REVUE FRANÇAISE D'AUTOMATIQUE, D'INFORMATIQUE ET DE RECHERCHE OPÉRATIONNELLE. RECHERCHE OPÉRATIONNELLE

T. M. O'DONOVAN

Direct solutions of M/G/1 priority queueing models

Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle, tome 10, n° V1 (1976), p. 107-111.

<http://www.numdam.org/item?id=RO_1976__10_1_107_0>

R.A.I.R.O. Recherche opérationnelle
(vol. 10, n° 2, février 1976, p. 107 à 111)

DIRECT SOLUTIONS OF M/G/1 PRIORITY QUEUEING MODELS (*)

by T. M. O'DONOVAN (1)

Abstract. — *This paper presents a general method for deriving expected conditional response times in priority queueing models. The method consists of applying Kleinrock's conservation law to subsystems of jobs with priority over all other jobs. The method is illustrated for the following queue disciplines: preemptive resume shortest processing time, non-preemptive resume shortest processing time and shortest remaining processing time.*

Three well-known queueing models are considered in which priority is assigned to jobs on the basis of their processing times. It is shown that the average waiting times in these models are easily evaluated by applying a conservation law to a subsystem of jobs.

Mathematical models of priority queues have been widely studied (*see* Jaiswal [3]). This paper is concerned with priority queues in which priority is assigned to jobs on the basis of their processing time requirements. Of these systems, the Non-preemptive Shortest Processing Time system is most widely used. In this system, jobs are served to completion. When a job is to be selected from among those waiting, the one with the shortest processing time is chosen. In the Preemptive Resume Shortest Processing Time system, an arriving job will preempt the job in service if and only if the processing time of the new arrival is less than the total processing time of the job then in service. Partially completed jobs can be removed from the processor and returned at a later time without waste of time or work already done. In the Shortest Remaining Processing Time system, an arriving job will preempt the job in service if and only if the processing time of the new arrival is less than the remaining processing time of the job then in service. When a job is to be selected from among those waiting, the one with the lowest remaining processing time is selected.

The expected conditional waiting times in M/G/1 models under these queue disciplines were derived by Phipps [6], Cohen [1] and Conway, Maxwell and Miller [2], respectively, by first evaluating this characteristic in models

(*) Reçu avril 1975.

(1) Department of Statistics, University College, Cork, Ireland.

with a finite number of priority levels and then letting the number of levels become infinite. In the case of the last system, Schrage and Miller [7] have given a direct derivation of this characteristic using a complicated busy period argument. Here it is shown that for each of these models, this characteristic is easily obtained directly by applying Kleinrock's Conservation Law [4] to a subsystem of jobs.

We consider M/G/1 queueing systems in which jobs arrive at rate λ and the processing times are independently sampled from a distribution having distribution function $F(.)$. At each epoch, a job in the system is either waiting for service, being served, or (under queue disciplines which permit interrupting a job in service before it is completed) in limbo (*see* Wolff [8]). The waiting time of a job is the time from the epoch the job arrives until the epoch its service begins. Let $W(t)$ be the expected waiting time of a job whose processing time requirement is t units. Let $1/\mu$ and m_2 be the first and second moments of the processing time distribution.

Define,

$$\rho = \lambda \frac{1}{\mu}, \qquad V = \frac{1}{2}\lambda m_2,$$

$$\lambda(t) = \lambda F(t), \qquad m(t) = \int_0^t x \frac{dF(x)}{F(t)},$$

$$\rho(t) = \lambda(t)\, m(t).$$

A CONSERVATION LAW

Kleinrock [4] has proved a Conservation Law for queueing systems subject to the following restrictions:

1. All jobs remain in the system until completely serviced.
2. The single service facility is always busy if there are any jobs in the system.
3. Preemption, if it occurs, is of the preemptive-resume type.

Consider the load on such a system at a given time point, i. e. the total processing time yet to be allocated to all the jobs in the system. It is obvious that this load is independent of queue discipline. Thus L, the expected load on the system at a random time point, is also independent of queue discipline. The expected load on the system at a random time point due to the job, if any, in service is well known to be independent of queue discipline and to have the value:

$$V = \frac{1}{2}\lambda m_2, \tag{1}$$

(*see* Wolff [8]). This holds not only for Poisson arrivals but also for general independant arrivals. Thus the expected load on the system due to jobs

waiting or in limbo is also independent of queue discipline. We will evaluate ti in a system with a First Come First Served queue discipline. If W is the expected waiting time in such a system, it follows from Little's relation [5] that the expected number of jobs waiting for service at a random time point is λW and so the expected load on the system due to these jobs is $\lambda W 1/\mu$. Assuming that the arrivals of jobs to the waiting line form a Poisson process, we have $W = L$.

Thus

$$L = \rho L + V. \tag{2}$$

This is Kleinrock's Conservation Law.

In this paper, we consider M/G/1 queueing systems under the following queue disciplines:

1. Preemptive Resume Shortest Processing Time.
2. Non-preemptive Shortest Processing Time.
3. Shortest Remaining Processing Time.

Let $W_i(t)$ $(1 \leqq i \leqq 3)$ be the value of the expected waiting time $W(t)$ under the corresponding queue discipline. We evaluate $W_i(t)$ as follows. In each system, we define a different subsystem of jobs S_i and let L_i be the expected load on the subsystem. It is immediately obvious that in each system $W_i(t)$ has two components:

a) The expected load L_i.

b) The delay caused by subsequent arrivals while this load is being cleared, whose processing times are less than t. Such jobs arrive in a Poisson process with rate $\lambda(t)$ and their expected processing time is $m(t)$. By delay cycle analysis [2], it follows that:

$$W_i(t) = \frac{L_i}{1-\rho(t)}. \tag{3}$$

In each system, L_i is evaluated by applying the Conservation Law (2) to the subsystem of jobs S_i. This is possible because the jobs in the subsystem S_i have priority over all other jobs and so condition (ii) for the Conservation Law is satisfied. Let L_i^{W+L} be the expected load on the subsystem S_i at a random time point due to jobs waiting or in limbo and let L_i^s be the corresponding expected load due to jobs in service. Then:

$$L_i = L_i^{W+L} + L_i^s. \tag{4}$$

THE PREEMPTIVE RESUME SHORTEST PROCESSING TIME SYSTEM

Let S_1 be the subsystem of jobs whose processing times are at most t. The arrival rate of such jobs is $\lambda F(t)$ and the second moment of their processing time distribution is $\int_0^t x^2 (dF(x)/F(t))$. Thus by (1), we have:

$$L_1^s = \frac{1}{2}\{\lambda F(t)\}\left\{\int_0^t x^2 \frac{dF(x)}{F(t)}\right\}.$$

By the argument used in the derivation of the Conservation Law it follows that:

$$L_1^{W+L} = \rho(t)\, L_1.$$

Thus from (4),

$$L_1 = \rho(t)\, L_1 + \frac{1}{2}\lambda \int_0^t x^2\, dF(x)$$

and so from (3),

$$W_1(t) = \frac{(1/2)\lambda \int_0^t x^2\, dF(x)}{(1-\rho(t))^2}$$

(*see* Cohen [1]).

THE NON-PREEMPTIVE SHORTEST PROCESSING TIME SYSTEM

Let S_2 be the subsystem of jobs whose processing times are at most t plus the job, if any, in service. Jobs whose processing times are at most t enter this subsystem on arrival and join the waiting line, if any. A job whose processing time exceeds t can only enter the subsystem if there are no jobs whose processing times are at most t in the subsystem. Such a job begins service as soon as it enters the subsystem and thus never joins the waiting line. From (1), it follows that:

$$L_2^s = V.$$

As before, the contribution to L_2^{W+L} of jobs whose processing times are at most t is $\rho(t)\, L_2$. Jobs whose processing times exceed t, never join the waiting line and so their contribution to L_2^{W+L} is zero. Thus from (4),

$$L_2 = \rho(t)\, L_2 + V$$

and so from (3):

$$W_2(t) = \frac{V}{(1-\rho(t))^2}$$

(*see* Phipps [6] and Cohen [1]).

THE SHORTEST REMAINING PROCESSING TIME SYSTEM

Let S_3 be the subsystem of jobs whose remaining processing times are at most t. Jobs whose processing times are at most t enter this subsystem on arrival and join the waiting line, if any. A job whose processing time exceeds t, will only begin to be served when there are no jobs in the subsystem whose remaining processing times are at most t. When its remaining processing time equals t, it then enters the subsystem and continues in service unless preempted by a subsequent arrival. Thus such a job never joins the waiting line. Since all jobs in the original system eventually join the subsystem and the time spent in service by a job in S_3 is distributed as a processing time truncated at t, we have from (1) that:

$$L_3^s = \frac{1}{2}\lambda\left\{\int_0^t x^2\,dF(x)+t^2(1-F(t))\right\}.$$

As before, the contribution to L_3^{W+L} of jobs whose processing times are at most t is $\rho(t)\,L_3$.

As shown above, jobs whose processing times exceed t never join the waiting line. Since L_3^{W+L} is independent of queue discipline and when jobs in S_3 are served First Come First Served, jobs whose processing times exceed t will never enter limbo while in S_3, the contribution of such jobs to L_3^{W+L} is zero.

Thus from (4),

$$L_3 = \rho(t)\,L_3+\frac{1}{2}\lambda\left\{\int_0^t x^2\,dF(x)+t^2(1-F(t))\right\}$$

and so from (3),

$$W_3(t) = \frac{(1/2)\lambda\left\{\int_0^t x^2\,dF(x)+t^2(1-F(t))\right\}}{(1-\rho(t))^2}.$$

(*see* Schrage and Miller [7]).

REFERENCES

1. J. COHEN, *The Single Server Queue*, North-Holland Publishing Company, 1969.
2. R. W. CONWAY, W. L. MAXWELL and L. W. MILLER, *Theory of Scheduling*, Addison-Wesley, 1967.
3. N. K. JAISWAL, *Priority Queues*, Academic Press, 1968.
4. L. KLEINROCK, *A Conservation Law for a Wide Class of Queueing Disciplines*, Nav. Res. Log. Quart., Vol. 12, 1965, pp. 181-192.
5. J. D. C. LITTLE, *A Proof for the Queueing Formula $L = \lambda W$*, Opns. Res., Vol. 9, 1961, pp. 383-387.
6. T. PHIPPS, *Machine Repair as a Waiting Line Problem*, Opns. Res., Vol. 4, 1956, pp. 76-86.
7. L. SCHRAGE and L. W. MILLER, *The Queue* M/G/1 *with the Shortest Remaining Processing Time Discipline*, Opns. Res., Vol. 14, 1966, pp. 670-684.
8. R. W. WOLFF, *Work-Conserving Priorities*, J. Appl. Prob., Vol. 7, 1970, pp. 327-337.

n° février 1976.