Efficient and Accurate Methodology for Solving Multiserver Retrial Systems

M. J. Doménech-Benlloch, J. M. Giménez-Guzmán,

J. Martínez-Bauset and V. Casares-Giner

We propose a novel methodology for solving retrial systems which is based on the aggregation of levels of the Markov model beyond a given one. Its evaluation concludes that is more accurate than previous approximations while requiring a low computational cost.

Introduction:

A common assumption when evaluating the performance of communication systems is that users that do not obtain an immediate service leave the system without retrying. However, due to the increasing number of users and the complexity of current systems the impact of retrials is no longer negligible. This is particularly true in mobile cellular networks [1].

In many cases the dimension of Markovian models makes it necessary to resort to approximate methodologies to solve them. In this letter we develop

a methodology to reduce the state space in such a way that the accuracy is not compromised and the computation cost is greatly reduced. Our work has been motivated by the study in [2], where the authors develop different models to study the impact of retries in a single service mobile cellular network. One of the main contributions of the study is an approximation to reduce the dimension of the state space, which the authors show is a good approximation to obtain accurate values for the the blocking probability (P_b). We postulate that this approximation was too simple to obtain accurate values for other common performance parameter used in retrial systems like the immediate service probability (P_{is}), the delayed service probability (P_{ds}) and the non-service probability (P_{ns}), being $P_{is} + P_{ds} + P_{ns} = 1$. To justify our postulate we study a finite population model that can be considered representative of the systems studied in [2]. Finite population models are more appropriate to evaluate the impact of retrials [2, 3], although our methodology can be equally applied to infinite population models.

Application Model:

We study a system in which a group of M users contend for access to a system with C servers, requesting an exponentially distributed service time with rate μ. When a new request finds all servers busy, it is moved to a retrial orbit with infinite capacity. Calls in the orbit retry after an exponentially distributed time with rate μ_r. The retry is successful if it finds a free server.

Otherwise, the user leaves the system with probability P_i or goes back to the retrial orbit with probability $(1-P_i)$. It is clear that P_i models the customer impatience.

The arrival process is modeled as an state-dependent Poisson process with rate $\lambda\ (k,m) = (M-k-m)\lambda$, being λ the individual user arrival rate when idle, and k (m) the number of users in service (retrying). The infinitesimal generator matrix $(\mathbf{Q})$ presents a tridiagonal structure being its elements also matrices, i.e. a quasi-birth-death (QBD) process. The stationary state probability vector π can be obtained by solving $\pi\mathbf{Q} = \mathbf{0}$. The desired performance parameters we can be computed by using the stationary state probability distribution.

Proposed Approximation Methodology:

We define the state space as $S := \{(k,m) : 0 \leq k \leq C; 0 \leq m \leq M-C\}$, where k is the number of occupied servers and m is the number of users retrying. We also define a threshold (Q) for aggregating states when there are Q or more users in the retrial orbit, i.e. states (k,Q) correspond to the situation where Q or more users are retrying. Figure 1 shows the state transition diagram of the proposed approximate model, where the first $Q-1$ columns remain the same as in the exact model. For column Q we approximate the arrival rate of new users by $\lambda(k,Q) = (M-k-\overline{m})\lambda$, where $\overline{m}$ denotes the average number of users retrying when the orbit holds Q or

more users. When a user executes a successful retrial, the number of users retrying can drop below Q with probability $(1-p)$ or not with probability p. Therefore, the retrial rate in states (k,Q) can be split in two contributing rates. The first one corresponds to transitions from (k,Q) to $(k+1,Q-1)$ and is approximated by $\alpha = \overline{m}\mu_r(1-p)$, and the second one from (k,Q) to $(k+1,Q)$ and is approximated by $\beta = \overline{m}\mu_r p$.

Parameters p and $\overline{m}$ can be estimated balancing the probability flux crossing each vertical cut of the state transition diagram and are given by

$$p = \frac{\lambda(C,Q)\pi(C,Q)}{\lambda(C,Q-1)\pi(C,Q-1) + \lambda(C,Q)\pi(C,Q)}$$

$$\overline{m} = \frac{\lambda(C,Q-1)\pi(C,Q-1) + \lambda(C,Q)\pi(C,Q)}{\mu_r[\sum_{k=0}^{k=C-1}\pi(k,Q) + P_i\pi(C,Q)]}$$

To find the values of p and $\overline{m}$ an iterative procedure must be followed starting with $p=0$ and $\overline{m}=Q$. We have used a relative precision of 10^{-4} for that procedure. In [2] the convergence of the iterative procedure was assumed. We evaluated a wide range of scenarios with different configuration parameters and the procedure converged in all cases.

Results:

The numerical evaluation is done by computing the relative error of the performance parameters respect to the exact model, defined as $\mid P_{xx}^{exact} - P_{xx}^{approx} \mid / P_{xx}^{exact}, xx \in \{b, is, ds, ns\}$. For the evaluation we have chosen

the following system parameter values: $M = 120$ users, $C = 30$ servers, $\mu^{-1} = 180$ seconds, $\mu_r^{-1} = 10$ seconds and $P_i = 0.5$. We define the offered load as $\rho = \lambda/(\lambda + \mu)$ and vary it from $\rho = 0.14$ to $\rho = 0.44$. We tested different algorithms [4, 5] for solving the resulting QBD, selecting the Algorithm 0 proposed in [4]. This choice only affects the computational cost not the accuracy.

Figure 2 shows the relative error of the performance parameters for two values of Q. $Q = 1$ represents the case described in [2] while $Q = 10$ represents a compromise between computation cost and accuracy. Note that using $Q = 1$ might not be a good choice because the relative error in P_{is}, P_{ds} and P_{ns} is not negligible. In general, the relative error in all performance parameters decreases as Q increases towards the exact model $Q = M - C$. As observed in Fig. 2, the precision is also a function of the system load, given that as more load is offered more users will be retrying.

In regard to the state space reduction, with $Q = 10$ we benefit from a 90% reduction while achieving a very low relative error. The computation cost savings for a systems with $\rho = 0.22$ ($P_b = 0.12$) is of 99.95% for $Q = 1$ and 99.5% for $Q = 10$, with respect to the exact model. Finally, note that a rule of thumb to determine a suitable value for Q could be to try with a value around $Q \simeq 0.15(M - C)$, independently of the system load, although lower values would be probably enough. We have checked this rule in several

practical scenarios and we found that it was a good choice in all of them.

Conclusions:

We propose a novel methodology to compute the value of typical performance parameters in systems with retrials. Our approximation methodology substantially improves the accuracy of previous approximations [2] with a very small computation cost increase. Moreover, when the computation of the exact model might not be feasible, our approach makes it possible to gradually increase the complexity of the approximate model until the relative accuracy between two successive approximations falls below a given precision objective.

Acknowledgements:

This work has been supported by the Spanish Government and the European Commission through projects TIC2003-08272 and TEC2004-06437-C05-01 (30% PGE and 70% FEDER), by the Spanish Ministry of Education and Science under contract AP-2004-3332 and by the Universidad Politécnica de Valencia under Programa de Incentivo a la Investigación. The authors express their sincere thanks to Jesús Artalejo and Vicent Pla for their insightful comments.

References

[1] ONUR, E., DELIÇ, H., ERSOY, C., and ÇAGLAYAN, M.U.: 'Measurement-based replanning of cell capacities in GSM networks', *Computer Networks*, 2002, **39**, pp. 749-767.

[2] MARSAN, M. A., DE CAROLIS, G., LEONARDI, E., LO CIGNO, R., and MEO, M.: 'Efficient estimation of call blocking probabilities in cellular mobile telephony networks with customer retrials', *IEEE J. Sel. Areas in Commun.*, 2001, **19**, (2), pp. 332-346.

[3] TRAN-GIA, P.,and MANDJES, M.,: 'Modeling of customer retrial phenomenon in cellular mobile networks', *IEEE J. Sel. Areas in Commun.*, 1997, **15**, (8), pp. 1406-1414.

[4] SERVI, L. D.: 'Algorithmic Solutions to Two-Dimensional Birth-Death Processes with Application to Capacity Planning', *Telecommunication Systems*, 2002, **21**, (2-4), pp. 205-212.

[5] GAVER, D. P., JACOBS, P. A., LATOUCHE, G.: 'Finite birth-and-death models in randomly changing environments', *Adv. Appl. Prob.*, 1984, **16**, pp. 715-731.

Authors' affiliations:

M. J. Doménech-Benlloch, J. M. Giménez-Guzmán, J. Martínez-Bauset and

V. Casares-Giner (Department of Communications, Polytechnic University of Valencia, Camí de Vera s/n, 46022, Valencia, Spain)

Corresponding e-mail:

mdoben@doctor.upv.es

Figure Captions:

Fig. 1 Approximate Markov model.

Fig. 2 Relative error for the performance parameters.

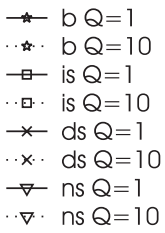


Figure 1

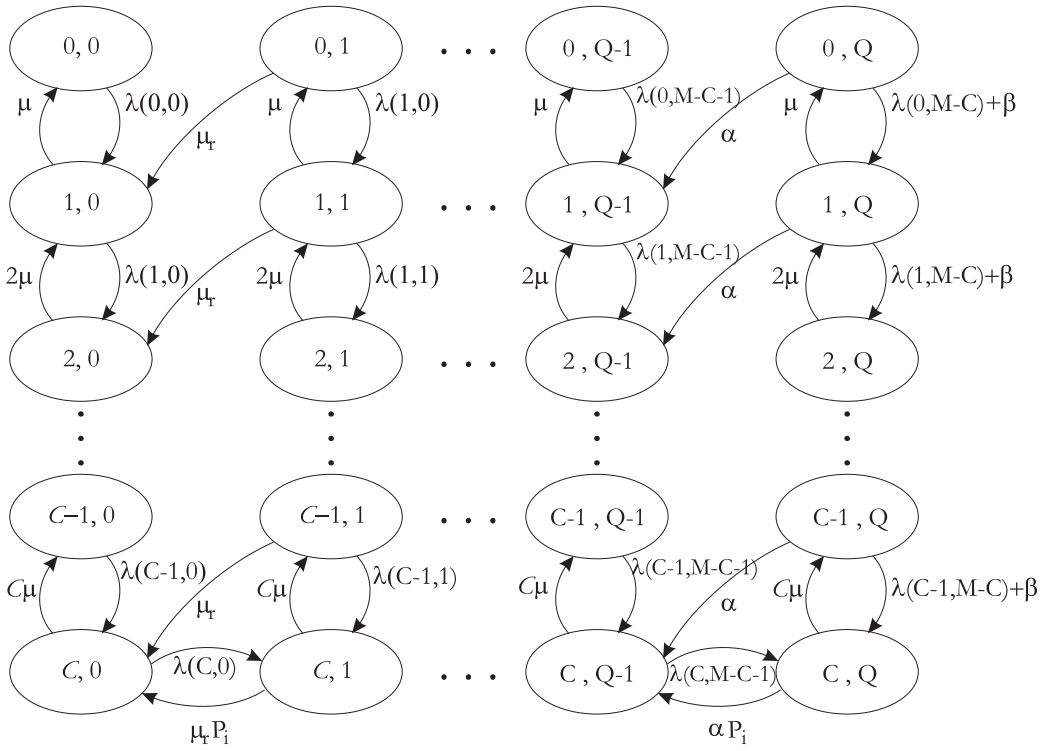
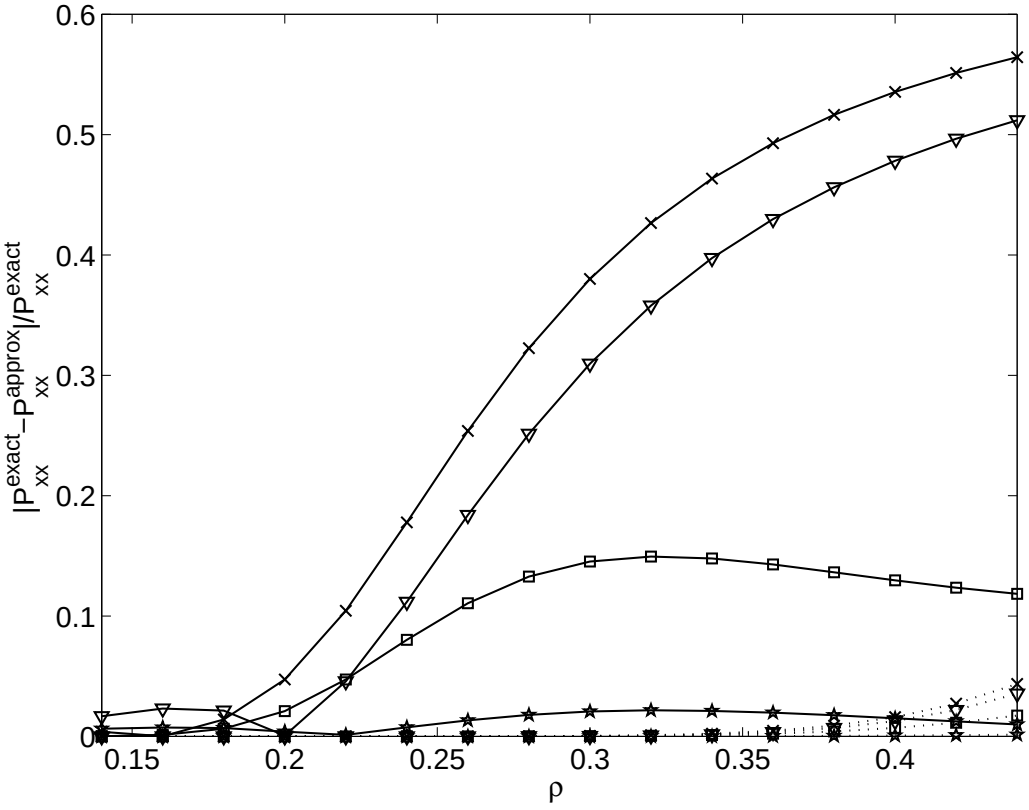

0, 0
0, 1
0 , Q-1
0 , Q
1, 0
1, 1
1 , Q-1
1 , Q
2, 0
2, 1
2 , Q-1
2 , Q
C–1, 0
C–1, 1
C-1 , Q-1
C-1 , Q
C, 0
C, 1
C , Q-1
C , Q
μ
λ(0,0)
μr
λ(1,0)
λ(0,M-C-1)
α
λ(0,M-C)+β
2μ
λ(1,1)
λ(1,M-C-1)
λ(1,M-C)+β
Cμ
λ(C-1,0)
λ(C-1,1)
λ(C-1,M-C-1)
λ(C-1,M-C)+β
λ(C,0)
λ(C,M-C-1)
μrPi
αPi

Figure 2


0.6
0.5
0.4
0.3
0.2
0.1
0
0.15
0.2
0.25
0.3
0.35
0.4
ρ
|P^exact_xx − P^approx_xx|/P^exact_xx