The Extended Connection-Dependent Threshold Model for Elastic and Adaptive Traffic

Vassilios G. Vassilakis, Ioannis D. Moscholios and Michael D. Logothetis
WCL, Dept. of Electrical & Computer Engineering, University of Patras, 265 00 Patras, Greece
E-mail: {vasilak, moscholios, m-logo}@wcl.ee.upatras.gr

Abstract-**In this paper we present a new model for the calculation of link occupancy distribution and blocking probabilities in a single-link multi-rate loss system. In this model *K* different service-classes compete for the available link bandwidth under the *complete sharing* policy. A service-class can be either of *elastic* or of *adaptive* type. Calls arrive to the link with several contingency bandwidth requirements, which depend on thresholds that indicate the occupied link bandwidth. It is possible for a call, while in service with a certain bandwidth, to reduce it, in order to allow the acceptance of new calls in the link. The accuracy of the new model is verified by simulation results.**

Keywords: Loss System; Call Blocking Probability; Elastic Traffic; Adaptive Traffic; Threshold Model; Recurrent Formula.

I. INTRODUCTION

The call-level quality of service (QoS) assessment in the multi-service environment of modern communications networks remains an open issue, due to the presence of *elastic* traffic of many applications. *Elastic* service-classes are able to reduce the required bandwidth per call, while increasing simultaneously the service time (e.g file transfer). One variation of *elastic* traffic is the *adaptive* traffic, and comes from applications which have a bandwidth requirement that can be reduced in some extent, but the service time is not altered (e.g. *adaptive* audio or video). Calls of fixed bandwidth and service time requirements compose the so-called *stream* traffic.

Despite the fundamental difference between the *elastic*/*adaptive* traffic and the *stream* traffic, the springboard of call-level modelling in both cases is a *stream* traffic model: the Erlang Multirate Loss Model (EMLM) [1], [2], where calls of *stream* service-classes arrive to a single link of certain capacity, according to a Poisson process, and compete for the available bandwidth under the complete sharing (CS) policy – i.e. calls do not have contiguous bandwidth requirements. The calls' service time can be arbitrarily distributed [1]. As far as the equilibrium state probabilities are concerned, the EMLM has a product form solution (PFS). This fact leads to an accurate calculation of the link occupancy distribution and, consequently, of Call Blocking Probabilities (CBP). Moreover, these calculations are recurrent, a feature of the EMLM that is most desirable, because it readily broadens the EMLM's applicability range to links of large capacities. Because of the excellent analysis of the EMLM, important extensions are found in the literature, extensions which cover the analysis of *elastic* traffic.

In [3] the retry models are proposed, in which a blocked call is allowed to retry once or more, requesting for less bandwidth each time. In the threshold models of [4], a call arrives to the link with several contingency bandwidth requirements and requests for bandwidth according to the total link occupied bandwidth. In these models the occupied bandwidth and the corresponding required service time are related is such a way that the *total offered traffic load* [5] remains constant; in this sense we talk of *elastic* traffic. The Connection-Dependent Threshold Model (CDTM) [6] generalizes the above mentioned retry and threshold models (as well as the EMLM) by individualizing the thresholds among the service-classes.

In another extension of the EMLM [7], when the link capacity is exceeded, weighted bandwidth reductions of the in-service calls are applied in order for a new call to be accepted in the link. A comparative evaluation of this model to the CDTM has been presented in [8]. It is shown that if only *elastic* service-classes are present in the link, both models perform equally well. In [9], a further extension of the model of [7] has been proposed (we call it *Extended EMLM* (*E-EMLM*)) that takes into account *adaptive* traffic as well. Calls arrive to the link with a peak bandwidth and service time requirement. When the peak bandwidth is not available, it is possible for a call to be accepted to the link with reduced bandwidth, by reducing accordingly the allocated bandwidth of the in-service calls. Simultaneously, the service time of *elastic* in-service calls is increased, while the service time of *adaptive* in-service calls remains unchanged. A comparative evaluation of this model to the CDTM has been shown in [10]. This work shows that if both *elastic* and *adaptive* service-classes are present in the link, the *E-EMLM* performs better than the CDTM with regard to CBP.

In this paper, we propose a new model named *Extended CDTM* (*E-CDTM*), based on [6], [7] and [9]. In the *E-CDTM*, calls of different service-classes, either *elastic* or *adaptive*, arrive to a link of limited bandwidth capacity. Each service-class has an associated set of thresholds, which indicate the occupied link bandwidth. Calls of a particular service-class, upon arrival, request for bandwidth according to the occupied link bandwidth at that moment and to the thresholds of the service-class that they belong to – i.e. calls of each service-class have several predefined contingency bandwidth requirements associated with thresholds. Each bandwidth requirement has a corresponding service time requirement; they are related in such a way that the *total offered traffic load* is the same for every pair of bandwidth and service time. As long as there is enough available bandwidth in the link, each call can be accepted with one of its several contingency bandwidth requirements. If, upon arrival of a new call, its requested bandwidth is not available in the link, this call can still be accepted but with reduced bandwidth. This is done by reducing the bandwidth of all

in-service calls. The bandwidth allocated to a call, however, cannot drop below a predefined percentage of the requested bandwidth. When the allocated bandwidth of an *elastic* call drops below the requested bandwidth, its service time is increased proportionally to the bandwidth reduction. When the allocated bandwidth of an *adaptive* call drops below the requested bandwidth, its service time is not altered. We compare the analytical CBP results of the *E-CDTM* to the simulation results in order to show its validity.

This paper is organized as follows: In Section II we present the new model and provide the recurrent formula for the calculation of link occupancy distribution and consequently of the CBP. In Section III we present an application example, where the new model is evaluated through simulation results. We conclude in Section IV.

II. The Extended Connection-Dependent Threshold Model

A. Description

We assume K service-classes accommodated to a transmission link of capacity C bandwidth units (b.u). A service-class can be either of *elastic* or of *adaptive* type. We denote by K_e the number of *elastic* service-classes and by K_a the number of *adaptive* service-classes, that is, $K=K_e+K_a$. Calls of service-class k arrive to the link according to a Poisson process with rate λ_k (k = 1,2,…,K). Each service-class k has an associated set of thresholds: $J_{k_1}, J_{k_2},\ldots, J_{k_{S_k}}$, where S_k is the number of thresholds of service-class k. The thresholds, which indicate the occupied link bandwidth, are used to determine the bandwidth that will be requested from a newly arriving call. Calls of a particular service-class, upon arrival, request for bandwidth according to the system state j (which is defined below) at that moment and the thresholds of the service-class that they belong to; they have S_k+1 predefined contingency bandwidth requirements $b_{k_0} > b_{k_1} > \ldots > b_{k_{S_k}}$. Each bandwidth requirement b_{k_l} (l=0,1,…,S_k) corresponds to a service time requirement which is exponentially distributed with mean $\mu_{k_l}^{-1}$. The product of *(service time)* by *(bandwidth per call)* remains constant for every pair $(b_{k_l}, \mu_{k_l}^{-1})$. As long as there is enough available bandwidth in the link, each call is accepted with one of its several contingency bandwidth requirements b_{k_l}. If, upon arrival of a new call, its requested bandwidth b_{k_l} is not available in the link, this call can still be accepted but with reduced bandwidth, while reducing the bandwidth allocated to the in-service calls.

The bandwidth reduction of service-class k calls is allowed up to a minimum bandwidth $b_k^{\min} = r_k^{\min} b_{k_l}$, where $r_k^{\min}$ is the minimum proportion of the requested bandwidth b_{k_l} that can be allocated to service-class k calls. In this paper we consider, for all service-classes, a common $r_k^{\min} = r^{\min}$, which constitutes a system parameter. The maximum bandwidth demand of all calls is defined as $T = C/r^{\min}$ [9].

We denote by s ($0 \le s \le C$) the total link occupied bandwidth. We denote by j ($0 \le j \le T$) the system state, which is the bandwidth that would be occupied if all in-service calls were granted the initially requested bandwidth, whereas it represents the total bandwidth demand of the system.

The percentage of bandwidth reduction of each call in a state j is determined by: $r(j) = \min(1, C/j)$ – i.e the actual bandwidth occupied from a call is $r(j)\, b_{k_l}$ .Note that for states $j \le C$ we have $r(j)=1$, so no bandwidth reduction occurs, while for $j=T$ we have $r(T)= C/T = r^{\min}$.

For states $C < j \le T$, there is a proportional bandwidth reduction $r(j) = C/j$ for all in-service calls. The reduction of the service rate is the same for *elastic* calls and it becomes $r(j)\,\mu_{k_l}$, while the service rate of *adaptive* calls remains unchanged and always equal to μ_{k_l}.

As an example of the above, consider the system of Fig. 1-2 with two service-classes. At a given time, with the system being in state, say j_a, a new call of service-class k=1 arrives to the link. Assume that at this point $J_{1_l} < j_a \le J_{1_{l+1}}$, which means that the new call requests for b_{1_l} b.u.; if $j_a + b_{1_l} \le C$ the call is accepted in the link with parameters $(b_{1_l}, \mu_{1_l}^{-1})$. Thus the system will pass to the state $j_b = j_a + b_{1_l}$ (see Fig. 1). Afterwards, a new call of service-class k=2 arrives to the link and if $J_{2_{l'}} < j_b \le J_{2_{l'+1}}$ (generally $l \ne l'$), this call will request for $b_{2_{l'}}$ b.u. Now, assume that $C < j_b + b_{2_{l'}} \le T$ and $j_b + b_{2_{l'}} = j_c$ In this case, the new call is accepted in the link with reduced bandwidth, less than the requested bandwidth $b_{2_{l'}}$. The bandwidth allocated to this call will be $r(j_c)\, b_{2_{l'}}$ (where $r(j_c) = C/j_c < 1$), while the call service time will be $\mu_{1_{l'}}^{-1} / r(j_c)$ (increased). At the same time the bandwidth allocated to all in-service calls is also reduced by the same factor $r(j_c)$. After the call acceptance the system will pass to the state $j_c = j_b + b_{2_{l'}}$ (see Fig. 2).

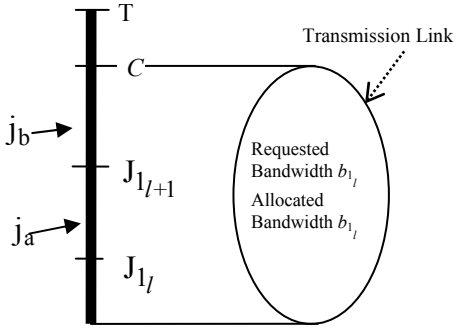

Fig. 1. System state with no bandwidth reduction for new arrival.

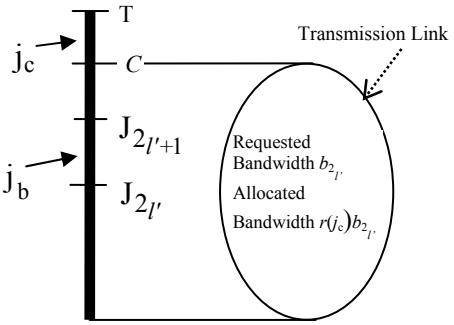

Fig. 2. System state with bandwidth reduction for new arrival.

B. Assumptions

In our model we assume the following:

1) *Local balance* between adjacent system states.

We denote by $a_{k_l} = \lambda_k / \mu_{k_l}$ the offered traffic load of calls with bandwidth requirement b_{k_l}, l=0,…, S_k. We denote by $Y_{k_l}(j)$ the average number of calls with bandwidth requirement b_{k_l} in state j ($0 \le j \le T$).

We assume the following local balance equations for *elastic* traffic:

i) Calls accepted with their maximum contingency bandwidth requirement b_{k_0}.

$$q(j-b_{k_0})\,\lambda_k = q(j)\,Y_{k_0}(j)\,\mu_{k_0} r(j), \quad j \le J_{k_0} + b_{k_0} \tag{1a}$$

ii) Calls accepted with other than the maximum bandwidth requirement b_{k_m}, m=1,…, S_k,.

$$q(j-b_{k_m})\,\lambda_k = q(j)\,Y_{k_m}(j)\,\mu_{k_m} r(j), \quad J_{k_{m-1}}+b_{k_{m-1}} < j \le J_{k_m}+b_{k_m} \tag{1b}$$

We assume the following local balance equations for *adaptive* traffic:

i) Calls accepted with their maximum contingency bandwidth requirement b_{k_0}.

$$q(j-b_{k_0})\,\lambda_k = q(j)\,Y_{k_0}(j)\,\mu_{k_0}, \quad j \le J_{k_0} + b_{k_0} \tag{2a}$$

ii) Calls accepted with other than the maximum bandwidth requirement b_{k_m}, m=1,…, S_k,.

$$q(j-b_{k_m})\,\lambda_k = q(j)\,Y_{k_m}(j)\,\mu_{k_m}, \quad J_{k_{m-1}}+b_{k_{m-1}} < j \le J_{k_m}+b_{k_m} \tag{2b}$$

2) *Migration approximation*: calls accepted in the system with other than the maximum bandwidth requirement are negligible within a space, called *migration space*. More precisely, the mean number of calls, $Y_{k_l}(j)$, with bandwidth b_{k_l} is negligible when $0 < j < J_{k_m} + b_{k_m}$ and $J_{k_{m+1}} + b_{k_m} \le j \le T$; the latter region is related to the variable $\delta_{k_m}(j)$, m=1,…, S_k.

3) *Upward approximation:* calls accepted in the system with their maximum bandwidth requirement are negligible within a space, called *upward space*. More precisely, the mean number of calls, $Y_{k_0}(j)$, with bandwidth b_{k_0} is negligible when $J_1 + b_{k_0} < j \le T$; the latter region is related to the variable $\delta_{k_0}(j)$.

By using the parameters $\delta_{k_l}(j)$ (l=0,…, S_k), eq. (1a) and (1b) are combined into (3), which is valid for *elastic* service-calsses:

$$q(j-b_{k_l})\,\delta_{k_l}(j)\,\lambda_k = q(j)\,Y_{k_l}(j)\,\mu_{k_l} r(j), \quad 0 \le j \le T \tag{3}$$

By using the parameters $\delta_{k_l}(j)$ (l=0,…, S_k), eq. (2a) and (2b) are combined into (4), which is valid for *adaptive* service-classes:

$$q(j-b_{k_l})\,\delta_{k_l}(j)\,\lambda_k = q(j)\,Y_{k_l}(j)\,\mu_{k_l}, \quad 0 \le j \le T \tag{4}$$

Similarly to [9] we combine (3) and (4) into the following local balance equation for both *elastic* and *adaptive* service-classes:

$$q(j-b_{k_l})\,\delta_{k_l}(j)\,\lambda_k = q(j)\,Y_{k_l}(j)\,\mu_{k_l}\,\Phi_{k_l}(j), \quad 0 \le j \le T \tag{5}$$

where the factor $\Phi_{k_l}(j)$ is the ratio of state-dependent weights $x(j)$:

$$\Phi_{k_l}(j) = x(j - b_{k_l}) / x(j) \tag{6}$$

The weights $x(j)$ denote the variation of the transition rates from one state to the other due to the bandwidth reduction, and are defined in (9), below.

C. System occupancy distribution – CBP Calculation

a) When $j \le C$ the total allocated bandwidth s is equal to the total bandwidth demand j.

$$j = \left\langle \sum_{k\in Ke}\sum_{l=0}^{S_k} Y_{k_l}(j) b_{k_l} \Phi_{k_l}(j) + \sum_{k\in Ka}\sum_{l=0}^{S_k} Y_{k_l}(j) b_{k_l} \Phi_{k_l}(j) \right\rangle \tag{7}$$

b) When $C < j \le T$ the total allocated bandwidth s is equal to the system capacity C.

$$C = \left\langle \sum_{k\in Ke}\sum_{l=0}^{S_k} Y_{k_l}(j) b_{k_l} \Phi_{k_l}(j) + r(j) \sum_{k\in Ka}\sum_{l=0}^{S_k} Y_{k_l}(j) b_{k_l} \Phi_{k_l}(j) \right\rangle \tag{8}$$

Substituting (6) into (7) and (8) results in (9), which calculates the weights $x(j)$ recurrently.

$$x(j) = \left\langle \frac{\sum_{k\in Ke}\sum_{l=0}^{S_k} Y_{k_l}(j) b_{k_l} x(j-b_{k_l})}{min(C,j)} + \frac{r(j)\sum_{k\in Ka}\sum_{l=0}^{S_k} Y_{k_l}(j) b_{k_l} x(j-b_{k_l})}{min(C,j)} \right\rangle \tag{9}$$

The following recurrent formula calculates the system occupancy distribution $q(j)$, $0 \le j \le T$:

$$min(C,j)\,q(j) = \left\langle \sum_{k\in K_e}\sum_{l=0}^{S_k} \alpha_{k_l} b_{k_l} \delta_{k_l}(j) q(j-b_{k_l}) + r(j) \sum_{k\in K_a}\sum_{l=0}^{S_k} \alpha_{k_l} b_{k_l} \delta_{k_l}(j) q(j-b_{k_l}) \right\rangle \tag{10}$$

$q(x)$=0 for $x < 0$ and $\sum_{j=0}^{T} q(j) = 1$.

The CBP of service-class k is given by:

$$B_k = \sum_{j=C-b_{k_{S_k}}+1}^{T} q(j) \tag{11}$$

III. EVALUATION

We evaluate the Extended CDTM through simulation by considering a link of C=80 b.u. and maximum bandwidth demand T=85 b.u. or T=90. One *elastic* service-class, s_1, and one *adaptive* service-class, s_2, are accommodated to the link with a maximum bandwidth per call requirement of 24 and 8 b.u., and offered traffic-load of α_1 = 0.25 erl and α_2 = 0.5 erl, respectively. Calls of *elastic* service-class are able to reduce their bandwidth from 24 to 16 b.u. in steps of four and calls of *adaptive* service-class from 8 to 6 b.u. unit by unit. More precisely, the thresholds of s_1 are: J_{1_0}=40, and J_{1_1}=55, whereas of s_2 are: J_{2_0}=35 and J_{2_1}=45. That is, when the system state, j, is less than or equal to 40 b.u. a call of service-class s_1 is accepted in the link with its maximum bandwidth requirement of 24 b.u. If 40< j≤55 then a call of service-class s_1 is accepted in the link with 20 b.u., while if 55<j≤80-16=64 a call of service-class s_1 is accepted in the link with 16 b.u. A call of service-class s_2 is accepted in the link with 8 b.u. when j≤35. If 35<j≤45 then a call of service-class s_1 is accepted in the link with 7 b.u., while if 45<j≤80-6=74 a call of service-class s_1 is accepted in the link with 6 b.u. When C<j≤T, calls of s_1 and s_2 seize $b_{k_1}r(j)$ and $b_{k_1}r(j)$ b.u. respectively. We consider eight traffic-load points (1, 2,…, 8) in the x-axis of Fig. 3-4, by increasing simultaneously a_1 and a_2 by 0.25 erl and 0.5 erl, respectively, so that the heaviest traffic-load (8) is a_1 = 2.0 erl and a_2 = 4.0 erl for s_1 and s_2 , respectively.

In Fig. 3-4 we present analytical and simulation CBP results of the model versus the offered traffic-load for T=85 and T=90. The simulation CBP results have been obtained as mean values from 8 runs. For each mean value, a confidence interval of 95% has been defined. However, the resultant reliability ranges of our measurements are small enough and therefore we present only the mean CBP values in the figures. As it was anticipated the resultant CBP are lower for T=90 than for T=85. The results show that the model's accuracy is absolutely satisfactory, especially for low offered-traffic load (points 1, 2, 3 in the x-axis of Fig. 3-4).

IV. CONCLUSION

We present a new model for the analysis of a single-link multi-rate loss system, where two types of traffic, *elastic* and *adaptive*, are present that may have several contingency bandwidth requirements according to thresholds. We provide a recurrent formula for the calculation of the link occupancy distribution and CBP. Simulations are used to verify the accuracy of the proposed model. We show by numerical examples that the accuracy of the new model is absolutely satisfactory.

ACKNOWLEDGMENT

Work supported by the research program Caratheodory of the Research Committee of the University of Patras.

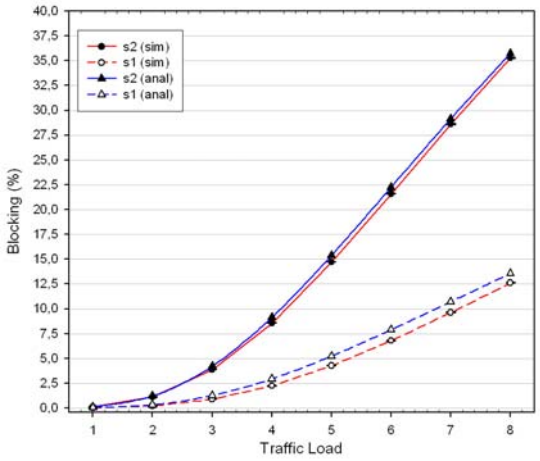

Fig. 3. CBP results versus the offered traffic-load (T=85).

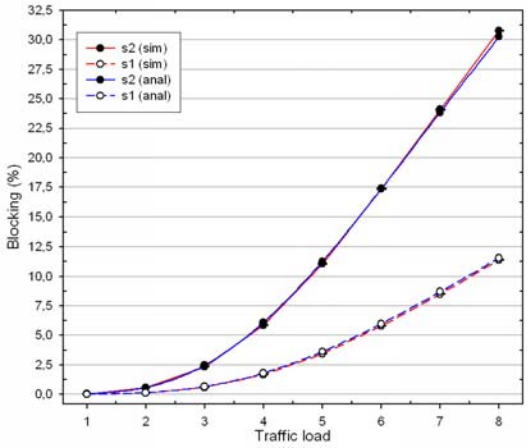

Fig. 4. CBP results versus the offered traffic-load (T=90).

REFERENCES

[1] J.S. Kaufman, *"Blocking in a shared resource environment"*, IEEE Transactions on Communications, vol.29 (10) (1981), pp.1474-1481.

[2] J.W. Roberts, *"A service system with heterogeneous user requirements"*, in: G. Pujolle (Ed.), Performance of Data Communications systems and their applications, North Holland, Amsterdam, 1981, pp. 423-431.

[3] J.S. Kaufman, *"Blocking in a completely shared resource environment with state dependent resource and residency requirements"*, in Proceedings of the IEEE Conference, INFOCOM'92, 1992, pp. 2224-2232.

[4] J.S. Kaufman, *"Blocking with retrials in a completely shared resource environment"*, Performance Evaluation, 15 (1992), pp. 99-113.

[5] H. Akimaru, K. Kawashima, *"Teletraffic – Theory and Applications"*, Springer-Verlag, 1993.

[6] I. Moscholios, M. Logothetis, G. Kokkinakis *"Connection Dependent Threshold Model: A Generalization of the Erlang Multiple Rate Loss Model"*, Performance Evaluation, vol. 48, issue 1-4, pp. 177-200, May 2002.

[7] G. Stamatelos, V. Koukoulidis, *"Reservation-based bandwidth allocation in a radio ATM network"*, IEEE/ACM Trans. Networking 5 (3) (1997), pp. 420–428.

[8] V. Vassilakis, I. Moscholios, M. Logothetis, *"Evaluation of Multi-rate Loss Models for Elastic Traffic"*, 3rd International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs'05), Ilkley, West Yorkshire, U.K., July 18-20, 2005.

[9] S. Racz, B. P. Gero, G. Fodor, *"Flow Level Performance Analysis of a Multi-service System Supporting Elastic and Adaptive Services"*, Performance Evaluation, vol. 49, no. 1/4, pp. 451-469, 2002.

[10] V. Vassilakis, I. Moscholios, M. Logothetis, *"Evaluation of Multi-rate Loss Models for Elastic and Adaptive Services"*, 12th Polish Teletraffic Symposium (PSRT '05), Poznan, Poland, September 19-20, 2005.