HETEROGENEOUS FINITE-SOURCE RETRIAL QUEUES

B. Almási, * G. Bolch, J. Sztrik
University of Debrecen, Debrecen, Hungary
* University of Erlangen, Erlangen, Germany
{almasi|jsztrik}@math.klte.hu

1 Introduction

Retrial queueing systems (or queueing systems with repeated attempts, returning customers, calls or orders) have wide practical use in designing local area networks and telecommunication systems. These queues are characterized by the following feature: a primary request finding all servers and waiting positions busy upon arrival leaves the service area, but after some random time he repeats his demand. Between retrials he is said to be in *'orbit '*. So the repeated attemps for service from the group of blocked customers are superimposed on the normal stream of arrivals of primary requests, thus the retrial queues can be considered as alternative to queues with losses that do not take repeated attemps into account.

In recent years, there has been considerable interest in retrial queues. For a systematic account of the fundamental methods and results, furthermore an accessible classified bibliography on this topic the interested reader is referred to, for example [2], [5], and references therein.

In many practical situations it is important to take into account the fact that the rate of generation of new primary calls decreases as the number of customers in the system increases. This can be done with the help of finite-source, or quasi-random input models. Queueing systems without retrials, that is systems with classical waiting lines and finite population have been review in detail by Takagi [15]. Since Kornyisev [12], which was the first paper devoted to finite-source retrial queues, there has been a rapid growth in the literature on this topic. A complete survey on related results can be found in Artalejo [2] for systems of type $M/G/1//K$ and $M/M/c//K$ in Kendall's notation. In addition, in the papers Falin, Artalejo [6], Falin [7] not only the outside observer's distributions of the systems in steady state, but also the stationary performance characteristics are considered on more detail. In particular, all main measures were expressed in terms of the server utilization. Arriving customer's distribution of the system state, busy period and waiting time processes (which is especially complex for retrial queues due to the overtaking) were investigated, too.

[1]Research is partially supported by NATO Scholarship for Senior Researchers 2002, German-Hungarian Bilateral Intergovernmental Scientific Cooperation, OMFB-DLR No. 21-2000, Hungarian Scientific Research Found OTKA T0-34280/2000 and FKFP grant 0191/2001.

Retrial queues with quasi-random input are recent interest in modelling magnetic disk memory systems [13], cellular mobile networks [16], computer networks [8], and local-area networks with nonpersistent CSMA/CD protocols with star topology [9], with random access protocols [10], and with multiple-access protocols [11].

An examination of these papers shows no investigations on models with heterogenous sources. Thus the aim of the present paper is to analyse a finite-source retrial queues with the following assumptions. There are K different sources of primary calls each consisting a single request. When source i is free at time t (i.e. is not being served and not waiting for service) it may generate a primary call during interval $(t, t+dt)$ with probability $\lambda_i dt + o(dt)$. If the server is idle at time of its arrival then the service starts. The service is finished during the interval $(t, t+dt)$ with probability $\mu_i dt + o(dt)$. During the service time the source cannot generate a new primary request. After service the source moves into a free state and can generate a new call. If the server is busy at time of the arrival of request from the ith source, then the source starts generating a Poisson flow of repeated calls with rate ν_i until it finds the server free. As before, after service the source becomes free and can generate a new primary call. All the times involved in the model are assumed to be mutually independent of each other.

Our objective is to give the main usual stationary performance measures of the system and to show the effect of different parameters on them. To achive this goal a tool called MOSEL (Modeling, Specification and Evaluation Language) developed at the University of Erlangen, Germany, see [4], is used to formulate and solve the problem. We show how this system can be modelled, and how easily performance measures can graphically be represented using IGL (Intermediate Graphical Language). This model is another extention of investigations for heterogeneous finite-source queueing systems that in the case of the classical service disciplines (first-come-first-served, polling, processor-shared, priority processor-shared) were treated, for example in Sztrik [14] and Takagi [15].

2 The $\vec{M}/\vec{M}/1//K$ retrial queueing model

2.1 The underlying Markov chain

Because of the exponentiality of the involved random variables the following process will be a Markov chain. The state of the system at time t can be described by the process $X(t) = ((\alpha_{C(t)}; \beta_1, ..., \beta_{N(t)}), t \geq 0)$ where $C(t) = 0$ if the server is idle, $C(t) = 1$ if the server is busy, and $\alpha_C(t)$ is the index of the request under service at time t. $N(t)$ is the number of sources of repeated calls at time t, and because of the heterogeneity of the sources we need to identify their indices that are denoted by β_j, $j = 1, ..., N(t)$ if there is a customer in the orbit, otherwise the second component is 0. Since its state space is finite the process $(X(t), t \geq 0)$ is ergodic for all values of the rate of generation of new primary calls, and from now on we assume that the system is in the steady state.

We define the stationary probabilities

$$P(0;0) = \lim_{t\to\infty} P(C(t) = 0, N(t) = 0)$$

$$P(j;0) = \lim_{t\to\infty} P(\alpha_1 = j, N(t) = 0), \qquad j = 1, ..., K$$

$$P(0; i_1, ..., i_k) = \lim_{t\to\infty} P(C(t) = 0, \beta_1 = i_1, ..., \beta_k = i_k), \qquad k = 1, ..., K-1$$

$$P(j; i_1, ..., i_k) = \lim_{t\to\infty} P(\alpha_1 = j, \beta_1 = i_1, ..., \beta_k = i_k), \qquad k = 1, ..., K-1.$$

The traditional way is to derive the related Kolmogorov equations for these probabilities and using the norming condition somehow we have to solve the set of equations. In our case these two steps are performed by the help of the tool treated in the next subsection.

Once we have obtained these limiting probabilities the **main system performance measures** can be derived in the following way.

1. The server utilization with respect to source j

$$U_j = \text{P (the server is busy with source } j \text{)}$$

that is, we have to summarize all the probabilities where the firsr component is j. Formaly

$$U_j = \sum_{k=0}^{K-1} \sum_{i_1,...,i_k \neq j} P(j; i_1, ..., i_k)$$

Hence the *server utilization*

$$U = E[C(t) = 1] = \sum_{j=1}^{K} U_j.$$

Let us denote by $P_W^{(i)}$ the steady state probability that request i is waiting (staying in the orbit). It is easy to see that

$$P_W^{(i)} = \sum_{j=0, j\neq i}^{K} \sum_{k=1}^{K-1} \sum_{i\epsilon(i_1,...,i_k)} P(j; i_1, ..., i_k).$$

Similarly, it can easily be seen, that the steady state probability $P^{(i)}$ that request i is in the service facility (it is under service or waiting in the orbit) is given by

$$P^{(i)} = P_W^{(i)} + U_i.$$

2. Mean response time of source i

The derivation of the following formulae are based on [15]. Let us denote by $E[T_i]$ the mean response time of customer i, and by γ_i the *throughput* of request i, that is, the mean number of times that request i is served per unit time. These are related by

$$\gamma_i = \frac{1}{E[T_i] + 1/\lambda_i} = \lambda_i(1 - P^{(i)}) = \mu_i U_i, \qquad i = 1, ..., K. \tag{1}$$

For $P^{(i)}$ we have

$$P^{(i)} = \frac{E[T_i]}{E[T_i] + 1/\lambda_i} = \gamma_i E[T_i] = 1 - \frac{\gamma_i}{\lambda_i} \qquad i = 1, ..., K. \tag{2}$$

which represents **Little's theorem** for request i in the service facility. It is easy to see that as a concequence of (1) we have

$$P^{(i)} = 1 - \frac{\mu_i}{\lambda_i} U_i$$

and

$$P_W^{(i)} = P^{(i)} - U_i = 1 - \frac{\mu_i + \lambda_i}{\lambda_i} U_i.$$

Alternatively, by the help of (2) we can express the mean response time $E[T_i]$ for request i in terms of U_i as

$$E[T_i] = \frac{P^{(i)}}{\lambda_i(1 - P^{(i)})} = \frac{1 - \frac{\mu_i}{\lambda_i} U_i}{\mu_i U_i} = \frac{\lambda_i - \mu_i U_i}{\lambda_i \mu_i U_i}.$$

3. Mean waiting time of source i

The mean waiting time of request i is given by

$$E[W_i] = E[T_i] - 1/\mu_i = \frac{1}{\gamma_i} - \frac{1}{\lambda_i} - 1/\mu_i = \frac{\lambda_i - (\mu_i + \lambda_i)U_i}{\lambda_i \mu_i U_i}.$$

At the same time we have another **Little's theorem** for request i waiting for service. Namely

$$P_W^{(i)} = \frac{E[W_i]}{E[T_i] + 1/\lambda_i} = \gamma_i E[W_i] \qquad i = 1, ..., K.$$

4. Mean number of calls staying in the orbit or in service

$$M = E[C(t) + N(t)] = \sum_{i=1}^{K} P^{(i)} = \sum_{i=1}^{K} (1 - \frac{\mu_i}{\lambda_i} U_i) = K - \sum_{i=1}^{K} \frac{\mu_i}{\lambda_i} U_i.$$

5. Mean number of sources of repeated calls

$$N = E[N(t)] = \sum_{i=1}^{K} P_W^{(i)} = \sum_{i=1}^{K} (1 - \frac{\mu_i + \lambda_i}{\lambda_i} U_i) = K - \sum_{i=1}^{K} \frac{\mu_i + \lambda_i}{\lambda_i} U_i.$$

6. Mean rate of generation of primary calls

$$\overline{\lambda} = \sum_{i=1}^{K} \gamma_i = \sum_{i=1}^{K} \lambda_i (1 - P^{(i)}) = \sum_{i=1}^{K} \mu_i U_i.$$

7. Blocking probability of primary call i

$$B_i = \frac{\lambda_i \sum_{j=1, j \neq i}^{K} \sum_{k=0}^{K-1} \sum_{i \neq i_1, ..., i_k} P(j; i_1, ..., i_k)}{\overline{\lambda}}.$$

Hence *blocking probability of primary calls*

$$B = \sum_{i=1}^{K} B_i$$

which is the fraction of primary calls which were blocked (i.e. met the server busy).

In particular, in the case of homogeneous calls, that is, when $\lambda_i = \lambda, \quad \mu_i = \mu, \quad \nu_i = \nu, \quad i = 1, ..., K$ the corresponding main performance measures treated in [6, 5] are the following

$$U_i = E[C(t)]/K, \quad i = 1, ..., K, \quad N = K - \frac{(\lambda + \mu)U}{\lambda},$$

$$\overline{\lambda} = \lambda E[K - C(t) - N(t)] = \mu U,$$

$$E[W] = N/\overline{\lambda} = K - (\mu U)^{-1} - \lambda^{-1} - \mu^{-1},$$

$$B = \frac{\lambda E[K - C(t) - N(t); C(t) = 1]}{\lambda E[K - C(t) - N(t)]}$$

as it can be seen in [6] with the appropriate changing of notations.

It should also be mentioned that all the performance measures can be expressed in the terms of the corresponding utilizations U_i, as it was stated in [6]. However, we must admit the distribution function and moments of the waiting times could not be solved.

2.2 MOSEL program

The crutial part of the whole modeling is the formulation of problem and derivation of the main performance measures. This can be done quite easily by the help of a tool called MOSEL . It automatically generates the state probabilities and using these a result file containing the performance measures specified in the model description file. It also generates a graphical representation of the results if the input file contains a picture part.

Since in our paper we concentrate on the effect of different parameters of the system on the main performance measures the technical details of programming are not treated.

3 Numerical examples

In this section several sample numerical results illustrate the power of the tool showing the influence of different parameters on the mean response time $E[T_i]$. In homogeneous case the results were validated by the Pascal program given in [5] in pages 272-274. For the easier representation of graphics we deals with 3 sources, but in the different setups the request generation, retrial and service rates play different roles. All the time we display the mean

response time $E[T_i]$ of call i as the function of the above mentioned rates. First we consider totally homogenous systems then systems with mixed group of rates are investigated. Finally rather complex situation is analyzed showing the unpredictable operational behaviour of the finite-source retrial system.

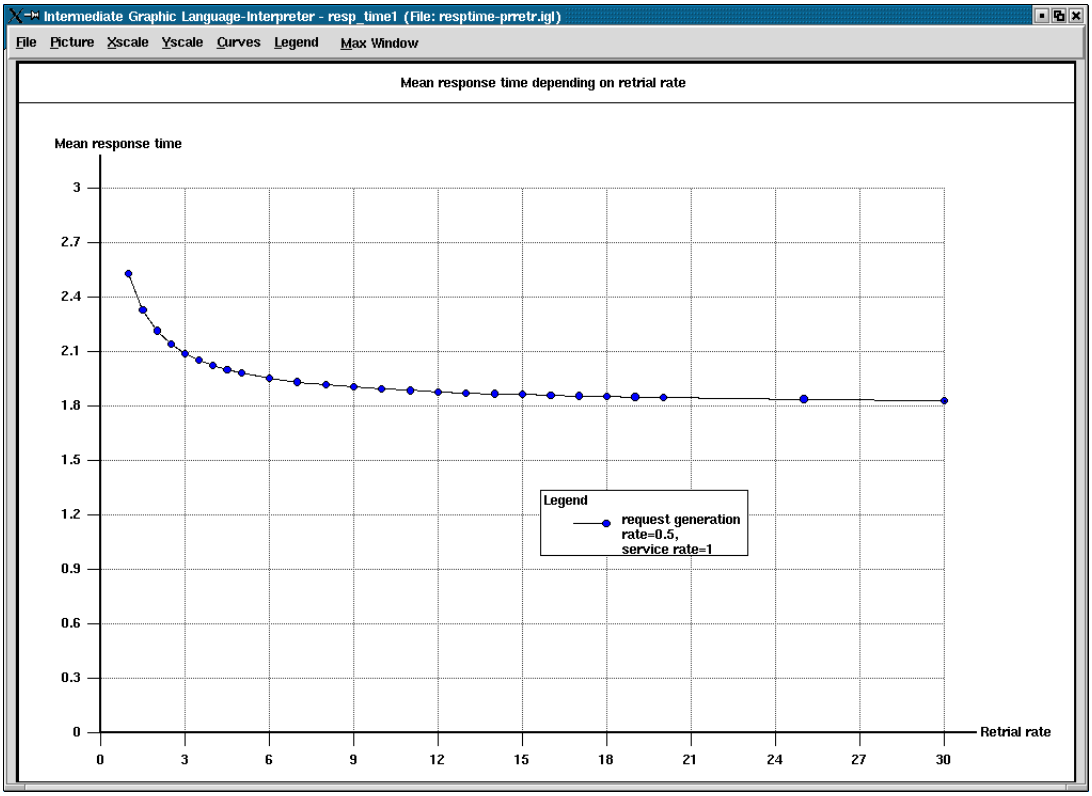

Figure 1: $E[T]$ versus retrial rate

3.1 Comments

In Figure 1 we can see how the mean reponse time $E[T]$ for the retrial system approaches the mean time for the classical system without retrial as the retrial rate increases. This is not suprising since in this case the retrial system becomes a FCFS finite-source queue. In Figure 2 we present 3 curves with the same different retrial rates as it was considered in [6] showing a suprising phenomenon of retrial queues having a maximum of $E[T]$. In Figures 3,4 different systems with mixed group of rates are investigated. Finally in Figure 5 a rather complex situation is considered as the function of primary request generation rate with fixed heterogeneous service and heterogeneous retrials.

Acknowledgements

We thank János Roszik for his programming work with MOSEL.

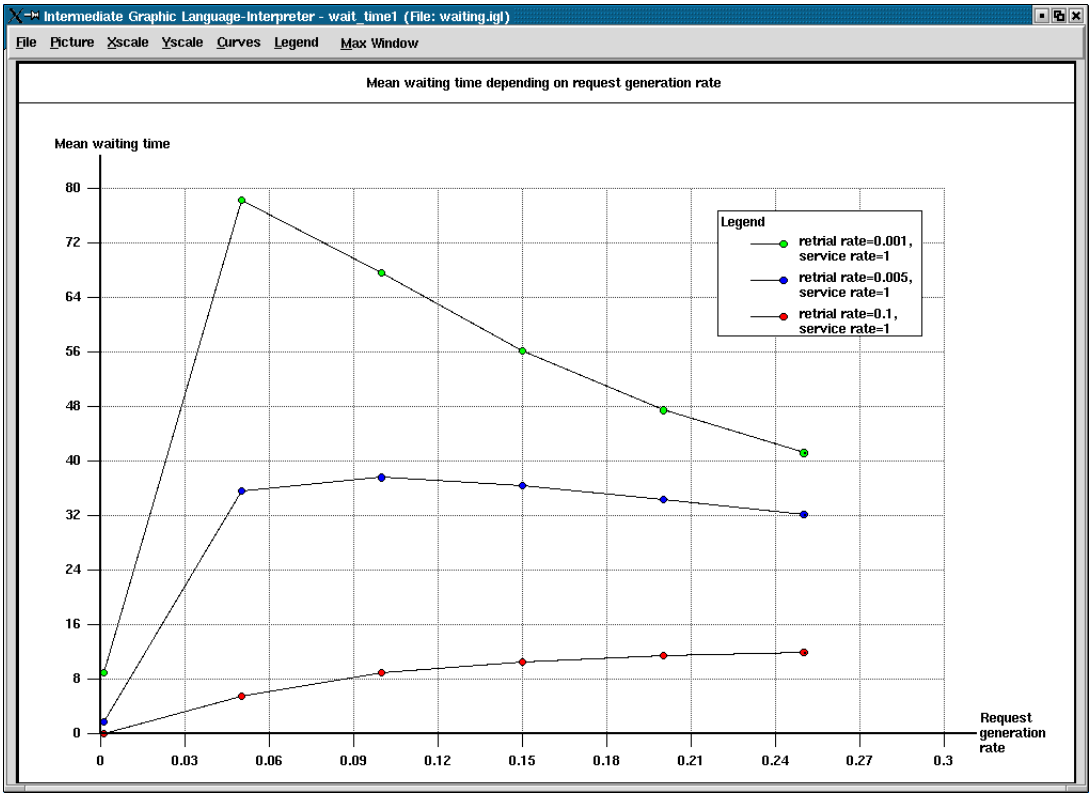

Figure 2: $E[T]$ versus primary request generation rate

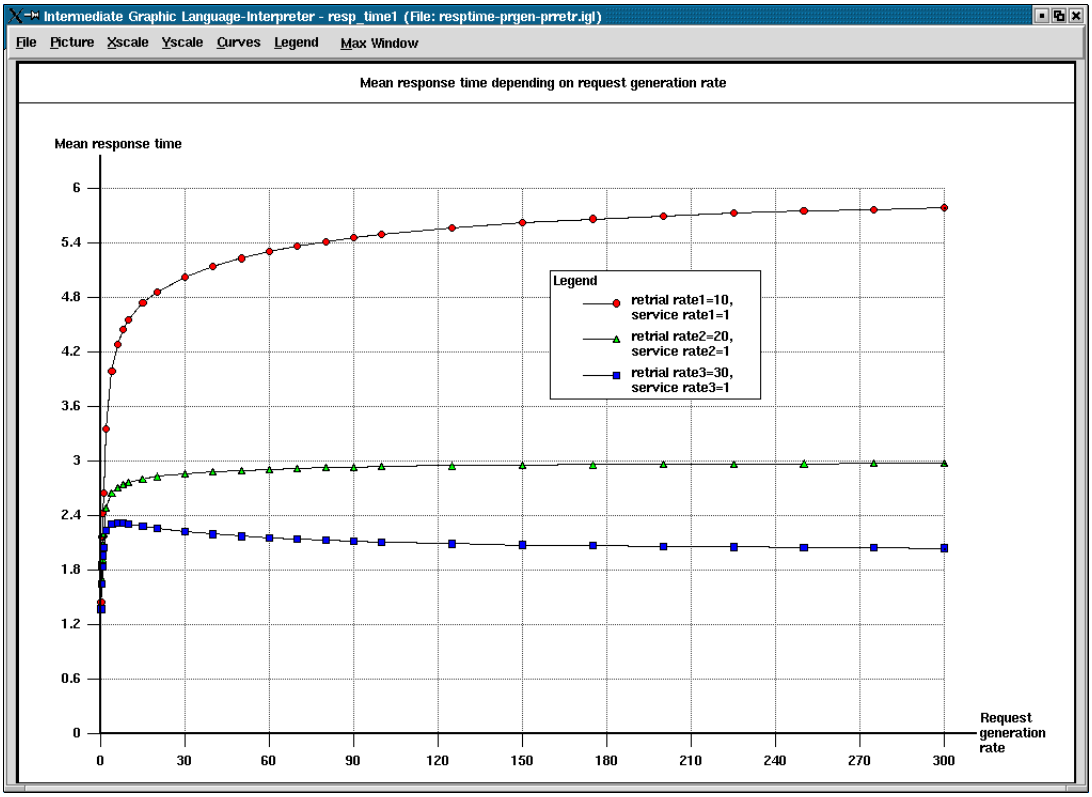

Figure 3: $E[T]$ versus primary request generation rate with homogeneous service and heterogeneous retrial

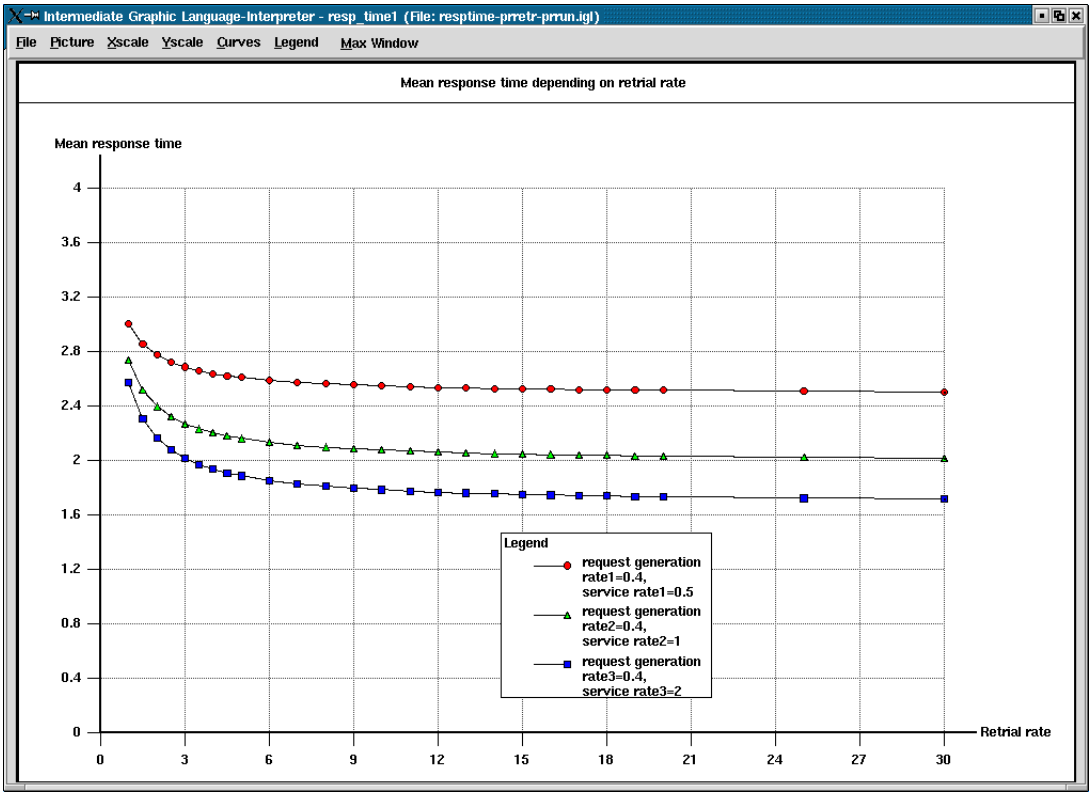

Figure 4: $E[T]$ versus retrial rate with homogeneous primary request generation and heterogeneous service

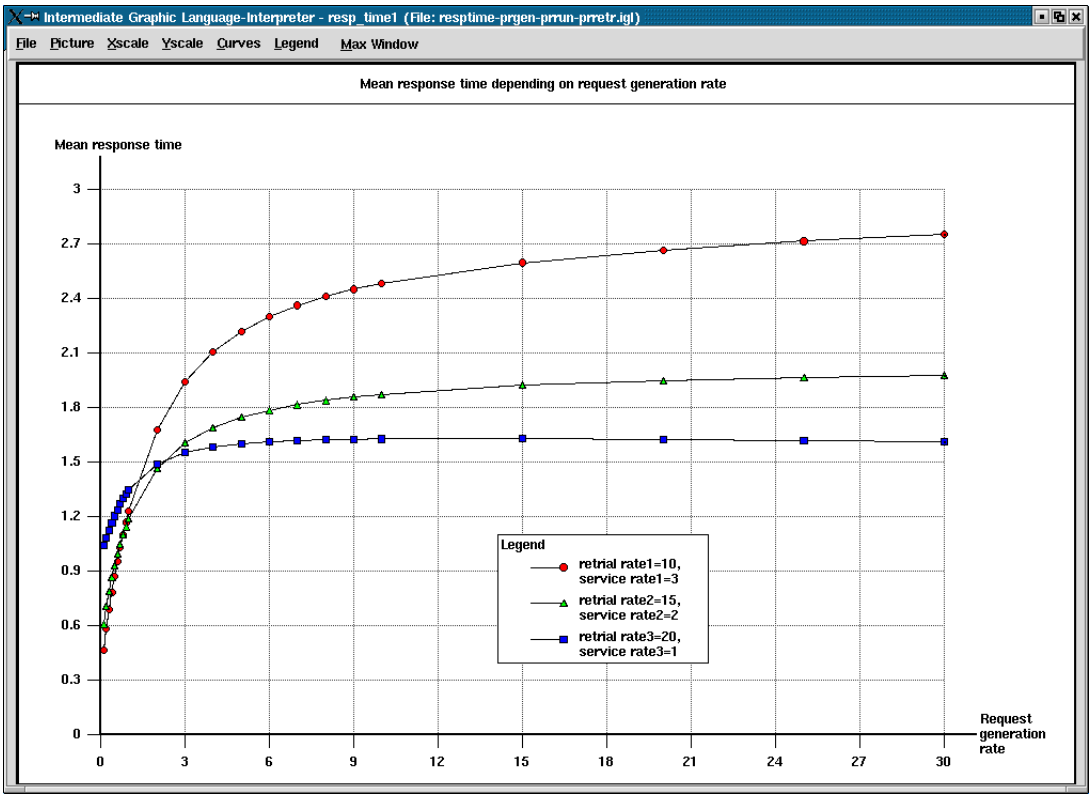

Figure 5: $E[T]$ versus primary request generation rate with heterogeneous service and heterogeneous retrial

References

[1] **Almási B., Bolch G., Sztrik J.** Performability Modeling of Non-homogeneous Terminal Systems Using MOSEL, *5th International Workshop on Performnability Modeling of Computer and Communication Systems, Erlangen, Germany, 2001* 37-41.

[2] **Artalejo J.R.** Retrial queues with a finite number of sources, *J. Korean Math. Soc.* 35(1998) 503-525.

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

[4] **Begain K., Bolch G., Herold H.** *Practical performance modeling, application of the MOSEL language,* Kluwer Academic Publisher, Boston, 2001.

[5] **Falin G.I. and Templeton J.G.C.** *Retrial queues,* Chapman and Hall, London, 1997.

[6] **Falin G.I. and Artalejo J.R.** A finite source retrial queue, *European Journal of Operational Research* 108(1998) 409-424.

[7] **Falin G.I.** A multiserver retrial queue with a finite number of sources of primary calls, *Mathematical and Computer Modelling* 30(1999) 33-49.

[8] **Houck D.J., Lai W.S.** Traffic modelling and analysis of hybrid fibercoax systems, *Computer Networks and ISDN Systems* 30(1998) 821-834.

[9] **Janssens G.K.** The quasi-random input queueing system with repeated attempts as a model for collision-avoidance star local area network, *IEEE Transactions on Communications* 45(1997) 360-364.

[10] **Kalmychkov A.I. and Medvedev G.A.** Probability characteristics of Markov local-area networks with random-access protocols, *Automatic Control and Computer Science* 24(1990) 38-45.

[11] **Khomichkov I.I.** Study of models of local networks with multiple-access protocols, *Automation and Remote Control* 54(1993) 1801-1811.

[12] **Kornyshev Y.N.** Design of a fully accessible switching system with repeated calls, *Telecommunications* 23(1969) 46-52.

[13] **Ohmura H. and Takahashi Y.** An analysis of repeated call model with a finite number of sources, *Electronics and Communications in Japan* 68(1985) 112-121.

[14] **Sztrik J.** A probability model for priority processor-shared multiprogrammed computer systems, *Acta Cybernetica* 7(1986) 329-340.

[15] **Takagi H.** *Queueing Analysis, A Foundation of Performance Evaluation, Vol. 2., Finite Systems,* North-Holland, Amsterdam, 1993.

[16] **Tran-Gia P. and Mangjes M.** Modeling of customer retrial phenomenon in cellural mobile networks, *IEEE Journal of Selected Areas in Communications* 15(1997) 1406-1414.