Bul. Ştiinţ. Univ. Baia Mare, Ser. B,
Matematică -Informatică, Vol. XIV(1998), Nr. 1, 81-88

Dedicated to Professor Iulian Coroian on his 60 th anniversary

APPLICATION OF THE QUEUEING THEORY TO DISCRETE EVENT SIMULATION

Alexa L. PETREAN

Abstract. Our aim in this article is to present the basics concepts that appear in discrete system modelling using the waiting lines and to show, both analitically and practically, the corresponding solving methods for the arising problems.

The use of the simulation approach represents one of the most modern and powerful technics of solving complex problems that appear in almost every fields of human activity. Simulation models of real-world systems can be classified as discrete change or continuous change models. In discrete simulation we suppose that the state variables of the system vary discretely at points in time, while continuous simulation implies that state indicators change in a continuous manner in time. For both models, the change of the system states means that an event occurs.

In this paper we will concentrate our attention on the discrete modeling of the waiting lines (queues), using for this purpose abstract concepts from the queueing theory. Waiting lines appear in almost every system that is modeled because most of them use some limited resource, whether it is the number of servers in a filling station or the number of available I/O channels on a general-purpose computing system.

The major problem of interest which appears in the queueing theory is to solve in both theoretical and practical way the queueing models, having imposed some financial and material restrictions. Queues arise because of the competition for the limited resources that exist in a system and can be distinguished by the manner in which customers are served.

In every waiting problems there are some common elements as follow: the customers (input units) which wait for the necessary service if that is not immediately available, the serving stations that work through one or more disposable channels, an input stream and an output stream of units. With this agreement, a simple queueing system can be illustrated like this:

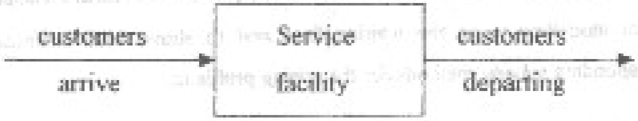

Fig.1 A queueing system

For any type of queueing system a number of common characteristics that need to be discussed can be distinguished [1]:

- the *arrival pattern*, concerns with the distribution of the input stream. The customers can arrive singly or in batches and in certain circumstances they can leave without receiving the service;
- the *service process*, reffers to the distribution of the service time requested. The input units can be served singly or grouped and the level of service can change or remains constant as the queue forms;
- the *number of parallel servers*. In most systems this is a finite number. The maximum capacity of a system has an important influence on the operation of the system;
- the *queue discipline* concerns itself with the discipline by which customers are selected from the waiting line for service. Queueing disciplines include: FIFO, LIFO, random (RAND) and priority (PRI) techniques.

In the next paragraphs of this article, we'll use Kendall's convention to describe a queueing system, which is based upon previous presented characteristics [4]. Thus, a queueing system is described by a series of symbols separated by slashes of form: A/B/C/D/E. Here A represent the interarrival time distribution, B the service time distribution, C the number of parallel servers, D the system capacity and E the queueing discipline. Corresponding to the first classification term the following important categories of interarrival distributions can be considered: M (exponential), D (deterministic), E_k (Erlang type k) and G (general). The same letters are used to denote similar service time distribution.

For each queueing model there are a number of items of interest. The most important of which we will consider are: the queue length, the time in the server station, the idle and the busy time of the server. The next paragraphs contain a briefly description of some queueing models with a large area of application to real-world systems, more precisely the M/M/1/∞/FIFO model, the M/M/k/∞/FIFO model and priority queueing models.

The M/M/1/∞/FIFO model

In this case there is a single-server system, whose interarrival and service times are exponentially distributed with parameters $1/\lambda$ and $1/\mu$ respectively. There is no restriction on system capacity and the queue follows a first-in, first-out discipline. We know that if the interarrival distribution is exponential, then the arrival process is Poisson. It is also known that a Poisson stream has the next important features:

- the arrivals probabilities of an arbitrary number of units in nonoverlapping time intervals do not depend on each other;
- the probability of two or more arrivals in Δt is $O(\Delta t)$ and can be neglected related to the probability of which one arrival appears;

- the probability that an arbitrary number of units arrives in the system in a time interval doesn`t depend of the interval position on the time axis but only on the interval length.

A very important value of interest in the analysis of any queueing system is the number of customers in the system. If we denote with S_n the state of the system when there are n customers present ($n \geq 0$) and with $P_n(t)$ the probability of state S_n at some time t, then the system remains in state S_n at time $t+\Delta t$ if and only if one of the following mutually exclusive events occur:

i) the system was in state S_n at time t and no arrivals or departures occur during the period $(t, t+\Delta t)$;

ii) the system was in state S_{n+1} at time t and one departure but no arrivals occur during the interval $(t, t+\Delta t)$;

iii) there was n units in system at time t and no arrivals or departures occur during the interval $(t, t+\Delta t)$.

From the above statements the next equation results [3]:

$$P_n(t+\Delta t)= \lambda P_{n-1}(t)\Delta t+(1-(\lambda+\mu)+O(\Delta t))P_n(t)+\mu P_{n+1}(t)\Delta t,\ n=1,2,\ldots \text{ or:}$$

$$\frac{P_n(t+\Delta t)+P_n(t)}{\Delta t} = \lambda P_{n-1}(t) - (\lambda + \mu)P_n(t) + \mu P_{n+1}(t),\ n = 1,2,\ldots$$

Taking the limit at both sides as $\Delta t \to 0$ gives:

$$P_n'(t)=\lambda P_{n-1}(t)-(\lambda+\mu)P_n(t)+\mu P_{n+1}(t),\ n=1,2\ldots \qquad (1)$$

For n=0, utilizing the similar procedure gives:

$$P'_0(t)=\lambda P_0(t)+\mu P_1(t) \qquad (2)$$

From (1), (2) a system of equations results, whose solution gives the probability to have n units in the system and that is easier to solve once the system is in a steady state ($P_n(t)=P_n$=constant). Under the assumption of steady state the system of equations becomes a set of simple difference equations of the form:

$$\begin{cases} P_1 = \dfrac{\lambda}{\mu} P_0 \\ P_{n+1} = \dfrac{\lambda+\mu}{\mu} P_n - \dfrac{\lambda}{\mu} P_{n-1} \end{cases}$$

which is easy to solve, using, for example, an iterative technique. The solution is given by: $P_n=\rho^n(1-\rho)$, $n=0,1,2,\ldots$where $\rho=\lambda/\mu$ represents the traffic intensity. Using this formula, the expressions of various measures of interest can be obtained. We have, for example:

- the expected number in the system $L=\rho/(1-\rho)$;
- the expected number in the queue $L_q=\rho^2/(1-\rho)$;
- the expected time in the system $W=1/(\mu(1-\rho))$;
- the expected time in the queue $W_q=\rho/(\mu(1-\rho))$.

The M/M/C/∞/FIFO model

In this case, there is a finite number $1<C<\infty$ of servers in the system, each with an independently and identically distributed exponential service time distribution with rate μ. The input units that arrive follow a Poisson process. The probability that $c\leq C$ customers remain in the servers in $(t, t+\Delta t)$ is now $1-c\mu\Delta t+O(\Delta t)$ and the probability of depart is $c\mu\Delta t+O(\Delta t)$.

According to a similar procedure to that used for the analysis of the M/M/1/∞/FIFO model, the following expression for the probability that n customers are in the system results [3]:

$$P_j = \begin{cases} \dfrac{\rho^j}{j!} P_0,\ 1 \leq j \leq C \\ \dfrac{\rho^j}{C!C^{j-C}} P_0,\ j > C \end{cases}$$

The value of P_0 yields from the condition: $\sum_{j=0}^{\infty} P_j = 1$ or:

$$P_0\left[\sum_{j=0}^{C-1}\frac{\rho^j}{j!}+\frac{1}{C!}\sum_{j=C}^{\infty}\frac{\rho^j}{C^{j-C}}\right]=1 \Leftrightarrow P_0=\frac{1}{\sum_{j=0}^{C-1}\frac{\rho^j}{j!}+\frac{C\rho^C}{C!(C-\rho)}}$$

From this relation, the formulas of the characteristics of the system can be obtained, but they have a more complicated expression in this case [2]. As a last remark, it is important to note that the requirement for the M/M/C/∞/FIFO model to reach steady state is λ/Cμ<1, rather than λ/μ<1 as with the previous model.

The priority queueing model

The particular feature of this model is that the discipline used to select a customer for the service is based on a priority system. There are at least two important reasons to apply priority disciplines. One refers to the minimization of the average cost of the system. That is possible, for example, if the high-cost units in the queue are served first. The second reason is to reduce the average number of customers in the system. Thus, if the service time is shorter for certain customers than for others and the priority scheme is based on the principle "begin with the unit that requires the least service", then it is possible to reduce the average number of customers in the system.

There is a set of various techniques used to assign priorities to the customers of a system, as follows:

- *shortest service first*, the highest priority is given to the unit that requires the least amount of service, supposing that the length of the time service is known;
- *round robin*, the server allocates for each customer a quantum of service time. If the quantum is not sufficient to finish the operation, the service is interrupted and the customer rejoins the queue in a cyclic manner;
- *willingness to pay*, when the customers are allowed to buy a high priority. In this case there is a number of levels of priority, each with a

corresponding rate and a customer is charged according to the level of priority desired.

The priority queueing disciplines are of one of the next two types: non-preemptive disciplines or preemptive disciplines. The non-preemptive technique specifies that a started service of a given customer cannot be interrupted until it is finished. For the preemptive scheme, if a new unit which enters the system has a higher priority than the one being served, the service is interrupted for the current unit and the higher-priority customer gains the control of the server. The interrupted customer rejoins the queue service and the portion of service that it received can be memorized or not. In the first case we say that a preemptive repeat discipline exists while the other one implies a preemptive resume discipline in the system.

Analyzing the queueing models presented above at first sight, the following statement can be invoked: as the queueing systems become more complex, the mathematics involved becomes nearly intractable [2]. Because of the complexity of the formulas that appear in the analytical treatments of the systems using queueing models, it is more convenient to apply the simulation techniques. Moreover, even if the system can be analyzed in an analytical mode, it is simpler to simulate it and to use the theoretical results to validate the simulation.

Below we give a brief description of the most important procedures that might be implemented in a general-purpose computing program which simulates a queueing system. Thus, four mainly routines can be distinguished here: ARRIVAL, SERVE, ADD and REMOVE.

The ARRIVAL routine simulates the arrival of an input unit to a waiting line and returns the time of the next arrival. The arrival process follows various distribution probabilities (for example a Poisson distribution).

The SERVE subroutine assigns required service time to the arrivals to a queueing system and must returns the service time required by the next arriving customer. In this case also the service time distribution follows a chosen law of probability.

The scope of the ADD subroutine is to enter the arriving customers into a finite length waiting line. Customers attempting to enter a full queue should be turned away. The implementation of this procedure strongly depends on the queueing discipline used in simulation (FIFO, LIFO, PRI, etc.)

Finally, the REMOVE routine deletes from the system the customer that was served. A particular attention should be given to the operation of removing a customer from an empty queue.

REFERENCES

1. Pooch, Udo W., and Wall James A., "*Discrete Event Simulation - A practical Approach*", CRC Press, 1993
2. Gross, D., and Harris, C., "*Fundamentals of Queueing Theory*", New York: John Wiley and Sons, 1974
3. Mihoc, Gh., Bergthaller, C., and Urseanu, V., "*Procese stochastice. Elemente de teorie şi aplicaţii*", Bucureşti, Ed. ştiinţifică şi enciclopedică, 1978
4. Kendall, D. G., "*Stochastic Processes Occuring in the Theory of Queues and Their Analysis by the Method of Imbedded Markov Chains*", Ann. Math. Statistics, 24 (1953): 338 - 354

Received 10.06.1998

Department of Electrotechnics
Faculty of Engineering
North University of Baia Mare
Dr. Victor Babeş st., no. 62/A
RO-4800 Baia Mare
ROMANIA
E-mail: axel@univer.ubm.ro