queueing theory

is very large; two major types of situation give rise to successful economic applications. The first according to them deals with the case where an organization controls a sufficiently large number of similar or identical service facilities, such as bank tellers, machine operators or telephones exchanges. Queueing theory may offer only small economic incentives for each facility by itself; it is the pooling of large numbers of individually small gains that make such applications economically worthwhile. The second type of application deals with the planning and design of single facilities involving large capital investments, such as the purchase and operation of part facilities, or computer installation.

The objectives of this paper therefore are:

- To highlight the main objective of queueing model
- To consider and compare M/M/1 queueing model with M/M/K/F queueing model
- To suggest which of the models is appropriate to allow customers to come into the system without any restriction.

A. Infinite Calling Population

The analysis of queues being considered is that of infinite calling population. Usually an infinite calling population can be assumed if the rate of arrival of future customers is not affected by the number of customers already on the queue.

Arrivals following random patterns are usually described by the statistical Poisson distribution (T) being referred to, represents the total time period under consideration. Robert ThieRauf-et al (1986) expressed this as:

$$f(n, \lambda, T) = \frac{\lambda^n e^{-\lambda T}}{n!}$$

It approximates a normal distribution and is skewed to one side. The Poisson assumption is appropriate whenever:

- The number of an arrival at any instant is very small but there are many opportunities for it to happen.
- Unit arrives at random independent of each other, with the same average rate λ .
- The probability of two or more arrivals in a very short time, say one second, is almost zero.

B. Objective of Queueing Model

The objective of queueing modeling analysis according to Grose et al (1974) and Griffin (1978) is to minimize total costs of waiting and providing services. It is therefore the responsibility of managers to balance these set to make any economic queueing system decision.

The optimum service facility that minimizes costs is given as:

$$\mu = \lambda + \sqrt{\frac{C_w \lambda}{C_s}}$$

Where

μ = Average service rate

λ = Average arrival rate

C_w = Cost of waiting in the queue.

C_s = Cost of service per unit of time

For the purpose of this study and to evaluate the model these notations and variables according to Lee (1966) and Vohra (2007) shall be used: -

λ = Average arrival rate

μ = Average Service rate

L_s = Average number of customers in the system

W_s = Average waiting time in the system

L_q = Average number of customers waiting for service in the queue

W_q = Average time a customer spends waiting for service

P = Traffic intensity or utilization factor

P_0 = Probability that the system is idle

P_1 = Probability of having exactly one customer in the system.

P_2 = Probability of having exactly 2 customers in the system.

S = Number of channels

σ = Standard Deviation

P_n = Probability of having n number of customers in the system

$\mu = \lambda$ = Upper limit of customers in the system

The mean arrival rate for a unit of time is represented by λ , and $1/\lambda$ is the mean service time, as well as the standard deviation of service times. This implies again that the service times are independently distributed, no matter when they are undertaken, no matter how many arrivals are waiting in the queue, and no matter what the previous service history has been.

The arrival rules follow a Poisson distribution while the service time is exponentially distributed.

To derive operating characteristics for this type of waiting line, we must also make several other assumptions. First, there must be only one service channels, which arrivals enter one at a time. Secondly, it is assumed that there is an infinite population from which arrivals originate. It is also assumed that there is an infinite room to hold arrivals waiting for service. Finally it assumed that arrivals are served on a first-come-first served basis.

C. M/M/1 Queuing Model

To evaluate the model, we begin with the first question whether the service station can handle the customer demand for service. To answer this question, Vohra (2007) posited that it lies in the value of μ and λ . If $\lambda > \mu$, the waiting line would increase without limit, leading to the breakdown of the system ultimately. For a workable system, it is necessary that $\lambda < \mu$. It is interesting to note that it is impossible for $\mu = \lambda$.

Two important relationships, known as little's formulae, exist that relate the average number of customers in the system, L_s , and the average customer waiting time in the system, W_s , as well as between the average number of customers waiting for service, L_q , and the average time a customer spends waiting for service, W_q (Lawrence and Pasternack 2002).

These formulas state that if a queuing system has a single waiting line, customers arrive at some finite mean rate, λ , and the necessary condition for steady state exists, then the following relationships hold:

$$L_s = \lambda W_s \quad \text{eq(1)}$$

$$L_q = \lambda W_q \quad \text{eq(2)}$$

For steady state models that involve a potentially infinite customer population, if all servers performed at the same rate, μ , then λ/μ represents the average number of customers being served. Hence,

$$L = L_q + \lambda/\mu \quad \text{eq(3)}$$

Therefore, knowing one of the values for L_s , L_q , W_s or W_q allows us to calculate the others.

The ratio $P = \lambda/\mu$ is defined as the traffic intensity or utilization factor. This indicates the proportion of times, or the probability, that the service station is busy. From this, the probability that the system is idle, that is, there are no customers in the system, equals $P_0 = 1 - P$.

We can show that the probability of having exactly one customer in the system is $P_1 = P_0$. Similarly, the probability of having exactly 2 customers in the system would be:

$$P_2 = pP_1 = p^2P_0 \quad \text{eq(4)}$$

To generalize, the probability of having exactly n customers in the system shall be:

$$L_s = \sum_{n=0}^{\infty} nP_n \quad \text{eq(5)}$$

This can be solved to obtain

$$L_s = \frac{\lambda}{\mu - \lambda} \quad \text{or} \quad \frac{P}{1 - P}$$

Expected number of customers in the queue shall be equal to the difference between the expected number of customers in the system and the expected number of customers being served. Now, since the server is busy, or is serving O unit, λ/μ of the time, and is serving O unit (i.e. it is idle) $1 - \lambda/\mu$ of the time, the expected number being served equals

$$1 \left(\frac{\lambda}{\mu} \right) + 0 \left(1 - \frac{\lambda}{\mu} \right) = \frac{\lambda}{\mu} \quad \text{eq(6)}$$

Consequently, expected number of customers in the queue.

$$L_q = L_s - \frac{\lambda}{\mu} = \frac{\lambda}{\mu - \lambda} - \frac{\lambda}{\mu} \quad \text{eq(7)}$$

$$= \frac{\lambda^2}{\mu(\mu - \lambda)} \quad \text{or} \quad \frac{P^2}{1 - P}$$

It is significant to note that L_q is the average length of all queues. The average length of non-empty queues, that is, those which contain at least one customer, L_q , is given by the following expression.

$$L_q = \frac{1}{1 - P} \quad \text{or} \quad \frac{\mu}{\mu - \lambda} \quad \text{eq(8)}$$

With an average arrival rate of λ , the average time between the arrivals is $1/\lambda$. Thus, the mean waiting time in queue, W_q , is the product of the average time between the arrivals and the average queue length.

$$\text{Symbolically, } W_q = 1/\lambda \cdot L_q \quad \text{eq(9)}$$

Substituting $\frac{\lambda^2}{\mu(\mu - \lambda)}$ for L_q and simplifying,

$$\text{we get } W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{P}{\mu - \lambda}$$

Similarly, the mean time in the system, W_s , is equal to the product of the average time between arrivals and the average number of customers in the system. Thus:

$$W_s = \frac{1}{\lambda} \times L_s \quad \text{eq(10)}$$

Putting $L_s = \lambda/(\mu - \lambda)$ in this equation and simplifying, we get

$$W_s = \frac{1}{\mu - \lambda}$$

Stated another way, since the mean service rate is μ , the average (expected) time for completing the service is $1/\mu$. Therefore, the expected time a customer would spend in the system shall be equal to the expected waiting time in the queue plus the average processing/ servicing time. Thus,

$$W_s = W_q + \frac{1}{\mu} \quad \text{eq(11)}$$

$$= \frac{\mu}{\mu(\mu - \lambda)} + \frac{1}{\mu} = \frac{1}{\mu - \lambda}$$

Which is the same as shown earlier.

Finally, the probability that a customer spends more than t units of time in the system, $W_s(t) = e^{-t}/W_s$. And the probability that a customer spends more than t units of time in the queue, $W_q(t) = pe^{-t}W_q$.

It is significant to note that although the time spent in the system is distributed exponentially and so is the service time, but the difference of the two, that is to say, the time spent in the queue, $W_q(t)$, is not exponentially distributed.

For the purposes of analysis, Lawrence and Pasternack (2002) summarized the performance measures for the M/M/I queues as follow:

$$P_0 = 1 - \left[\frac{\lambda}{\mu} \right] \quad \text{eq(12)}$$

$$P_n = 1 - (\lambda/\mu) (\lambda/\mu)^n$$

$$L = \frac{\lambda}{(\mu - \lambda)}$$

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

$$W = \frac{1}{(\mu - \lambda)}$$

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)}$$

$$P_w = \frac{\lambda}{\mu}$$

$$P = \frac{\lambda}{\mu}$$

D. M/M/K/F/ Queuing Model

An M/M/K/F model assumes a Poisson arrival process at mean rate λ , K servers, each having an exponential service time distribution with mean rate μ ; and an upper limit of F customers who can be present in the system at any one time. Customers attempting to enter when F people are already present (when the system is full) are denied entry, or blocked. The model assumes that blocked customers will leave the system forever.

The implication of blocking arriving customers is that, although an average of λ customers per hour may attempt to join the system, some will be turned away. Hence, only a fraction of these λ potential arrivals make it through the system. The effective arrival rate is denoted by λ_e . It is λ_e that must be used in little formulas

for the M/M/K/F case.

Arriving customers are blocked only if F customers are already in the system. Since the probability of F customers in the system is P_F , the following relationship exists between λ_e and λ ;

$$\lambda_e = \lambda (1 - P_F) \quad \text{eq(13)}$$

For finite queues with Markovian arrival and service processes, balance equations can be used to determine formulas for the steady-states performance measures. Because the queue is finite, the restriction that $\lambda < K\mu$ does not have to hold in order to achieve steady – state results.

Performance measures formulas for M/M/K/F are given below:

$$P_0 = \frac{1}{1 + \sum_{n=1}^K \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^K}{K!} \sum_{n=K+1}^F \left(\frac{\lambda}{K\mu} \right)^{n-K}} \quad \text{eq. (14)}$$

$$P_n = \frac{(\lambda/\mu)^n}{n!} P_0 \quad n = 1, 2, 3, \dots, K$$

$$\frac{(\lambda/\mu)^n}{K! K^{n-K}} P_0 \quad n = K+1, K+2, K+3, \dots, F$$

$$= \frac{P_0 (\lambda/\mu)^K (\lambda/K\mu) \left[(1 - (\lambda/K\mu)^{F-K} - (F-K) (\lambda/K\mu)^{F-K} (1 - (K\mu)) \right]}{K! (1 - (\lambda/K\mu)^2) + \sum_{n=0}^{K-1} n P_n + K \left[1 - \sum_{n=0}^{K-1} P_n \right]}$$

$$L_q = L - \left[\sum_{n=0}^{K-1} n P_n + K \left[1 - \sum_{n=0}^{K-1} P_n \right] \right]$$

$$W = \frac{L}{F - 1 - \lambda \sum_{n=0} P_n}$$

$$W_q = \frac{L_q}{F - 1 - \lambda \sum_{n=0} P_n}$$

$$P_w = 1 - \lambda \sum_{n=0}^{K-1} P_n$$

$$P = \frac{F - 1 - \lambda \sum_{n=0} P_n}{K\mu}$$

II. Conclusion and Recommendation

Queuing problems are of considerable interest in this world as a result of its perturbing consistency in both private and public service systems. Queuing theory rests upon some assumptions and complex mathematics. It encompasses a very large group of models

with each relating to a different type of queuing situations.

Problematic queuing system that are long lines arising from infinite calling population can lead to the customer's overall satisfaction with the service transaction. Finite Calling Population experiences frustration and opportunity costs because of deflecting source population units since limited waiting space exists. When we view the cost of attracting customers with the cost of blocking them from coming into the system, the effects of the later is far greater than the former.

The result of our generalized queuing model have been considered and compared. Our model therefore recommends M/M/I queuing model which allows customers to come in without any restriction. In spite of the identified problems, this model is considered appropriate if it could be properly managed.

References

- [1] Aremu, M.A., "Managing Queue to enhance customers' satisfaction in Nigerian Banking Industry", In the Nigerian Academic Forum, A multidisciplinary Journal, published by National Association of The Academics (N.A.A) Vol. 9, No. 1st October, 2005.
- [2] Aquilano, N., Chase, R., "Fundamentals of Operations Management", Prentice Hall, New Jersey, 1991.
- [3] Banjoko, S., "Techniques of Production Management", New Jersey: Prentice Hall Inc. 2nd edition pp. 415-431, 1989.
- [4] Buffa, E.A., "Operations Management Problems and models", Canada: John Wiley, 1972.
- [5] Chary, S.N., "Production and Operations Management. 3rd Edition. Tata McGraw-Hill Publishing Company Ltd. New Delhi, 2007.
- [6] Daellenbach, George, J.A., "Introduction to operations Research Techniques (Allyn and Bacon Inc). Boston, 1978.
- [7] Dervitsiotis, K., "Operations Management New York: USA Mc Graw-Hill", Book Company, New York, 1981.
- [8] Koontz, O. Donnell et al., "Management 7th Edition", MC. Graw Hill, pp. 238, 1983. Microsoft Operating Frame Work (2009).
- [9] Olaniyi, T.A., "An Appraisal of Cost of Queuing on Nigeria Banking Sector: A Case Study of First Bank of Nigeria Plc: Ilorin. In Ilorin Journal of Business and Social Sciences Vol. 9, No. 1& 2, 2004.
- [10] Proctor, R.A., "Queues and the Power of simulation: helping with business decisions and problems. Manage Decisions. 32:50 -5, in hospital pharmacy, Vol. 36, No. 3, pp. 275-279, 1994.
- [11] Skinner, W., "The focus Factory" Harvard Business in Aquiline, N and Chase R.: Fundamentals of Operations Management (1991). Boston; Richard D. Irwin, U.S.A., 1974.



Mustapha Adeniyi Mudashiru received his B.sc (Hons.) degree in Economics from University of Ibadan, Ibadan. Nigeria in 1984. M.B.A. from Obafemi Awolowo University, Ile-Ife. Nigeria in 1997. M.sc. Management Science from University of Ilorin, Ilorin. Nigeria in 2003. Ph.D. degree in Business Administration from University of Ilorin, Ilorin. Nigeria in 2010. He was a part-time lecturer, University of Ado-Ekiti, Ekiti State, Nigeria. Kwara State

College of Education Ilorin (Sandwich Programme) 1999-2004. Part-time lecturer, Imo State University Owerri, (Ilorin centre, Ilorin). 2000-2004. He is currently a Deputy Director, Independent National Electoral Commission (INEC). His research interest is in Operations Research.