means that if the available link bandwidth is not enough to accept the whole batch (group of calls), then a part of it is accepted and the rest of it is discarded. The batch Poisson arrival process with the partial batch blocking discipline has been studied in the case of the EMLM. This process is important, not only because in several applications calls arrive as batches, but also because it can represent quasi-random arrivals and overflow traffic [7-10]. In [7], the batch size distribution is restricted to be geometric, while in [8] the batch size distribution is general. In both, the model has a PFS and the link occupancy distribution is calculated by accurate recursive formulas. In [9] an asymptotic analysis of the model of [8] is considered, while in [10] the model is extended to cover multiple links, state dependent batch arrivals, both complete and partial batch blocking disciplines as well as other policies apart from the *complete sharing* policy.

In the proposed BP-ON-OFF model, we focus on the call-level analysis (not burst-level) following the methodology of [8]. We prove that the BP-ON-OFF model has a PFS and show that the link occupancy distribution can be calculated by an accurate two-dimensional recursive formula. Then, we provide analytical formulas for various performance measures such as call-level blocking and link utilization. Regarding call-level blocking, it is important to note the distinction between time and call congestion probabilities. These probabilities coincide in the case of Poisson arrivals (Poisson Arrivals See Time Averages, PASTA property [11]), but not in the case of batch Poisson arrivals. Time congestion probability is determined by the proportion of time that the system is congested. An observer, who is not part of the system, can measure this probability. Call congestion probability is determined by the proportion of arriving calls that find the system congested. An observer who is part of the system (i.e. an arriving call) can measure this probability.

Section 2 presents the BP-ON-OFF model; in subsection 2.1 the service system is described and in subsection 2.2 the system is analysed and the formulas for the call-level performance measures are derived. Section 3 is the numerical/application section; analytical results are compared to simulation results in order to validate the accuracy of the proposed formulas. We conclude in section 4.

2. THE PROPOSED BP-ON-OFF MODEL

2.1. The service system

We consider a link with a pair of capacities C and C^*, accommodating K independent services of ON-OFF-type calls under the *complete sharing* policy. The first capacity, named real (real link), corresponds to state ON while the second one, named fictitious (fictitious link), corresponds to state OFF. Calls of service k (k=1,…,K) require b_k bandwidth units (b.u.) and arrive to the link according to a batch Poisson process, with an arrival rate λ_k, an exponentially distributed holding time and a batch size distribution B_r^k (r=1,2,…) where B_r^k is the probability that an arriving batch of service k consists of r calls. Note that if B_r^k =1 for r=1 and B_r^k =0 for r =2, 3, …, then the arrival process is Poisson and the resultant model is the ON-OFF model of [3]. Regarding the partial batch blocking, if l out of r calls of an arriving batch can be accepted then the other r-l calls will be blocked and lost without further affecting the system. The l accepted calls of an arriving batch enter the system via state ON only. At the end of an ON-period, there are two possibilities:

(i) With probability σ_k, the l accepted calls release the $l*b_k$ b.u. that hold and are transferred to state OFF, in which they seize fictitious bandwidth ($l*b_k$ b.u.) of the fictitious capacity C^*,
(ii) With probability 1- σ_k, one out of the l accepted calls departs from the system.

On the occurrence of the first possibility and at the end of an OFF period each call of service k returns to state ON with probability 1 (a call cannot depart from the system from state OFF). To return to state ON, each call requires again b_k b.u. If C=C^*, there is always available bandwidth for that call in state ON, i.e. no burst blocking occurs. If C<C^*, then if there is available bandwidth in the real link, i.e. if $j_1 + b_k \leq C$, (where j_1 is the occupied real link bandwidth), the call returns to state ON and a new burst begins, otherwise the burst is blocked and the call remains in state OFF for another OFF-period.

A new call of service k is accepted in the system with b_k b.u. if it meets the following constraints:

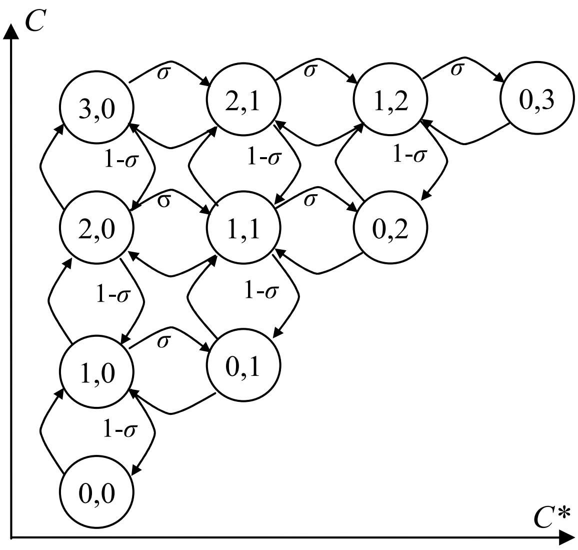

Fig. 1a: State transition diagram of ON-OFF model.

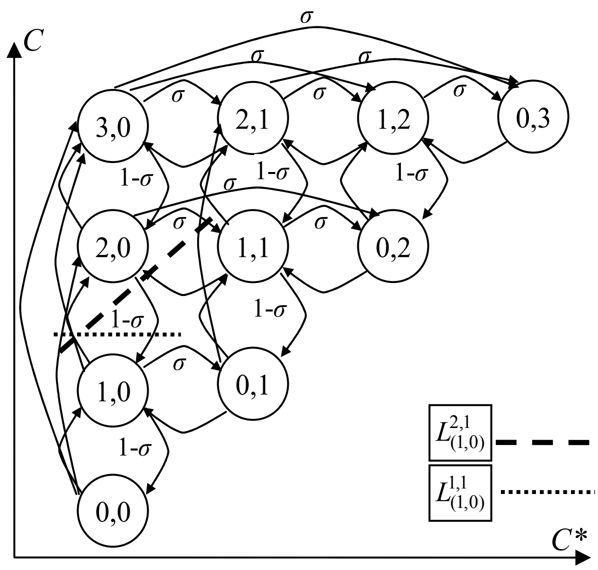

Fig. 1b: State transition diagram of BP-ON-OFF model.

$\hat{B}_l^{(k)} = \sum_{r=l+1}^{\infty} B_r^{(k)}$ we can write eq.(3) as follows:

$$f^{(up)}(L^{1,k}_{\vec{n}}) = \sum_{l=0}^{n_k^1} P(\vec{n}^{\,1}_{k-l})\lambda_k \hat{B}_l^{(k)} \tag{4}$$

The downward probability flow across the level $L^{1,k}_{\vec{n}}$ (due to departures from the system) is:

$$f^{(down)}(L^{1,k}_{\vec{n}}) = \mu_{1k}(n_k^1+1)(1-\sigma_k)P(\vec{n}^{\,1}_{k+1}) \tag{5}$$

where μ_{ik}^{-1} is the mean service time of a service k call in state i, exponentially distributed.

Similarly the upward probability flow across the level $L^{2,k}_{\vec{n}}$ (due to transitions from state ON to state OFF) is given by:

$$f^{(up)}(L^{2,k}_{\vec{n}}) = \sum_{l=0}^{n_k^2} P(\vec{n}^{\,2}_{k-l})\lambda_k \sigma_k \hat{B}_l^{(k)} \tag{6}$$

while the downward probability flow across the level $L^{2,k}_{\vec{n}}$ (transitions from OFF to ON) is:

$$f^{(down)}(L^{2,k}_{\vec{n}}) = \mu_{2k}(n_k^2+1)(1-\sigma_k)P(\vec{n}^{\,2}_{k+1}) \tag{7}$$

Note: For boundary states $\vec{n}$ (e.g. state (0,3) in Fig. 1a), both upward and downward probability flows across the levels $L^{1,k}_{\vec{n}}$, $L^{2,k}_{\vec{n}}$ are zero.

By equating eq.(4) to eq.(5) and eq.(6) to eq.(7) the following local flow balance equations result:

Having found a PFS for $P(\vec{n})$ we continue by deriving a two dimensional recursive formula for the calculation of the distribution of $\vec{j}=(j_1, j_2)$, denoted as $G(\vec{j})$. Such a recursive formula is of great merit in calculating the call-level performance measures.

Theorem

The distribution of $\vec{j}=(j_1, j_2)$, $G(\vec{j})$, is calculated by the following recursive formula:

$$G(\vec{j}) = \frac{1}{j_s}\sum_{i=1}^{2}\sum_{k=1}^{K} p_{i,k} b_{i,k,s} \sum_{l=1}^{\lfloor j_s / b_k \rfloor} \hat{B}_{l-1}^{k} G(\vec{j} - lB_{i,k}) \quad \text{with } \vec{j} \in \Omega \Leftrightarrow \{(j_1 + j_2 \le C)\} \tag{13}$$

where: s refers to the links (i.e. $s = 1 \Rightarrow$ real link, $s = 2 \Rightarrow$ fictitious link), $b_{i,k,s} = \begin{cases} b_k & if\ \ s = i \\ 0 & if\ \ s \neq i \end{cases}$,

B is a $(2K \times 2)$ matrix with entries $b_{i,k,s}$. The $(i, k)^{\text{th}}$ row of B is denoted by $B_{i,k} = (b_{i,k,1}\ b_{i,k,2})$;

e.g. for K=2 services, with bandwidth requirement per call b_1, b_2, $B = \begin{bmatrix} b_{1,1,1} & b_{1,1,2} \\ b_{2,1,1} & b_{2,1,2} \\ b_{1,2,1} & b_{1,2,2} \\ b_{2,2,1} & b_{2,2,2} \end{bmatrix}$ and

$\lfloor y \rfloor$ denotes the largest integer less than or equal to y.

Proof

Based on eq.(12) we can write $n_k^i P(\vec{n})$, for $n_k^i \ge 1$ as follows:

$$n_k^i P(\vec{n}) = \sum_{l=1}^{n_k^i} p_{i,k} P(\vec{n}^{\,i}_{k-l}) \hat{B}_{l-1}^{(k)} \tag{14}$$

We multiply both sides of eq.(14) by $b_{i,k,s}$ and sum over i and k:

$$P(\vec{n}) \sum_{i=1}^{2}\sum_{k=1}^{K} n_k^i b_{i,k,s} = \sum_{i=1}^{2}\sum_{k=1}^{K}\sum_{l=1}^{n_k^i} p_{i,k} b_{i,k,s} P(\vec{n}^{\,i}_{k-l}) \hat{B}_{l-1}^{(k)} \tag{15}$$

The occupied bandwidth of link s, denoted as j_s, can be expressed as follows:

$$j_s = \sum_{i=1}^{2}\sum_{k=1}^{K} n_k^i b_{i,k,s} \tag{16}$$

and therefore eq.(15) can be written as follows:

$$P(\vec{n}) j_s = \sum_{i=1}^{2}\sum_{k=1}^{K}\sum_{l=1}^{n_k^i} p_{i,k} b_{i,k,s} P(\vec{n}^{\,i}_{k-l}) \hat{B}_{l-1}^{(k)} \tag{17}$$

We consider now the set of states: $\Omega_{\vec{j}} = \left\{ \vec{n} \in \Omega_{\vec{j}} : \vec{n}B = \vec{j},\ n_k^i \ge 0,\ k = 1,...,K,\ i = 1,2 \right\}$. Summing eq.(17) over $\Omega_{\vec{j}}$ and interchanging the order of summation we have:

Substituting the new left hand side to eq.(22) we have:

$$E(n_k^i \left| \vec{j} \right.) = \frac{p_{i,k} \sum_{l=1}^{\lfloor j_s/b_k \rfloor} \hat{B}_{l-1}^k G(\vec{j} - lB_{i,k})}{G(\vec{j})} \tag{24}$$

(b) The average number of service k calls in state ON (i=1) or state OFF (i=2), denoted as $\overline{n}_k^i$, can be calculated by:

$$\overline{n}_k^i = \sum_{\vec{j} \in \Omega} E(n_k^i \left| \vec{j} \right.) G(\vec{j}) = \sum_{\vec{j} \in \Omega} p_{i,k} \sum_{l=1}^{\lfloor j_s/b_k \rfloor} \hat{B}_{l-1}^k G(\vec{j} - lB_{i,k}) \tag{25}$$

(c) Call congestion probability of service k, denoted as C_{b_k}, which is the probability that an arriving call of service k cannot be accepted in the system, is calculated (in the case of $C=C^*$) by the formula:

$$C_{b_k} = \frac{(p_{1k} + p_{2k})\hat{B}_k - (\overline{n}_k^1 + \overline{n}_k^2)}{(p_{1k} + p_{2k})\hat{B}_k} \tag{26}$$

where: $\hat{B}_k$ denotes the average size of arriving batches, given by: $\hat{B}_k = \sum_{r=1}^{\infty} r B_r^k$.

(d) Time congestion probability of service k, denoted as P_{b_k}, is the probability that the real and fictitious occupied link bandwidth ($j_1 + j_2$) is at least $C - b_k + 1$ and is given directly through eq.(13):

$$P_{b_k} = \sum_{\{\vec{j} \mid j_1 + j_2 + b_k > C\}} G^{-1} G(\vec{j}) \tag{27}$$

where $G = \sum_{\vec{j} \in \Omega} G(\vec{j})$ is the normalization constant.

(e) In order to calculate the utilization of the real (s=1) or the fictitious link (s=2), the following formula can be used:

$$\overline{G}_s = \sum_{\tau=0}^{C_s} \tau G_s(\tau) \tag{28}$$

where $C_s = C$ if s =1 and $C_s = C^*$ if s =2, while $G_s(\tau)$ is the marginal link occupancy distribution of link s given by:

$$G_s(\tau) = \sum_{\{\vec{j} \mid j_s = \tau\}} G(\vec{j}) \tag{29}$$

3. NUMERICAL EXAMPLES

Through an application example, we compare the analytical results, obtained by the proposed BP-ON-OFF model, with simulation results. The latter are mean values of 7 runs with 95% confidence

4. CONCLUSION

We propose a teletraffic model for bursty services of ON-OFF type accommodated in a single link of certain capacity. Calls of each service arrive at the link according to a batch Poisson arrival process and compete for the available bandwidth under the complete sharing police. This arrival process is important, not only because calls arrive as batches in several applications, but also because it can represent quasi-random arrivals and overflow traffic. The batch size can be generally distributed while the batch blocking discipline is the partial batch blocking, i.e. if the available bandwidth is less than that the total amount of bandwidth required by the batch, then a part of it is accepted and the rest of it is discarded. Accepted calls enter the system via state ON and can alternate between ON-OFF states. When a call is transferred to state OFF it releases its bandwidth in state ON, which becomes available to other calls (link utilization increase). After presenting the service system we prove that the proposed batched Poisson ON-OFF model has a product form solution and provide an accurate recursive formula for the link occupancy distribution calculation. Based on the link occupancy distribution we determine various call-level performance measures such as the time and call congestion probabilities and link utilization. Simulation results validate our analytical methodology.

ACKNOWLEDGEMENT

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

REFERENCES

1. J. S. Kaufman, "Blocking in a shared resource environment", IEEE Trans. Commun. COM-29 (10) (1981) 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. M. Mehmet Ali, "Call-burst blocking and call admission control in a broadband network with bursty sources", Performance Evaluation 38 (1999) 1-19.
4. R. Guerin, H. Ahmadi, M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks", IEEE J. Select. Areas Commun.-9 (7) (1991) 968-981.
5. I. Moscholios, P. Nikolaropoulos, M. Logothetis, "Call level blocking of ON-OFF traffic sources with retrials under the complete sharing policy", Proc. 18th International Teletraffic Congress (ITC), Berlin, (2003) pp. 811-820.
6. I.Moscholios, M. Logothetis and G. Kokkinakis, "Call-burst blocking of ON-OFF traffic sources with retrials under the complete sharing policy", Performance Evaluation 59/4, pp. 279-312, 2005.
7. E. A. van Doorn and F. J. M. Panken, "Blocking Probabilities in a Loss System with Arrivals in Geometrically Distributed Batches and Heterogeneous Service Requirements", ACM/IEEE Trans. Networking (1), pp. 664-667, 1993.
8. J. S. Kaufman and K. M. Rege, "Blocking in a shared resource environment with batched Poisson arrival processes", Performance Evaluation 24, pp. 249-263, 1993.
9. J.A. Morrison, "Blocking probabilities for multiple class batched Poisson arrivals to a shared resource", Performance Evaluation 25, pp. 131-150, 1996.
10. G. L. Choundhury, K. K. Leung and W. Whitt, "Resource-sharing models with state-dependent arrivals of batches", Computations with Markov Chains*,* W. J. Stewart, ed., Kluwer, Boston, pp. 255-282, 1995.
11. H. Akimaru and K. Kawashima, Teletraffic – Theory and Applications, 2nd edition, Springer-Verlag, Berlin, 1993.