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

GEORGE P. COSMETATOS
GREGORY P. PRASTACOS

An approximate analysis of the D/M/1 queue with deterministic customer impatience

Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle, tome 19, n° 2 (1985), p. 133-142.

<http://www.numdam.org/item?id=RO_1985__19_2_133_0>

© AFCET, 1985, tous droits réservés.

L'accès aux archives de la revue « Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle » implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/legal.php). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

R.A.I.R.O. Recherche opérationnelle/Operations Research
(vol. 19, n° 2, mai 1985, p. 133 à 142)

AN APPROXIMATE ANALYSIS OF THE D/M/1 QUEUE WITH DETERMINISTIC CUSTOMER IMPATIENCE

by George P. COSMETATOS (1)
and Gregory P. PRASTACOS (2)

Abstract. — *This paper analyzes the single server queue with regular arrivals, negative exponential service times, and deterministic customer impatience. Such queueing systems appear frequenbly when modeling inventory systems for perishable products, or data communication systems. In addition they may be regarded as realistic alternatives to the standard D/M/1 queue operating in heavy traffic conditions.*

Approximations for measures that describe the behavior of the system under normal or heavy-traffic conditions are derived and tested. The results indicate that the approximations developed are very accurate.

Keywords: Queue; Perishable Inventory; Heavy traffic.

Résumé. — *Cet article analyse la file d'attente à une station de service avec arrivées régulières, temps de service exponentiels négatifs, et impatience déterministe du client. De tels systèmes de file d'attente apparaissent dans la modélisation des systèmes de gestion des stocks de produits périssables ou des systèmes de communication de l'information; de plus, ils peuvent être considérés comme des alternatives réalistes au système standard D/M/1 dans des conditions de trafic lourd.*

Des approximations des mesures décrivant le comportement du système dans des conditions de trafic normal ou lourd sont dérivées et testées. Les résultats indiquent que les approximations développées sont très précises.

Mots clés : Files d'attente; Produits périssables; Trafic lourd.

1. INTRODUCTION

In this paper we examine the class of $D/M/1$; m queueing systems, which are characterized by single deterministic arrivals, exponential service times, one server, FIFO queue discipline, and customer impatience. In these systems, a customer arriving for service can stay in the system for at most m

(*) Received in January 1983.
(1) Department of Management Science, Imperial College of Science and Technology, London, U.K.
(2) The Athens School of Economics and Business Science, Patission 76, Athens 104, Greece.

R.A.I.R.O. Recherche opérationnelle/Operations Research, 0399-0559/85/03 133 10/$ 3.00
© AFCET-Gauthier-Villars

periods; if his service is not completed by that time, then the customer leaves the system (or, is rejected by the system) never to return.

Such queueing systems appear frequently in inventory theory, or in data communications. As an example, a perishable inventory system for a product with finite lifetime of m periods can be modeled as a queueing system where the customer (item) can stay in the system (on the shelf) for at most m periods; if not used by then, it has to be descarded (see, e. g. Brodheim et al. [2], Cosmetatos and Prastacos [10] and Nahmias [11]). Similarly, in data communications a buffer receiving, storing, and issuing information can be modeled as a queueing system where all requests have to be satisfied in a certain sequence, and within the time period specified by the user requirements (see, e. g., Doll [12]).

In addition, this system represents a realistic alternative to the standard $D/M/1; \infty$ system operating under conditions of heavy traffic. Under these conditions, the « no-limit » system would lead to infinitely long waiting times, unless some form of customer impatience or discouragement was imposed.

Two measures of practical interest in the study of such a system are:

(*a*) The rejection probability, or probability of a customer reneging (i. e., departing prior to service completion).

This is a function of the mean service time $\rho = 1/\lambda$, and of the total time m allowed in the system, and we denote it, therefore, by $W(\rho, m)$. Since customers arrive singly and deterministically, $W(\rho, m)$ is also the average number of reneging customers per time period. It also follows that the average number of customers whose service is completed per time period is $1 - W(\rho, m)$.

(*b*) The average server utilization, denoted by $U(\rho, m)$, and defined as the proportion of time over which the server is busy with customers who will eventually have their service completed. Since the proportion of customers who have their service completed is $1 - W(\rho, m)$, it follows that

$$U(\rho, m) = [1 - W(\rho, m)] . (1/\lambda) = \rho [1 - W(\rho, m)] \tag{1}$$

We are interested in evaluating analytically the performance of the system in terms of the above measures.

In the queueing theory literature significant work has been done for systems with finite waiting time. Gnedenko and Kovalenko [6] studied an $M/M/1; m$ system (exponential inter-arrival and service times, single server, and finite time in the system); they derived a simple expression for the rejection probability $W(\rho, m)$, which in the notation given above reads:

$$W(\rho, m) = (1 - \rho) / [\exp(m(1 - \rho)/\rho) - \rho] \tag{2}$$

Cohen [3], Loris-Teghem [7], and more recently Gavish and Schweitzer [5]

considered generalizations of the $M/M/1$; m system by relaxing the assumptions of exponential inter-arrival or service time distributions. However, the results pertaining to the $M/M/1$; m queue do not meet our objectives, as outlined above.

A similar system that has been examined in the literature is the one where the arriving customer reneges (is rejected) if his *queueing* time reaches a fixed length, say m_q (i. e., the customer remains in the system once he starts service, independent of the service length). Considerable work exists on systems of the $GI/G/1$; m_q type (see, for example, Stanford [8]. Of particular interest for our purposes is the analysis of the $GI/M/1$; m_q queue by Finch [4], and the results that he obtained for the case where interarrival times have an Erlang distribution with parameter k. We will refer to his work in more detail in the next sections.

The paper is organized as follows: In section 2 we present a Markov Chain formulation of the $D/M/1$; m system that leads to exact calculation of the performance measures. We indicate that this formulation is not practical if m is large, since numerical computations become complex. In section 3 we derive an approximation of the system performance; this approximation is tested and evaluated in section 4. Finally, in section 5 we present a heavy-traffic approximation for an arriving customer's average queueing time, and draw conclusions.

2. FORMULATION

Consider the following system:

(*a*) Customers arrive singly at a one-server queueing system. The arrival rate is constant, equal to one customer per time period.

(*b*) An arriving customer who finds the server idle starts being served immediately; if upon arrival the customer finds the server busy, then he joins a queue. In either case, he cannot remain in the system longer than m time periods: if by the end of time m he has not completed service, then the customer is rejected from the system or, alternatively, reneges, never to return.

(*c*) Fully completed service times (or, alternatively, service requirements) have a negative exponential distribution. The average service requirement is constant, equal to $1/\lambda$, and we assume that $1/\lambda = \rho < 1$. It follows that, when the server is busy with customers who will eventually have their service completed, departures from the system have a Poisson distribution with mean λ.

(*d*) The queue discipline is FIFO; in a single-server queueing system, this

is equivalent to saying that customers leave the server in the order of their arrival.

Suppose that the number of customers is observed at the end of every period. The number of customers in the system indicates the total time spent by the oldest customer in the system. If the number of customers at the end of a period is equal to m, this means that the oldest customer has already spent m periods in the system, so he is removed from the system. This is equivalent to refusing entrance to the newly arriving customer at the beginning of the next period.

Let X_t denote the number of customers in the system at the end of period t, and Z_t the number of customers served during period t. Clearly, $\{X_t\}$ is a Markov Chain, whose transition probabilities $P_{i,j}$ are given by:

$$P_{i,j}=\begin{cases} =P(Z_t=i-j+1), & \text{if } 0<j\leq i+1 \text{ and } i\leq m-1 \\ =P(Z_t\geq i-j+1), & \text{if } j=0 \text{ and } i\leq m-1 \\ =P(Z_t=i-j), & \text{if } 0<j\leq m \text{ and } i=m \\ =P(Z_t\geq i-j), & \text{if } j=0 \text{ and } i=0 \\ =0, & \text{if } j>i+1 \end{cases} \tag{3}$$

where

$$P(Z_t=k)=e^{-\lambda}\lambda^k/k!, \quad k=0, 1, \ldots$$

To see these probabilities, we have to realize that the transition from any state i to state 0 is not necessarily observed when all $(i+1)$ customers are served (including the new arrival), but (and most probably) later (i. e., at the end of the next period), when *more* than i customers could have been served. A similar analysis can be found in Brodheim et al. [2].

If m is not too large, then we can obtain the steady-state probabilities $\{\pi_0, \pi_1, \ldots, \pi_m\}$ by solving the following system of equations:

$$\pi_j=\sum_{i=0}^{m}\pi_i P_{i,j}, \quad j=0, 1, \ldots, m \quad \text{and} \quad \sum_{i=0}^{m}\pi_i=1 \tag{4a}$$

We can then compute the measures of performance as follows:

$$W(\rho, m)=\pi_m \quad \text{and} \quad U(\rho, m)=\rho(1-\pi_m) \tag{4b}$$

However, the exact numerical calculation of π_i can be very cumbersome if m is large, and also does not provide us with the insight of an analytical (closed form) solution which is very useful for policy selection or sensitivity analysis. Brodheim, et al. [2] encountered the same problem in the context of a perishable inventory system.

An alternative approach for the analysis of the $D/M/1$; m queue would be to revise assumption (*b*) by considering instead that an arriving item reneges

(is rejected) if his queueing time reaches a fixed length, say m_q. If we denote by $U_k(\rho, m_q)$ the average server utilization in the $E_k/M/1$; m_q system, then from [4] we have:

$$U_k(\rho, m_q)=\rho\left[\sum_{r=1}^{k} a_r-1/(1-\rho)\right]\Big/\left[\sum_{r=1}^{k} a_r-\rho/(1-\rho)\right] \tag{5}$$

where

$$a_r=(\rho k+z_r)\exp(z_r m_q/\rho)/[(\rho k+z_r+k(z_r-1))(1-z_r)] \tag{6}$$

and where $\{z_r; r=1, 2, \ldots, k\}$ are the k non-zero roots of the equation:

$$(\rho k+z)^k(z-1)+(\rho k)^k=0 \tag{7}$$

The rejection probability, $W_k(\rho, m_q)$, is then, by (1):

$$W_k(\rho, m_q)=1-U_k(\rho, m_q)/\rho=1\Big/\left[\sum_{r=1}^{k} a_r-\rho/(1-\rho)\right] \tag{8}$$

However, the above formulas for $U_k(\rho, m_q)$ and $W_k(\rho, m_q)$ present serious computational difficulties, since they involve the calculation of k roots in the polynomial (7). The computational difficulty inherent in the solution of (7) will be overcome in the next section, where simple approximations for the above quantities, as well as for their limit as $k \to \infty$, are developped.

Before proceeding with the analysis, the following deserves to be mentioned: $W_k(\rho, m_q)$ in (8) is the probability of an item having to remain for time m_q in the queue rather than for time m in the system, as required by the original assumption (*b*). Intuitively, one would expect the two probabilities to be approximatively equal if m_q is defined as

$$m_q=m-1/\lambda=m-\rho \tag{9}$$

This intuitive relationship draws some support from the comparison between the two rejection probabilities $W(\rho, m)$ in (2) and $W_1(\rho, m_q)$: When $k=1$, the non-zero root of (7) is $z_1=1-\rho$, so that

$$W_1(\rho, m_q)=\rho(1-\rho)/(\exp(m_q(1-\rho)/\rho)-\rho^2) \tag{10}$$

a result first established by Barrer [1]. Equating the two probabilities in (2) and (10) yields:

$$m_q=m+\rho \ln \rho/(1-\rho) \tag{11}$$

which upon expanding ln ρ into a series of powers in $1-\rho$ indicates that (9) is approximately valid when ρ is relatively close to unity. Of course, the expression in (11) need not necessarily hold when the arrival rate is constant: both (9) and (11) will be tested in section 4, when some analytical values of $W(\rho, m)$ are compared with the corresponding approximate $W_\infty(\rho, m_q)$ values.

3. DERIVATION OF THE APPROXIMATIONS

From the definition of $U_k(\rho, m_q)$ it follows that $U_k(\rho, 0)$ corresponds to the utilization of the standard $E_k/M/1$ no-queue system, and $U_k(\rho, \infty)$ corresponds to that of the standard $E_k/M/1$ queue system with no reneging.

RESULT 1:

$$W_k(\rho, m_q)=f(\rho k)^k/(1+\rho k)^k=f\,W_k(\rho, 0) \tag{12}$$

where $f=f_k(\rho, m_q)$ is a weighting factor, to be specified later, with limiting values of 1 and 0 for $m_q=0$ and ∞, respectively, regardless of k and ρ.

Proof: It is not difficult to show that for given k and ρ, the average server utilization in (5) is an increasing function of m_q. We therefore proceed in writing a general relationship of the form:

$$U_k(\rho, m_q)=f\,U_k(\rho, 0)+(1-f)U_k(\rho, \infty) \tag{13}$$

From [9] we know that

$$U_k(\rho, 0)=\rho\,[1-(\rho k)^k/(1+\rho k)^k]$$

Given that $U_k(\rho, \infty)=\rho$, we get from (13)

$$U_k(\rho, m_q)=\rho\,[1-f(\rho k)^k/(1+\rho k)^k] \tag{13a}$$

and by using (1), the result follows.

We will now proceed to derive approximations for $f_k(\rho, m_q)$.

RESULT 2: The following equations provide approximations for $f_k(\rho, m_q)$:

(*a*) $$f_k(1, m_q)\simeq 3(k+1)^{k+1}\,[k^k(6km_q+8k+4)] \tag{14}$$

(*b*) $$f_k(\rho, m_q)\simeq \{(1-\rho^2)/[\exp(m_q(1-\rho)/\rho)-\rho^2]\, 3(k+1)(1+1/k)^k/[2(3km+k+2)]-2/(m+1)\}^+ \tag{15}$$

Proof: To show (14), we examine the behavior of (7) when $\rho \to 1$. It turns out that, in that case, one of the k roots of (7), say z_1, tends to zero whereas all other roots have negative real parts, whose absolute value is large. It follows, from (6), that when m_q is not too small, the terms $\{a_r; r=2, 3, \ldots, k\}$ are negligible, so that by (8)

$$W_k(1, m_q)\simeq [a_1-\rho/(1-\rho)]^{-1} \tag{16}$$

On expanding the left-hand side of (7) into a series of powers in z, it is not difficult to verify that when $\rho \to 1$,

$$z_1\simeq 2\rho k(1-\rho)/(2\rho k-k+1) -(1-\rho)^2\,[4\rho k(k-1)(3\rho k-k+2)]/[3(2\rho k-k+1)^3] \tag{17}$$

By introducing the above into (6) and expanding the exponential into a power series of z_1, we obtain after some algebra on (16):

$$W_k(1, m_q) \simeq 3(k+1)/(6km_q+8k+4) \tag{18}$$

and therefore (14) follows from (13*a*) with $\rho=1$. We note that according to either (9) or (11), m_q in (14) or (18) is equal to $m-1$.

To derive (15), we calculated $W_k(\rho, m_q)$ in (8) analytically for $k=2, 3, 4$, and tabulated values of $f_k(\rho, m_q)=W_k(\rho, m_q)/W_k(\rho, 0)$ over a wide range of m_q and ρ values. We found that $f_k(\rho, m_q)$ depends on all three parameters, but the difference

$$f_k(\rho, m_q)-f_1(\rho, m_q)$$

appeared to be insensitive to ρ, and therefore could be approximated by

$$f_k(1, m_q)-f_1(1, m_q).$$

Assuming this is also true for higher values of k (it will be validated in the next section), then

$$f_k(\rho, m_q) \simeq \{ f_1(\rho, m_q)+[f_k(1, m_q)-f_1(1, m_q)] \}^+ \tag{19}$$

where $\{ x \}^+ = \max(x, 0)$. Upon using (10) and (12) with $k=1$ to obtain $f_1(\rho, m_q)$, and (14) with $m_q=m-1$ to obtain $f_k(1, m_q)$, (15) follows, and therefore the result is proven (Note that $f_1(1, m_q)$ may also be obtained as the limit of $f_1(\rho, m_q)$ when $\rho \to 1$, using (10)).

A direct consequence of Result 2 is the estimation of the rejection probability in the $D/M/1$; m system which is obtained by letting $k \to \infty$ (deterministic arrivals):

RESULT 3: The rejection probability in the $D/M/1$; m system can be approximated by:

$$W_\infty(\rho, m_q) \simeq \{ (1-\rho^2)/[\exp(m_q(1-\rho)/\rho-\rho^2] + [3e/(6m+2)]-[2/(m+1]\}^+ . \exp(-1/\rho) \tag{20}$$

4. EVALUATION OF THE APPROXIMATION

In order to evaluate (20), we solved the steady-state equations in (4*a*) for $m=2, 3, 4, 6$ and 10, and $0.30 < \rho \leq 0.99$. We then compared the analytical values for the probability of a customer reneging (as obtained from (4*b*)) with the approximate rejection probabilities calculated from (20). Table 1 gives the percentage errors incurred; these are defined relative to the arrival quantity per period (equal to 1 for the system considered here) as 100· (approximate value-analytical value). We assume that m_q is given by (9). It can be

seen that the errors are quite small: almost always they are less than 1%, and in the important (realistic) region ($\rho > 0.80$) they are less than 0.67%. They also appear to be rather insensitive to m, and, as expected, they are found to tend to 0 when ρ tends to 1. Similar errors occurred under the assumption that m_q is given by (11). We can, therefore, assume that $m_q = m - \rho$, this being a simpler expression.

TABLE I

Évaluation of formula (20) *: percentage errors*

ρ \ m	2	3	4	6	10
0.30	− 0.14	− 0.00	− 0.00	− 0.00	− 0.00
0.40	− 0.85	− 0.09	− 0.00	− 0.00	− 0.00
0.50	− 1.25	− 0.50	− 0.10	− 0.00	− 0.00
0.60	− 1.26	− 0.72	− 0.51	− 0.05	− 0.00
0.70	− 1.02	− 0.31	− 0.34	− 0.35	− 0.02
0.80	− 0.67	0.03	0.17	− 0.03	− 0.21
0.90	− 0.31	0.14	0.32	0.38	0.21
0.95	− 0.14	0.10	0.21	0.30	0.30
0.99	− 0.01	0.02	0.05	0.08	0.09

5. THE AVERAGE QUEUEING TIME IN HEAVY TRAFFIC

In his analysis, Finch [4] also derives the queueing-time (excluding service) distribution function. Letting $Q_k(\rho, m_q)$ denote the average queueing time of an arriving customer, in time periods, we have, in the notation used previously,

$$Q_k(\rho, m_q) = \rho\left[\sum_{r=1}^{k} \{(1-z_r)a_r(1-\exp(-z_r m_q/\rho))\}/z_r - m_q/(1-\rho)\right] \Big/ \left[\sum_{r=1}^{k} a_r - \rho/(1-\rho)\right] \tag{21}$$

RESULT 4: Under conditions of heavy traffic ($\rho \to 1$) the average queueing time of an arriving customer is given by:

$$Q_k(1, m_q) \simeq m_q(4+2k+3km_q)/(4+8k+6km_q) \tag{22}$$

Proof: It was pointed out in section 3 that when $\rho \to 1$, one of the roots of (7), say z_1, tends to zero, whereas the remaining $k-1$ roots have negative real parts whose absolute value is large. It is not difficult to show that, if m_q is not too small, the terms in (21) that correspond to $r > 1$ are negligible in comparison to the one for $r = 1$. In order to prove (22) we then expand

$\exp(-z_1 m_q/\rho)$ into a power series of z_1, replace z_1 by its expression given in (17) and finally calculate the limit of (21) as ρ tends to 1.

To illustrate the implications of the above results, consider a queueing system $D/M/1;\infty$ which operates under conditions of heavy traffic, the average queueing time of an arriving customer being infinitely large. Imposing an upper limit, m_q, on each customer's queueing time has two consequences: It reduces the average queueing time to

$$Q_\infty(1, m_q) \simeq m_q(2+3m_q)/(8+6m_q) \tag{23}$$

time periods (obtained from (22) for $k \to \infty$) but causes a proportion

$$W_\infty(1, m_q) \simeq 3/(8+6m_q) \tag{24}$$

of arriving customers to spend m_q time periods in the queue and then be denied service (obtained from (18) for $k \to \infty$).

Letting c_q be the unit cost of queueing and c_w be the unit cost of refusing service, it can be seen that the expected total cost per arrival is equal to

$$C = [m_q(2+3m_q)/(8+6m_q)]c_q + [3/(8+6m_q)]c_w \tag{25}$$

Differentiating (25) with respect to m_q, we get

$$dC/dm_q = 2[(8+24m_q+9m_q^2)c_q - 9c_w]/(8+6m_q)^2 \tag{26}$$

Thus, if $c_w/c_q < 8/9$, dC/dm_q is positive over the whole range of m_q values and a system where no queueing is allowed (i. e., $m_q=0$) would yield the lowest total cost. If, however, $c_w/c_q \geq 8/9$, then (26) is minimized for:

$$m_q = (\sqrt{8+9(c_w/c_q)} - 4)/3 \tag{27}$$

and the optimum operating characteristics of the system can easily be derived.

Alternatively, if management imposes a limit such as, for example, that the rejection probability should not exceed 6 %, this would imply, by (24), that $m_q=7$. $Q_\infty(1, 7)$ would, by (23), become equal to 3.22 time periods, and, therefore, the customers who do receive service (i. e., 94 % of the total) would each have to remain on average in the queue for

$$[Q_\infty(1, 7) - 7W_\infty(1, 7)]/[1 - W_\infty(1, 7)]$$

or less than 3 time periods. The 6 % rejection probability would also imply, by (27) with $m_q=7$, that the ratio c_w/c_q is implicitly assumed to be of the order of 70.

REFERENCES

1. D. Y. BARRER, Queueing with Impatient Customers and Ordered Service, *Operations Research*, vol. 5, 1957, p. 650-656.

2. E. BRODHEIM, C. DERMAN, G. P. PRASTACOS, On the Evaluation of a Class of Inventory Policies for Perishable Products such as Blood, *Management Science*, vol. 21, 1975, p. 1320-1326.
3. J. W. COHEN, Single Server Queues with Restricted Accessibility, *Journal of Engineering Mathematics*, vol. 3, 1969, p. 265-284.
4. P. D. FINCH, Deterministic Customer Impatience in the Queueing System GI/M/1, *Biometrica*, vol. 47, 1960, p. 45-52.
5. B. GAVISH, P. J. SCHWEITZER, The Markovian Queue with Bounded Waiting Time, *Management Science*, vol. 23, 1977, p. 1349-1357.
6. B. GNEDENKO, I. N. KOVALENKO, *Introduction to Queueing Theory*, Israel Program for Scientific Translation, 1968.
7. J. LORIS-TEGHEM, On the Waiting Time Distribution in a Generalized Queueing System with Uniformly Bounded Sojourn Times, *Journal of Applied Probability*, vol. 9, 1972, p. 642-649.
8. R. E. STANFORD, Reneging Phenomena in Single Channel Queues, *Mathematics of Operations Research*, vol. 4, 1979, p. 162-178.
9. L. TAKACS, *Introduction to the Theory of Queues*, Oxford University Press, New York, 1962.
10. G. P. COSMETATOS, G. P. PRASTACOS, An approximation for a Perishable Inventory System with fixed size periodic replacements, *Adv. Management Studies*, vol. 2, 1983, p. 233-239.
11. S. NAHMIAS, Perishable Inventory Theory: A Review, *Operations Research*, vol. 29, 1982, p. 680-708.
12. R. DOLL, *Data Communication Networks*, Prentice-Hall, 1977.