Advances in Stochastic Modelling

Editors
J.R. Artalejo
A. Krishnamoorthy

Notable Publications, Inc., New Jersey, USA.

Advances in Stochastic Modelling

Edited by J. R. Artalejo and A. Krishnamoorthy

456 p. 26 cms.

ISBN 0-9665847-3-2

NOTABLE PUBLICATIONS, INC.
1049 Hillcrest Drive, Neshanic Station, NJ 08853, USA
notable@curium.com

Current printing (last digit) :
10 9 8 7 6 5 4 3 2 1

Printed in India at Gnanodaya Press, Chennai 600 034, India.

The distribution of the maximum orbit size of an $M/G/1$ retrial queue during the busy period

M. J. Lopez-Herrero[1] *and M. F. Neuts*[2]

Abstract. In this paper we study the distribution of the maximum orbit size before emptiness in a stable $M/G/1$ retrial queue. By a recursive scheme, the computation of that distribution is reduced to solving systems of linear equations.

***Key words and phrases*:** Single server queue, repeated attempts, busy period, phase type distribution.

1. Introduction

We consider a single server queueing system in which an arriving customer who finds the server busy is obliged to join a pool of unsatisfied customers, called the 'orbit'. From the orbit, each customer reapplies for service after a random amount of time.

The evolution of such queues exhibits an alternating sequence of idle and busy periods. And, in contrast to the standard $M/G/1$ queue, is possible to have an idle server while the orbit, and consequently the system, is not empty. We define the *busy period* of a retrial queue as the period of time between an epoch when an arriving customer finds an empty system and the first departure epoch at which the system is again empty.

Queues in which retrials are allowed have been widely used to model problems in telephone, computer and communication systems. A complete description of situations in which retrials arise is found in [**4**]. In addition, a complete bibliography is given in [**1**] and [**2**].

---

[1]School of Statistics, Complutense University of Madrid, Madrid-28040, Spain. Email : *lherrero@estad.ucm.es*

[2]Department of Systems and Industrial Engineering, The University of Arizona, Tucson, AZ 85721, USA. Email : *marcel@sie.arizona.edu*

The busy period is clearly related to the regeneration cycle of the process. The maximum of the orbit size during a busy period offers us an estimate of the level of congestion encountered over a long period of time. Moreover, it facilitates the choice of a truncation index in the numerical computation of the distribution of the busy period.

We study the distribution of the maximum number of customers in orbit during a busy period of the $M/G/1$ retrial queue. The corresponding problem, the distribution of the maximum queue length for the standard $M/G/1$ queue, has received much attention. For the Poisson queue, that distribution was determined by Neuts [7] by using taboo probabilities. Cohen [3] essentially solved the problem for $M/G/1$ and $G/M/1$ queues. Serfozo [10], [11] examines the asymptotic behaviour of the maximum of birth and death processes and related queues. For $M/M/s$ queues, an approximation to the maximum queue length distribution is developed in [6]. In Neuts [9], the author presents algorithms for various distributions, among which that of the maximum queue length during the busy period, as aids in computing waiting time distribution under different disciplines for the $M/PH/1$ queue.

The related literature on retrial systems is limited to Gomez-Corral [5] which deals with the asymptotic behaviour of the maximum orbit length over n regeneration cycles of the positive recurrent $M/G/1$ queue with constant retrial rate.

The paper is organised as follows. In Section 2, we describe the model. In Section 3, the distribution of the maximum orbit size is expressed in terms of the solutions to systems of linear equations. In Section 4, we present various numerical examples with service time distributions of phase type. These are used for the sake of versatility and simplifying algorithmic properties. Finally, conclusions are included in Section 5.

The authors plan to extend this work to the busy period distribution by using truncated models of the initial $M/G/1$ retrial system. The distribution of the maximum orbit size during the busy period is useful in selecting the truncation index for these models.

2. Model description

We consider a single server queue to which customers arrive according to a Poisson process with rate λ. Any arriving customer who finds the server busy joins the orbit. Each customer in orbit, independently of the rest of the customers in orbit, generates a Poisson stream of rate $\mu > 0$ of repeated requests for service. The service times are independent with common probability distribution function $B(x)$ ($B(0) = 0$), hazard function $b(x)$, Laplace-Stieltjes transform

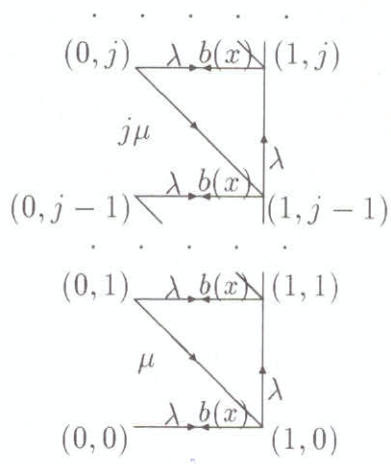

FIGURE 1. State space and transitions

$B(\cdot)$, and mean β_1. The flow of primary arrivals, the intervals between successive repeated attempts and the service times are all assumed to be mutually independent.

At any arbitrary time t, the system is described by the process $Y(t) = (C(t), N(t))$, where $C(t)$ is 0 or 1, according to whether the server is free or busy, and $N(t)$ is the number of customers in orbit at time t. As is well known, the stability condition $\rho = \lambda\beta_1 < 1$ guarantees that the limiting density of the process $Y(t)$ exists and is positive. Figure 1 describes the transitions among the states in the process $Y(t) = (C(t), N(t))$.

3. The maximum orbit size during a busy period

We restrict attention to stable queues, that is with $\rho < 1$. We consider the Markov chain embedded at departures epochs. Note that the busy period corresponds to the first return time of the state 0 in that chain.

Let M be the maximum orbit size attained during the busy period. We wish to determine the probability distribution of M. We recall that the probability density $\{c_k\}$ of the number of arrivals during a service time is given by

$$c_k = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^k}{k!} dB(t), \qquad k \geq 0. \tag{1}$$

Let $x_i^{(k)}$, $1 \leq i \leq k$, be the conditional probability that, starting in the state i, the embedded chain reaches state 0 before reaching state $k+1$. In other words,

$x_i^{(k)}$ is the probability that the busy period ends without exceeding k customers in orbit, given that we start at a departure with i customers remaining in orbit.

Then the probability distribution of M is related to the $\{x_i^{(k)}\}$ as follows

$$P\{M = 0\} = c_0,$$

$$P\{M \leq k\} = c_0 + \sum_{i=1}^{k} c_i x_i^{(k)}, \qquad k > 0. \tag{2}$$

The first equation (2) is obtained by noting that the orbit remains empty during the busy period if and only if no one customer arrives during the service time of the unique customer served. When $k > 0$, the event $\{M \leq k\}$ occurs when at least $i \leq k$ customers arrive during the first service time and subsequently the system empties out before reaching the state $k + 1$.

Equations for the probabilities $x_i^{(k)}$ are derived as follows: For $k = 1$, the preceding departure left one customer in orbit. The next customer to receive service may come from the orbit or may be a primary arrival. Moreover, during that customer's service there can be at most one arrival. Taking all alternatives into account, we obtain that $x_1^{(1)}$ satisfies the equation:

$$x_1^{(1)} = \frac{\lambda}{\lambda+\mu} c_0 x_1^{(1)} + \frac{\mu}{\lambda+\mu}\left(c_0 + c_1 x_1^{(1)}\right),$$

which leads to

$$x_1^{(1)} = \frac{\mu c_0}{\lambda(1-c_0) + \mu(1-c_1)}. \tag{3}$$

For $k \geq 2$, we use a similar argument. We distinguish between the alternatives where the next customer served is primary arrival or comes from the orbit. In either case, we partition the subsequent event on the number of arrivals that may occur during that service so that, at its end, there are j customers in the orbit and that, thereafter, the orbit size never reaches $k+1$ before the systems empties out. The following linear equations are so obtained for the quantities $x_i^{(k)}$, $1 \leq i \leq k$:

$$x_1^{(k)} = \frac{\mu c_0}{\lambda+\mu} + \sum_{j=1}^{k} \frac{\lambda c_{j-1} + \mu c_j}{\lambda+\mu} x_j^{(k)},$$

$$x_i^{(k)} = \frac{i\mu c_0}{\lambda+i\mu} x_{i-1}^{(k)} + \sum_{j=i}^{k} \frac{\lambda c_{j-i} + i\mu c_{j-i+1}}{\lambda+i\mu} x_j^{(k)}, \qquad \text{for } 2 \leq i \leq k. \tag{4}$$

For any fixed level $k \geq 1$, the system (4) can be written in matrix form, as

$$\mathbf{x}^k = T\mathbf{x}^k + A, \tag{5}$$

where $\mathbf{x}^k$ is the column vector with components $x_i^{(k)}$.

The symbols T, A represent, respectively, the $k \times k$ matrix and the k dimensional column vector

$$T = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1,k-1} & b_{1k} \\ a_2 & b_{22} & \cdots & b_{2,k-1} & b_{2k} \\ 0 & a_3 & \cdots & b_{3,k-1} & b_{3k} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & a_k & b_{kk} \end{pmatrix}, \qquad A = \begin{pmatrix} a_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

where

$$a_i = \frac{i\mu c_0}{\lambda + i\mu}, \qquad \text{for } 1 \leq i \leq k,$$

$$b_{ij} = \frac{\lambda c_{j-i} + i\mu c_{j-i+1}}{\lambda + i\mu}, \qquad \text{for } 1 \leq i \leq j \leq k.$$

The matrix $I - T$ is of upper Hessenberg form. The numerical solution of the system (5) by Gauss elimination is straightforward. Substitution into equation (2) yields the probability distribution of the maximum orbit size.

Note that, as is intuitive, when the retrial rate μ tends to infinity, reattempts are made so often that it appears that the next customer served is not in orbit but is next in the usual waiting line. Consequently, when $\mu \to \infty$, M should have the same probabilistic behaviour as the maximum number (excluding the customer in service) during the busy period of the standard $M/G/1$ queue. Letting μ approach infinity in (3) and (4), these equations converge to the corresponding ones (40) and (41) in [**9**].

4. Computational aspects and numerical results

Here, we present numerical examples on the behaviour of M. In these, we restrict ourselves to single server retrial queues with a service time distribution of phase type.

Probability distributions of phase type (PH-distributions) are related to finite-state Markov processes, have an appealing algebraic formalism which yields useful computational simplifications in algorithmic approaches. The hyperexponential and the generalised Erlang distributions, two types of probability distributions commonly used in applications, are of phase type. PH-distributions were introduced by Neuts; their definition and basic properties are discussed in [**8**]. One of those interesting properties refers to the computation of the probability density $\{c_n\}$ of the number of arrivals during a service time. For a PH service time distribution, that density is itself of (discrete) phase type and can be recursively computed without numerical integrations (see Theorem 2.2.8 in [**8**]).

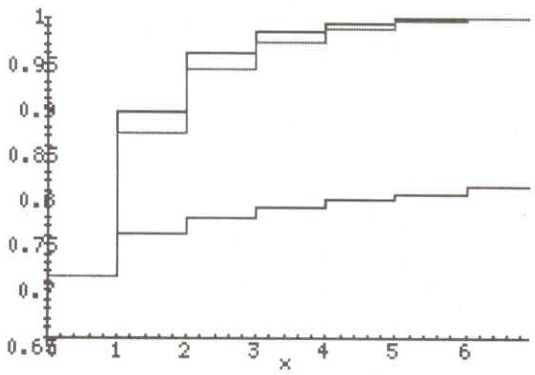

FIGURE 2. $F_M(x\,|\mu\,)$ when $\rho = 0.4$

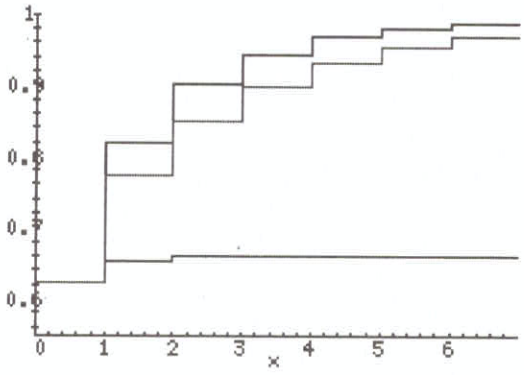

FIGURE 3. $F_M(x\,|\mu\,)$ when $\rho = 0.6$

We first consider an $M/M/1$ retrial queue with mean service time one, so that $\lambda = \rho$. In Figures 2, 3, and 4, for $\rho = 0.4, 0.6$, and 0.8 we display together the cumulative distribution functions of M, $F_M(x\,|\mu\,)$, corresponding to the values 0.1, 1.0 and ∞ of the retrial rate μ.

Note that when $\mu = \infty$, the retrial model is the classical $M/M/1$, and the explicit expression for $F_M(x\,|\infty\,)$ is

$$F_M(x\,|\infty\,) = 1 - \left(\frac{\rho^{-[x+1]} - \rho}{1-\rho}\right)^{-1}, \qquad x \geq 0,$$

which can easily be derived from the corresponding distribution of the maximum queue length in [**10**]. This function corresponds to the highest curves in Figures 2-4. Moreover, when μ increases, the cumulative functions tend to $F_M(x\,|\infty\,)$ and, at each point, their differences are smaller for low traffic intensities.

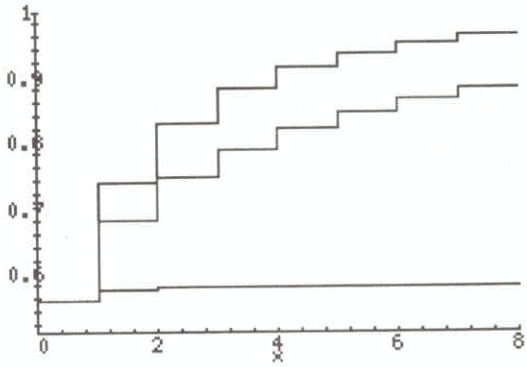

FIGURE 4. $F_M(x\,|\mu\,)$ when $\rho = 0.8$

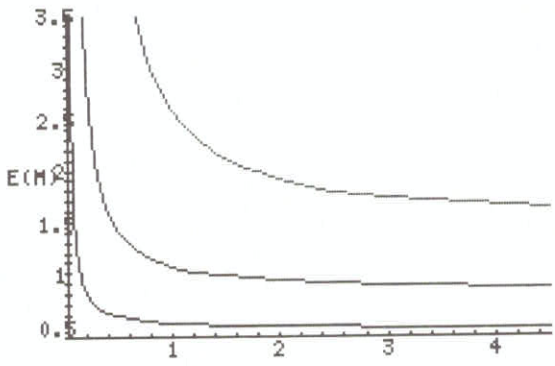

FIGURE 5. $E[M]$ verses μ

Figures 2, 3, and 4 show that the orbit congestion decreases with increasing retrial rates, and also with decreasing traffic intensities. Also, for each retrial intensity, the density $P\{M = k\}$, exhibits a decreasing behaviour for $k > 0$.

Figures 5-6 represent the expected value $E[M]$ of the maximum orbit size respectively as functions of μ and ρ. All curves in Figure 5, which from top to bottom, correspond to $\rho = 0.8,\ 0.6,\ 0.4$, have a decreasing shape. As is to be expected, for models with high retrial rates the orbit is less congested than for those having low retrial rates. Curves in Figure 6 correspond, from left to right, to the retrial rates $\mu = 0.1,\ 1.0,\ 10$, and show that the expected orbit occupancy increases with the traffic intensity. Table 1 also deals with the Markovian retrial model. It shows the influence of the retrial and ergodicity parameters on the coefficient of variation of M, $CV(M) = \frac{\sqrt{Var[M]}}{E[M]}$, the ratio between the

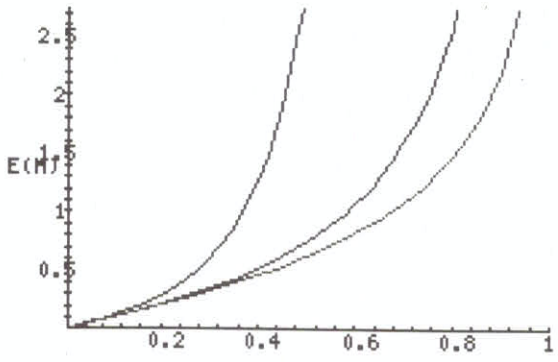

FIGURE 6. $E[M]$ versus ρ

ρ	$\mu = 0.1$	$\mu = 0.75$	$\mu = 1$	$\mu = 5$	$\mu = 10$	$\mu = 10^4$
0.05	4.635331	4.594469	4.592811	4.588803	4.588299	4.587795
0.1	3.436423	3.347250	3.343175	3.333147	3.331868	3.330585
0.2	2.666801	2.522401	2.513928	2.491988	2.489069	2.486110
0.3	2.346957	2.198643	2.187174	2.155741	2.151348	2.146840
0.4	2.119749	2.034521	2.021493	1.983318	1.977682	1.971815
0.5	1.874118	1.947580	1.934785	1.892947	1.886299	1.979259
0.6	1.601955	1.908379	1.898933	1.858047	1.850651	1.842617
0.7	1.377041	1.905208	1.905713	1.874929	1.867179	1.858361
0.8	1.230175	1.934797	1.961528	1.964034	1.956781	1.947524
0.9	1.130518	2.000914	2.105744	2.227503	2.223843	2.215151

TABLE 1. $CV(M)$ versus ρ and μ

standard deviation and the mean of M. Except for $\mu = 0.1$, the coefficient of variation tends to be large when ρ approaches either 0 or 1. For a fixed traffic intensity the coefficient of variation decreases for increasing retrial rates. Next, we let the service times have Erlang distributions with r exponential stages, each of mean length one.

Tables 2 and 3 summarize the obtained numerical results for $E[M]$ and $CV(M)$ by varying the retrial and ergodicity parameters. Each cell gives values of both measures for models corresponding to Erlang laws with 2 and 5 stages.

From Table 2 we see that, for each Erlang model, $E[M]$ increases with the traffic intensity, and decreases with increasing values of μ, as in Figures 5-6.

μ r=2 r=5	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	$\rho = 0.9$
0.05	0.281506	1.313578	7.996056	51.48338	172.3657
	0.236596	0.738313	2.513088	15.43906	57.52662
0.1	0.244000	0.812346	3.076184	20.09706	72.89178
	0.220955	0.584025	1.453398	5.491348	18.75898
0.25	0.221287	0.583044	1.429308	5.107928	16.40142
	0.211355	0.497743	1.006000	2.410408	5.140113
0.5	0.213554	0.513974	1.071280	2.698021	6.009824
	0.208104	0.469348	0.878723	1.792963	3.134000
1.	0.209639	0.480080	0.916920	1.923564	3.436290
	0.206467	0.455117	0.817960	1.537648	2.440461
2.5	0.207272	0.459796	0.830266	1.557986	2.447340
	0.205481	0.446550	0.782214	1.398387	2.097828
5.	0.206479	0.453027	0.802177	1.449359	2.184130
	0.205151	0.443688	0.770396	1.354021	1.993959
10^4	0.205686	0.446251	0.774433	1.346495	1.947936
	0.204822	0.440824	0.758628	1.310629	1.894817

TABLE 2. $E[M]$ versus μ and ρ and $Erlang_r$ service times

On another hand, when the service distribution has more stages the orbit is less congested. That finding also agrees with our intuition.

Table 3 is similar to Table 1; it also shows that $CV(M)$ decreases when number of stages increases.

In Table 4 we display the first indices for which the cumulative function of the maximum orbit size during the busy period exceeds 0.999. Each cell contains the indices for the queues with Markovian or Erlang service times distributions used in the previous analysis. These indices are useful in selecting truncation indices for other numerical computations. The entries in Table 4, show that maximum orbit content is essentially the same for models with low traffic intensity or with high retrial rates, i.e., for systems that are not congested.

Finally, we consider six retrial systems, $S_1 - S_6$, all with the same arrival rate $\lambda = 0.8$ and mean service time $\beta_1 = 1$. There are three choices for the retrial parameter $\mu = 0.1,\ 0.5$ and 1.0. The traffic intensity ρ equals 0.8 so that these systems are well saturated.

μ r=2 r=5	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	$\rho = 0.9$
0.05	2.576045	2.005558	1.469353	1.114489	1.016687
	2.420112	1.897986	1.636400	1.244867	1.054964
0.1	2.500418	1.992611	1.693489	1.268929	1.102950
	2.370848	1.838369	1.660037	1.522560	1.354494
0.25	2.434875	1.922872	1.760768	1.654509	1.519193
	2.334241	1.777971	1.611167	1.636023	1.749387
0.5	2.407161	1.879926	1.732217	1.762711	1.860306
	2.320427	1.750483	1.576303	1.620256	1.808372
1.	2.391739	1.851856	1.701068	1.766270	1.966187
	2.313160	1.734650	1.553709	1.590895	1.799337
2.5	2.381897	1.831983	1.674553	1.743736	1.971534
	2.308675	1.724327	1.537715	1.575189	1.778391
5.	2.378509	1.824728	1.664009	1.731129	1.960778
	2.307159	1.720732	1.531925	1.566912	1.768511
10^4	2.375066	1.817117	1.652438	1.715546	1.942984
	2.305626	1.717056	1.525884	1.557927	1.757077

TABLE 3. $CV(M)$ versus μ and ρ and $Erlang_r$ service times

The service times distributions are of phase type with representation (α, R), where α is an $r-$dimensional vector and R an $r \times r$ matrix. Specifically,

$S_1 : r = 1, \alpha = 1, R = 1.$
$S_2 : r = 2, \alpha = (1, 0), R(j, j) = -2, R(j, j+1) = 2.$
$S_3 : r = 5, \alpha = (1, 0, 0, 0, 0), R(j, j) = -5, R(j, j+1) = 5.$
$S_4 : r = 2, \alpha = (0.8, 0.2), R(1, 1) = -1.6, R(2, 2) = -0.4.$
$S_5 : r = 5, \alpha(j) = 1/5, R(j, j) = -3/j, j = 1, \cdots, 5.$
$S_6 : r = 5, \alpha(j) = 2^{-(j-2)}/5, R(j, j) = -1/2^{j-2}, j = 1, \cdots, 5.$

The first model is a Markovian queue, the second and third have Erlang service time distributions. The distributions for models $S_4 - S_6$ are hyperexponential. The columns in each percentile group correspond respectively to the retrial rates $\mu = 0.1,\ 0.5$ and 1.0. In spite of the fact that all models have the same traffic intensity, the entries in Table 5 show that, in general, their orbit behaviours are different. However, the distribution of M for models S_4 and S_5

μ	r=1 r=2 r=5	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$
0.05		7 5 4	23 14 9	75 39 20	293 144 63
0.1		5 4 3	15 10 7	43 24 14	157 81 39
1.		4 3 3	7 6 5	14 11 9	36 25 18
10		4 3 3	7 5 5	12 9 8	25 19 16
100		4 3 3	6 5 5	11 9 8	23 19 16

TABLE 4. 99'9% percentile of M

Model	$P(M=0)$	75%			90%			95%		
S_6	.675570	98	1	1	163	11	6	181	24	13
S_4	.6	124	3	2	142	16	8	151	26	14
S_5	.581929	119	4	2	132	17	9	139	25	14
S_1	.555555	117	5	3	127	18	9	133	25	14
S_2	.510204	114	7	3	121	19	9	125	24	13
S_3	.476113	111	9	4	117	19	9	120	24	13

with $\mu = 0.5, 1$, is essentially the same. Also note, that for each model, orbit congestion increases with low retrial rates.

5. Concluding remarks

A queueing system with repeated attempts is considered. The distribution of the maximum orbit size during the busy period, is determined from the solutions

Model	99%			99.9%		
S_6	213	54	38	254	94	75
S_4	166	42	29	187	62	48
S_5	151	38	26	168	54	40
S_1	143	36	24	157	49	36
S_2	134	33	21	144	43	31
S_3	127	31	20	136	39	28

TABLE 5. Selected percentiles of M for $\mu = 0.1, 0.5, 1$

of a recursive system of linear equations. Our results can be used to select an useful truncation index for those applications based on truncated models of the initial $M/G/1$ retrial system.

Acknowledgements. The work of the first author was supported by the DGES through project 98-0837. The research of Marcel Neuts was supported in part by Grant Nr. DMI-9988749 from the National Science Foundation.

Bibliography

[1] Artalejo, J.R. (1999). Accessible bibliography on retrial queues. *Math. Comput. Model.* **30**, 1-6.

[2] Artalejo, J.R. (1999). A classified bibliography of research on retrial queues: Progress in 1990-1999. *Top* **7**, 187-211.

[3] Cohen, J.W. (1967). The distribution of the maximum number of customers present simultaneously during a busy period for the queueing systems $M/G/1$ and $G/M/1$. *J. Appl. Prob.* **4**, 162-179.

[4] Falin, G.I. and Templeton, J.G.C. (1997). *Retrial Queues.* Chapman and Hall, London.

[5] Gomez-Corral, A. (2001). On extreme values of orbit lengths in $M/G/1$ queues with constant retrial rate. *OR Spektrum* **23**, 395-409.

[6] McCormick, W.P. and Park, Y.S. (1992). Approximating the distribution of the maximum queue length for $M/M/s$ queues. In *Queuing and Related Models*, ed. U.N. Bhat and I.V. Basawa, Clarendon Press, Oxford, 240-261.

[7] Neuts, M.F. (1964). The distribution of the maximum length of a Poisson queue during the busy period. *Oper. Res.* **12**, 281-285.

[8] Neuts, M.F. (1981). *Matrix Geometric Solutions in Stochastic Models. An Algorithmic Approach.* Dover Publications Inc, New York.

[9] Neuts, M.F. (1977). Algorithms for the waiting time distributions under various queue disciplines in the $M/G/1$ queue with service time distribution of phase type. In *Algorithmic Methods in Probability,* TIMS Studies in the Management Sciences, no. **7**. North-Holland Publishing Co., London, 177-197.

[10] Serfozo, R.F. (1988). Extreme values of birth and death processes and queues. *Stochastic Process. Appl.* **27**, 291-306.

Serfozo, R.F. (1988). Extreme values of queue lengths in $M/G/1$ and $GI/M/1$ systems. *Math. Oper. Res.* **13**, 349-357.