INTERNATIONAL JOURNAL OF NATURAL AND APPLIED SCIENCES (IJNAS), VOL. 3, NOS.1& 2 (2008); P. 105 – 111, 3 FIGS.

A single queue single server Markov processes model

O. A. Ofem[*1], E. E Williams[1], A. E. Edim[1] and S. S. Akpan[1].

ABSTRACT

Several research efforts as well as developments have chosen queueing theories as a vital tool in the measurement and analysis of performance evaluation of system components that arrives at different time intervals, waiting to be serviced by a server. In this paper, we considered the important issue of planning the queueing system to minimize system cost. This is a non-trivial task, since it involves several sets of variables: the mean arrival rate of items in the queue, the mean service rate of the server to the items in the queue, and the queueing model type to be used and the efficiency of the model in terms of performance evaluation. The task is further complicated due to the presence of system performance constraints, and the inter-dependence among the variables. Our first contribution in this paper is the formulation of this problem in terms of the variables, constraints and the optimization criterion. Our second contribution is in identifying the dependences among the variables and breaking down the problem into tractable sub-parts. In this process, we extensively used domain knowledge to strike a balance between tractability and practicality.

INTRODUCTION

A queue is a linear list of items waiting to be serviced by a server (John, 1944). The emergence of a queue is as a result of numerous items waiting to be service by a scarce resource known as the server. However, there is a direct relationship between the components of the queueing system, i.e. items waiting in a line requesting for service from the server and the server itself. However, the problem of interest here is to deal with the rate at which items in the queue are being serviced. The typical questions of interest are:

What is the time required to render service to an item in the queue?

What is the length of items in the queue?

To what extent is the server in the queue utilized?

The answers to the foregoing questions cannot be supplied easily since no simple characterization of the items requesting for service from a server by a general purpose system exists.

Often the best we can do is to determine the properties of an average system.

A better characterization is provided by determining the probability that a system of a given type (model) has a given property. The behaviour of a class of systems with respect to this property can then be represented by appropriate probability distribution. For example, it might be determined experimentally that the probability p(t) of a server rendering service to an item in the waiting queue in time t or less is approximated by the exponential function

$P(t) = 1\text{-}e^{-t/T}$

where T is an average execution time.

The probability distribution function therefore characterizes one aspect of the queueing system performance with respect to the server. Probabilistic or statistical parameters of this kind are often used in the analysis and synthesis of servers performance in the system.

As a result, there is inherent uncertainty in the behaviors of the system components. There may be even more uncertainty in the behaviour of the entire system. Although it may have been designed with a certain type of behaviour as an objective, the exact behaviour of the system must often be determined after a design has been completed, a process called performance evaluation is clearly important to the prospective user of the new system, it is equally important to the system designer.

PERFORMANCE EVALUATION

The goal of performance evaluation is to determine functions of the form $\varphi(X_1, X_2, ----, X_n) = \varphi(X)$, where X is a set of design parameters. (including items in the waiting queue) and φ is a performance measure such as service time, waiting time, or resource utilization (server utilization)(Lee, 1966).

It is generally desirable to be able to write φ as algebraic expression involving X. Such an expression is said to be an analytic-model of φ. Tractable analytic models have been developed for certain aspect of system performance evaluation, but accurate models of this kind are quite rare.

[*]Corresponding author

Manuscript received by the Editor July 10, 2006; revised manuscript accepted May 21, 2007.

[1]Department of Mathematic/Statistic & Computer Science ,University of Calabar, Calabar, Nigeria

The difficulties of purely analytical approaches to performance evaluation arise from the fact that the components of a system can interact in complex ways. Communication between system components is asynchronous, and contention for shared system resource frequently results.

In case where φ cannot be expressed in tractable form, it may be possible to compute φ(x) systematically for specific (numerical) value of x. such approaches may be termed numerical or experimental and fall into two main groups:

(a) Computer based simulation.

(b) Performance measurement on an actual system.

Queueing theory is a branch of applied probability theory concerned with processes that involves sharing of limited resources (server), the resource limitations result in waiting line or queues forming at the resources (server). The origins of queueing theory are usually traced to the analysis of congestion in a telephone system made by the Danish engineer A. K. Erlang [1878 – 1929] in Takacs (1969). A queueing system is a collection of queues that are waiting for service by a set of servers. The manner in which the queues are formed and serviced is determined by suitable probability distribution.

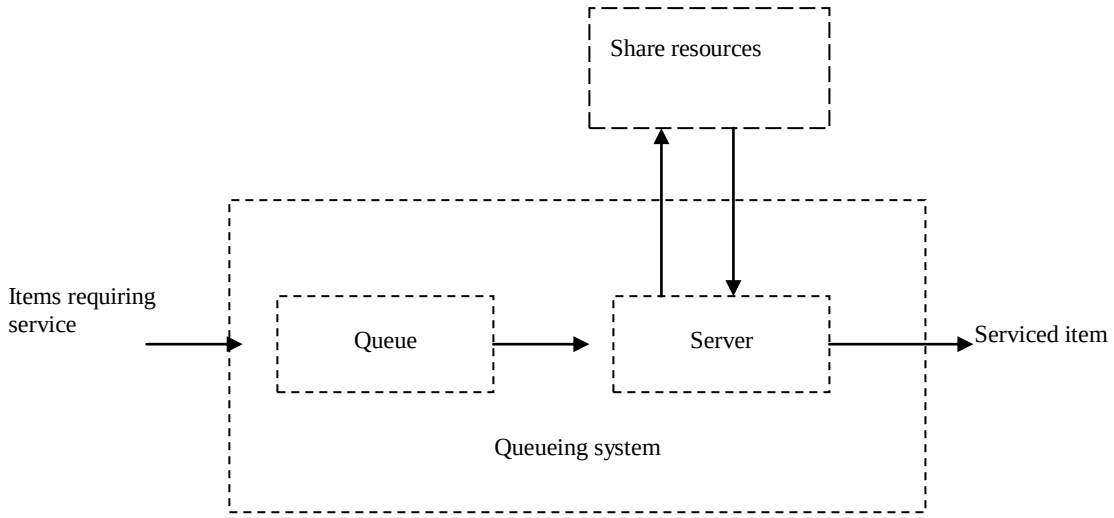

Fig. 1. A simple queueing model consisting of a single queue and single server.

Fig. 1. shows the simplest queueing system consisting of a single queue and a single server.

The appropriateness of queueing theory for performance evaluation stems from the fact that the queueing system consists of a set of limited resources, such as the server, power supply, and the efficiency of the sales person, which must be shared among competing items in the waiting line. Unfortunately, the analysis of queueing systems is extremely difficult unless rather restrictive assumptions are made.

PROBLEM FORMULATION

In this section, we articulate the problem of planning and designing of the system model in the course of our discussion, we present the optimization objective, the problem constraints, and the variables involved. Our primary goal is to minimize the waiting time to items in the queue and maximize sales of product. This is because system cost is a central consideration in the design of any system. In this paper, our research topic is as follows: from a statistics survey, it was observed that in Total filling station, at Ekpo Abasi, Calabar South,

vehicles in a queue were being service by a single server in a manner that can be approximated by the single server single queue markov processes (M/M/1) model. Arriving vehicles are in the waiting line (queue) until they are fully serviced by the server. New vehicles arrive at the system at an average rate of 10 per minute, and it is found that the server is, on the average, idle 25 percent of the time. We ask two questions:

1. What is the average time T that each vehicle spends in the queue?
2. What is the average number of vehicles N in the queue waiting to be serviced by the server?

Our optimization objective can thus be stated as:

Minimize the waiting time of each vehicle in the waiting line and maximize sales of product in the system. Obviously, when the waiting time for each vehicle in the queue is minimize, then total sale is bound to increase, this would ultimately increase the profit made from the total volume of sale of the product at a given time interval.

OVERALL SOLUTION STRATEGY

As depicted in fig. 1, there is significant inter-dependence in determining the sets of variables in the model. Given this, even the formulation of the problem in its entirety is complex. We initially tried this approach, but soon realized that the problem was better address by breaking it up into smaller parts.

We do not seek to design a single algorithm, which determines all the variables and satisfies all the constraints at one go. The sets of variables in the model suggest a natural division of the problem lines, however, due to the inter-dependence, it is still challenging to determine the order in which a set of variables are to be solved for. That is, the order in which the dependences are resolved are non-trivial.

We begin the analysis by making the following observations:

(i) The most basic queueing system is the single queue single server model depicted in Fig. 1. The parameters that define the behaviour of this system are the rate at which vehicles requiring service arrive and the rate at which vehicles are serviced.

(ii) It is assumed that vehicles on the queue are serviced on a first come first served basis. The mean arrival and service rate are denoted by λ and μ, respectively. The actual arrival and service rate varies randomly around these average values and are therefore characterized by probability distributions.

The way in which vehicles arrive at the queueing system (the arrival process) is often modeled by a Poisson probability distribution function, which has the form

$$Pp(n,t) = (\lambda t)^{n}_{e}{}^{-\lambda t} - \quad \text{- eqn.} \qquad (1)$$

where Pp (n , t) is the probability of exactly n vehicles arriving in time period of length t. If vehicles are not removed from the queue, Pp (n, t) represents the probability that the queue length – increases by n in time t.(Morse, 1958).

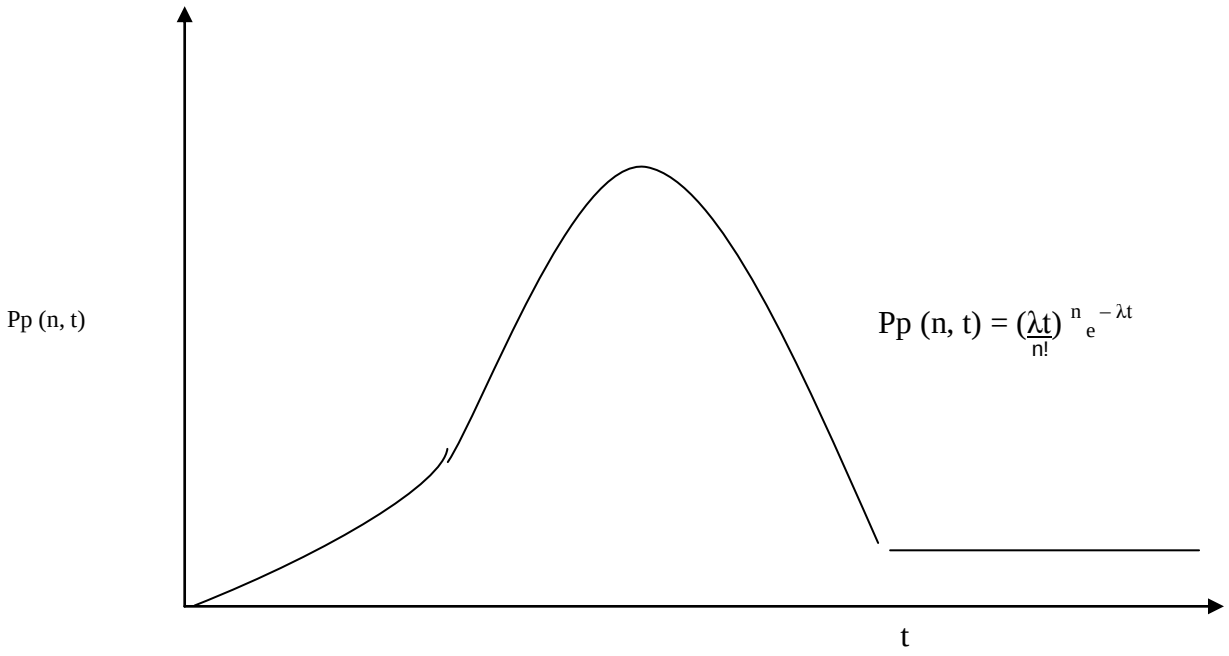

Fig. 2. A plot of Pp (n, t) against t

Fig. 2. shows a plot of Pp (n, t) as a function of t for fixed values of n and λ. This curve, which rises quickly to a peak and then decreases asymptotically to zero, typifies the Poisson distribution. Many physical arrival processes, including the arrival of calls at a telephone exchange, and the arrival of vehicles at Total filling station, along Ekpo Abasi, Street in Calabar Metropolis, can be quite accurately modeled by a Poisson distribution.

Arrival process

Another important characteristic of an arrival process is the distribution of the time period between two consecutive arriving items (vehicles). The inter - arrival time distribution p_1 (t) is defined to be the probability that at least one item (vehicle) arrives during the period of length t.

Clearly, for a Poisson arrival process

$$p_1 (t) = 1 - Pp\ (o, t) \qquad (2)$$

Therefore, on setting n = o in equation (1) and substituting it into equation (2) we obtain

$$p_1 (t) = 1 - e^{-\lambda t} \qquad (3)$$

This is the negative exponential distribution, which has the form shown in Fig. 3.

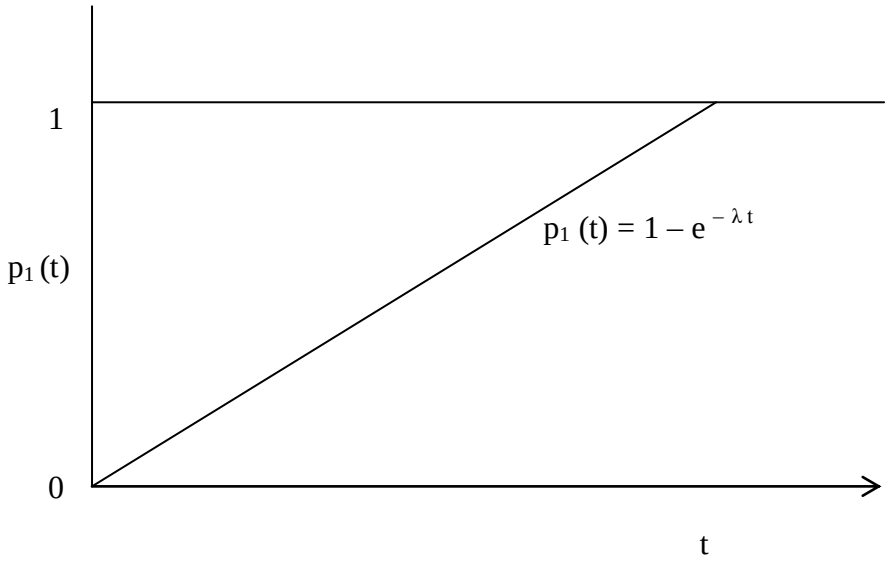

Fig. 3. Exponential distribution for fixed λ

Hence the inter-arrival times of a Poisson arrival process are characterized by the exponential distribution of equation (3). The probability density function corresponding to equation (3) is $\lambda e^{-\lambda t}$, while its mean value is $1/\lambda$. Exponential distributions are particularly simple and convenient to use in analytic models. It is therefore common to model the behaviour of the server (the service process) by an exponential distribution also.

Let $P_s(t)$ be the probability that the service required by an item (vehicle) is completed in time t or less after its removal from the queue. Then the service process is often characterized by the expression $P_s(t) = 1 - e^{\mu t}$. The performance of a queuing system can be measured by the following parameters:

(i) The average number of items (vehicle) waiting in the queue, including the items (vehicles) waiting for service and those actually being served. The parameter is called the mean queue length and is denoted by LQ.

(ii) The average time that arriving items spend in the queue system, both waiting for service and being served. This is called the mean waiting time and is denoted by t_Q. which is sometimes called the system delay or response time.

We now consider the problem of calculating L_Q and t_Q for a single queue, single server system in which the arrival distribution is Poisson (or equivalently, the inter arrival time distribution is expenential) and the service time distribution is exponential. This type of queueing system is called M/M/1 model, where the two Ms denote that the arrivals and service processes are Markov processes (essentially the same as Poisson processes) while the 1 denotes the number of servers.

The state of the queue can be specified by $P_Q(n, t)$, which is the probability that at time t there are exactly n items (vehicles) in the queuing system either waiting for service or being served(Stidham, 1974). When the system has been in operation for some time, the system can be expected to reach a state of equilibrium in which $P_Q(n, t)$ can be assumed to be independent of t, so we write $P_Q(n, t) - P_Q(n)$ (Stidham, 1974). It can be shown that under these conditions

$$P_Q(n) = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) \quad (4)$$

Provided $\lambda < \mu$. If $\lambda > \mu$, then the queue grows indefinitely. Equation (4) is called the steady state or balance equation for the queueing system.

The quantity $\rho = \lambda / \mu$ is the mean utilization of the server and is called the traffic intensity. Equation (4) can be rewritten in terms of ρ thus:

$$P_Q(n) = \rho^n (1 - \rho) \quad (5)$$

The mean queue length L_Q can be immediately expressed in terms of $P_Q(n)$ as follows:

$$L_Q = \Sigma n P_Q(n)$$
$$n \geq 1$$

Substituting from eqn (5) and starting the summation at n = 0, we obtain

$$L_Q = (1 - \rho) \Sigma \rho^n$$
$$n \geq 0$$
$$= \rho (1 - \rho) \frac{d}{d\rho} \Sigma \rho^n$$
$$n \geq 0$$

The summation in this last expression is an infinite geometric progression equal to $1/(1 - \rho)$

$$L_Q = \rho(1 - \rho) \frac{d1}{d\rho} \frac{}{1 - \rho}$$

Using the formula

$$\frac{d\,1}{dx\mu} = \frac{-1}{\mu^2} \frac{d\mu}{dx}$$

We obtain

$$\frac{d1}{d\rho} = \frac{1}{1 - \rho(1 - \rho)^2}$$

$$L_Q = \frac{\rho}{1 - \rho} \quad (6)$$

This defines the mean queue length. Finally we turn to the parameter t_Q, which is the mean time items spend in the queuing system. t_Q and L_Q may be related intuitively as follows: An average item x passing through the queue system should encounter the same number of waiting item L_Q when it enters as it leaves behind when it depart from the system. The number left behind is λt_Q, which is the number of items that enter the system at rate λ during the period t_Q when x is present.

Hence we could conclude that $L_Q = \lambda t_Q$ that is

$$t_Q = \frac{L_Q}{\lambda} \quad (7)$$

Equation (7) is known as the little's equation.

It is valid for all queuing systems, not just the M/M/1 model use here, combining equations (6) and (7) yields the desire expression for t_Q:

$$t_Q = \frac{1}{\mu - \lambda} \quad (8)$$

The quantities L_Q and t_Q defined by equation (6) and equation(8) refer to items (vehicles) that are either waiting for access to the server or are actually being served.

The mean number of items (vehicles) waiting in the queue excluding those being served is denoted by L_w, while t_w denote the mean time spent waiting in the queue excluding service time. (The subscript w stands for waiting).

The mean utilization of the server in an M/M/1 queueing system, I.e. the mean number of items (vehicles) being serviced, is

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

Hence subtracting this from L_Q yields L_w thus: $L_w = L_Q - \rho$ (9)

Using equation (6) to replace L_Q, we obtain

$$L_w = \frac{\rho^2}{1 - \rho} = \frac{\lambda^2}{\mu(\mu - \lambda)} \quad (10)$$

Similarly,

$$t_w = t_Q - \frac{1}{\mu}$$

where $1/\mu$ is the mean time it takes to service an item (vehicle).Substituting for t_Q from equation (8) yields

$$t_w = \frac{\lambda}{\mu(\mu - \lambda)} \quad (11)$$

Comparing equation (10) and equation (11) we see that $t_w = L_w/\lambda$, so that littlies equation holds for both the Q and the W subscripts.

To answer our questions, we assumed that steady state condition prevails from which it follows that T is t_Q, and N is L_w. Since the system is busy 75 percent of the time, $\rho = 0.75$, but $\lambda = 10$ vehicles / min; hence the service rate $\mu = \lambda/\rho = 10 / 0.75$

$$= \frac{10}{75} \times \frac{100}{1} = 40/3 \text{ vehicles / min}$$

Substituting into equation (8)

We get $T = t_Q = \frac{1}{\mu-\lambda} = \frac{1}{\frac{40}{3} - \frac{10}{1}}$

$$= \frac{1}{\frac{10}{3}} = \frac{3}{10} = 0.3\text{mins}$$

$$\therefore T_Q = 0.3\text{mins}$$

$$N = L_Q = \lambda t_Q = 10 \times 0.3 = 3 \text{ vehicles}$$

$$L_W = L_Q - \rho = 3 - 0.75$$

$$= 2.25 \text{ vehicles}$$

$$= 2 \text{ vehicles approximately}$$

From the evaluation of the above analysis, the mean time $T_w = 0.3$ mins. and L_w (number of vehicles waiting in the queue) is 2.

CONCLUSION

Since our $\lambda<\mu$, then our queue does not grow indefinitely. T_Q = 0.3min. Thus the time it takes for a vehicle to wait in the queue is minimal.

The length of the queue is 2 vehicles which implies that the queue length was not long at all. Hence, from the analysis to our problem, our solution is optimal and practical to a real life business situation.

RELATED WORK

The related topics to this research work are: scheduling resource use in computer systems. Jobs queueing up in the computer system waiting to be serviced by the CPU. Multi servers multi queues system, and single queue multi server systems.

ACKNOWLEDGEMENT

This first author would like to acknowledge Prof. Zsolt Lipscey, of the Department of Mathematics/Statistics and Computer Science, University of Calabar, for the fruitful brainstorming session, and Dr. Omidiora, E. O. of the Department of Computer Science and Engineering, Ladoke Akintola University of Technology, Ogbomoso, Nigeria, for the help in reviewing the paper.

REFERENCES

John, P.H.(1944).*Computer architecture and organization*. Mc Graw Hill, New York.

Lee, A. M.(1966) .Applied Queueing theory, St Martin's Press, New York,

Morse, P. M.(1958).Queues, Inventories, and Maintenance, Wiley, New York.

Stidham, S.((1974). A Last Word on L = λw", OR 22, (2): 417 – 421

Takacs, L.((1969).Erlang's Formula, AM. Math. Stat, 40: 71 – 78.