METHODS FOR COMPUTING B-ISDN LINK BLOCKING PROBABILITIES

Sándor Molnár
Technical University of Budapest
Department of Telecommunications and Telematics
Sztoczek u. 2
H-1111 Budapest, Hungary
Tel: + 36 1 1813 500/24 71
Fax: + 36 1 1812 302
E-mail: molnar@bme-tel.ttt.bme.hu

Søren Blaabjerg
Ellemtel Telecommunication Systems Laboratories
Scheelevägen 17
S-223 70 Lund
Sweden
Tel: + 46 46 18 57 66
Fax: + 46 46 18 57 61
E-mail: Soren.Blaabjerg@eua.ericsson.se

Abstract

In the future B-ISDN will support a very wide variety of services. It should be suitable for both business and residential customers providing all sort of voice, data and video communications. The ATM (Asynchronous Transfer Mode) has been chosen as an attractive multiplexing and switching technique for transfer of future communication services having a broad spectrum of traffic statistics and quality of service requirements.

Modeling the traffic in B-ISDN requires more general models than the well-known traffic models applied in the traditional telephone networks. To get such appropriate models for the investigation of the call scale traffic in B-ISDN the loss system with general arrival and service processes (G/G/c) should be considered. The exact evaluation of this system is very complicated and even for some subclasses is limited to rather small system sizes.

In this paper we describe two simple approximation procedures. Both are based on matching the mean and variance of the occupancy distribution, and it is pointed out how the concept of generalized peakedness makes the evaluation of the variance very simple for a large subclass of the G/G/∞ systems. The results for the infinite server system is finally used to compute approximate blocking probabilities for the finite capacity system.

1 INTRODUCTION

The computation of blocking probabilities in telephone networks has a long tradition initiated with the pioneering work of Erlang, see [BHJ48]. The key issue of computing accurate blocking probabilities in the ATM based B-ISDN is the appropriate traffic modeling. Decades of experiences has shown that in the traditional telephone networks the Poisson process is a reasonable model of the call request process. Also the assumption of exponential holding time has been accepted as a good model of the service process.

The ATM as a high speed network transport mechanism of B-ISDN supports a very large spectrum of services. The arrival process characteristics of new connection request differs substantially from service to service and shows an inherent different nature from those in the telephone networks. The Poisson process will not be appropriate to model the arrival process in B-ISDN and it is also expected that the holding time in most cases will differ significantly from the exponential distribution. Here is a challenge arises to the traffic modeling to find accurate models for the arrival and service processes. This need actualizes the investigation of networks with general arrival and service processes (G/G/c queues) both as loss and delay systems. The teletraffic theory has accepted this challenge and so far many sophisticated models have been developed such as Markov modulated models or phase-type renewal processes and so on for cell, burst and call levels of the ATM traffic [COST92]. However a bottleneck of the applicability of these complex models have been realized.

So far one of the few approaches to obtain the exact steady state distribution of a G/G/c system is by restricting the arrival process to be of the type Markovian Arrival Process, see [Neu79] and to assume the holding time distribution to be of phase type. Then the system can be analyzed through the usual Markovian approach [Neu81] and [Han83]. However, the size of the state space very quickly exceeds any in practice manageable size even when the sparseness of the transition matrix is taken into account, thereby limiting the feasible models to those with a rather limited number of phases in the arrival process, a very limited number of phases of the holding time distribution and only a small number of servers. Furthermore any exact solution approach is far to complicated to be used for dimensioning purposes.

Finding only the mean and variance of the distribution of the number of occupied servers in an infinite server group, [Eck83] presents formulas for computing the peakedness factor (defined as the fraction variance to mean) for a general stationary arrival process and general holding time distribution. The fact that it is relatively easy to obtain the mean and variance of the distribution, but difficult to obtain the distribution itself, calls for approximations.

The objective of the present paper is three fold. First, we shall investigate how the occupancy distribution and peakedness can be obtained in the G/G/∞ system. In the case of computing the exact occupancy distribution we will only consider phase type renewal arrival processes. The peakedness study is greatly simplified by the work of Eckberg, which give us an efficient tool to compute the variance of the occupancy distribution in the G/G/∞ systems. Secondly, we study two approximations for the occupancy distributions based on matching the mean and variance. The first one is the classical BPP approximation while the second one is the distribution which is obtained by maximizing the entropy subject to matching of the mean and variance. It is shown how the maximum entropy method enforces a distribution which is a discrete version of a normal density function restricted to the positive line. Finally, approximations for the blocking probability of the finite capacity system is suggested based on the derived occupancy distribution of the infinite system.

2 THE OCCUPANCY DISTRIBUTION IN THE G/G/∞ CASE

In this section the model under study is an infinite server group which is offered traffic from a stationary process *S* with arrival intensity *m*. All servers act independently of each other and have the same holding time distribution *H* which is assumed general but with finite mean $1/\mu$ and variance σ^2.

The main interest is the stationary distribution of the number of occupied servers, the occupancy distribution. This distribution itself is not possible to compute in an exact way for a general stationary arrival process. We restrict the arrival process to be renewal and the interarrival time to be of phase-type. The occupancy distribution for this case can be obtained by applying the results of [RNe80]. The peakedness of the occupancy distribution can be computed for general stationary arrival process, as shown in [Eck83].

2.1 The Exact Occupancy Distribution in Case of Phase Type Arrivals

To obtain an exact occupancy distribution Ramaswami and Neuts have investigated the infinite-server queue with phase type input and general holding time distribution [RNe80].

A phase type distribution is defined as a distribution function for the time until absorption in an *(n+1)* state Markov chain with *n* transient state and one absorbing state. A phase type renewal process can be obtained by resetting the Markov chain according to an initial probability vector after each transition to the absorbing state. We consider as a model of an arrival process this PH renewal process.

[RNe80] derives the basic system of differential equations of the queue from which we can get an infinite system of differential equations for the time dependent occupancy distribution. Truncated at a sufficient large value of the index *k* they can be solved numerically. (See [RNe80] for details.)

The main advantages of this approach that it produces both the transient and the steady state occupancy distributions, moreover, the complexity of the solution is independent on the holding time distribution. However, for steady state analysis, this solution approach is rather complex, and we have experienced difficulties in the numerical calculations for cases in which the steady state probability of an empty system is very small.

2.2 The Peakedness of the Occupancy Distribution

The peakedness (variance to mean ratio) of the occupancy distribution has been studied in [Eck83] assuming only that the arrival process is stationary. According to formula (3) in [Eck83] then the peakedness *Z* of the complementary holding time distribution *H* is

$$Z(H) = 1 + 2\mu \int_{0-}^{\infty} H_2^c(x)\, dU(x) - \frac{m}{\mu} \tag{2.1}$$

where $U(x)$ denotes the renewal function of the arrival process, which is defined as the expected number of arrivals in a time window starting just after an arrival with size of x and $H_2^c(x)$ denotes the autocorrelation function of the complementary holding time distribution.

Here we would like to emphasize the importance of introducing this generalized peakedness which

offers a very efficient way to get the variance in a large subclass of the G/G/∞ queues. As an application of the generalized peakedness a closed form expression can be derived for queues with the Coxian holding time distribution as shown in [MBla94].

3 APPROXIMATIONS FOR THE DISTRIBUTION

Here we shall consider two different approximation procedures both based on matching the mean and variance of the occupancy distribution in the infinite server case.

3.1 The BPP Approximation

The BPP arrival process is a state dependent Poisson process characterized by two parameters α and β such that the Poissonian arrival intensity when k servers are occupied is $\alpha + k\beta$. In the case $\beta = 0$ it reduces to the plain Poisson process, for $\beta < 0$ it represents a process of less variability than Poisson (finite source model), and for $\beta > 0$ it represents a process of higher variability than Poisson. The mean and variance of the occupancy distribution turns out to be:

$$M = \frac{\alpha}{\mu - \beta} \qquad V = \frac{\alpha\mu}{(\mu - \beta)^2} \tag{3.1}$$

where $1/\mu$ is the mean holding time.

With the mean M and variance V given (from e.g the Eckberg approach) we suggest to approximate the occupancy distribution with the BPP distribution with the same mean and variance. The BPP parameters α and β then should be chosen as:

$$\beta = 1 - \frac{1}{Z} \qquad \alpha = M(1 - \beta) \tag{3.2}$$

assuming a mean holding time of 1.

3.2 The Maximum Entropy Approximation

The concept of entropy has played a significant role in various technical disciplines like thermodynamics, statistical mechanics and information theory. The idea also appears in the mathematical theory of interconnecting networks and in the queueing theory [Ben65], [Kou86]. In queueing theory we only know of its existence in the 1-server case and usually it has been applied in a way where only average quantities like mean queue length and utilization has been matched. Here we shall apply this technique on a many server loss system where also the variance is matched.

Consider a stationary stochastic process $X(t)$ in a discrete state space and let p_i be the probability of being in state i. The entropy of $X(t)$ with stationary distribution $\{p_i\}$ is defined as:

$$H(\boldsymbol{p}) = -\sum_i p_i \ln p_i$$

If a certain value comes out with probability 1, the entropy is 0, and in all other cases it is larger than 0. If the outcome is constrained to a finite set of values, the distribution which maximizes the entropy is

the uniform distribution in which all outcome are equally likely. The less we know of the system, the larger the entropy is.

The idea here is, as an approximation for the occupancy distribution, to take the one which maximizes the entropy under the constrains that

- it should be a proper probability distribution i.e $\sum_{i=0}^{\infty} p_i = 1$
- the mean should be correct i.e $\sum_{i=0}^{\infty} i p_i = E(X) = \frac{m}{\mu}$
- and the second moment should be correct i.e. $\sum_{i=0}^{\infty} i^2 p_i = E(X^2) = \frac{m}{\mu}\left(Z(H) + \frac{m}{\mu}\right)$

If this result is applied to a single server queue and the mean queue length is matched, the queue length distribution which maximizes entropy is geometrical thus yielding the exact distribution for the queue length in the M/M/1 case.

When matching the first and second moment in the infinite server case, a straightforward application of the theorem in Chapter 8.4.1 of [Har92] shows that the occupancy distribution which maximizes entropy is:

$$p_i = P\{X = i\} = g x_1^i x_2^{i^2} = g e^{i \ln x_1} e^{i^2 \ln x_2}$$

Thus the maximum entropy approach enforces a distribution which comes from sampling the normal density function at non-negative integer values. The 3 equations needed to match the mean, variance and obtain a proper distribution is:

$$\sum_{i=0}^{\infty} g x_1^i x_2^{i^2} = 1 \qquad \sum_{i=0}^{\infty} g i x_1^i x_2^{i^2} = E\{X\} \qquad \sum_{i=0}^{\infty} g i^2 x_1^i x_2^{i^2} = E\{X^2\}$$

and we have only been able to solve them by heuristic methods. However, applying the corresponding continuous approach fitting a normal density function restricted to the positive line yields a system of equation which can be solved in an exact way as demonstrated in [MBla94]. Our numerical results indicates that the difference between the continuous and the discrete results are very small.

4 ESTIMATION OF BLOCKING PROBABILITIES

4.1 Truncation of the occupancy distribution

So far we have only considered the system with *infinitely* many servers. In order to arrive at the occupancy distribution of the finite system with c servers we first assume that the system under consideration can be modelled as a time reversible Markov process. For this case a corollary from [Kel79] (corollary 1.10) shows that the exact equilibrium distribution for the finite server case can be obtained by truncating and renormalizing the distribution related to the infinite server case.

In the case where the arrival process is a state dependent Poisson process (this includes the BPP case) this approach is exact.

For all cases where the system is not a reversible Markov process we will approximate the occupancy distribution in the finite case by the same formula.

As blocking probability we use the traffic congestion defined as:

$$\text{traffic congestion} = \frac{\text{offered traffic} - \text{carried traffic}}{\text{offered traffic}}$$

where the offered traffic is the mean number of occupied servers in the infinite system (mean holding time = 1), and the carried traffic is the mean of the truncated distribution.

4.2 When occupancy distribution is a Normal Distribution

When the occupancy distribution is a truncated normal distribution, we are able to find a closed form expression for the traffic congestion.

We have the infinite server occupancy distribution $f(x) = \begin{cases} g e^{-\alpha x} e^{-\beta x^2} & \text{for } x > 0 \\ 0 & \text{for } x < 0 \end{cases}$ implying that

$$\text{offered traffic} = g\sqrt{\frac{\pi}{\beta}} e^{\frac{\alpha^2}{4\beta}} \left\{ \left(-\frac{\alpha}{2\beta}\right) \Phi\left(\frac{-\alpha}{\sqrt{2\beta}}\right) + \frac{1}{\sqrt{2\beta}} \varphi\left(\frac{-\alpha}{\sqrt{2\beta}}\right) \right\} \tag{4.1}$$

where Φ is the standard normal distribution and φ is the corresponding density function. Thereby

$$\text{carried traffic} = \frac{\int_0^c x g e^{-(\beta x^2 + \alpha x)}\, dx}{\int_0^c g e^{-(\beta x^2 + \alpha x)}\, dx}$$

$$= \frac{\frac{-\alpha}{2\beta}\left(\Phi\left(\frac{-\alpha}{\sqrt{2\beta}}\right) - \Phi\left(-\sqrt{2\beta}c - \frac{\alpha}{\sqrt{2\beta}}\right)\right) + \frac{1}{\sqrt{2\beta}}\left(\varphi\left(\frac{\alpha}{\sqrt{2\beta}}\right) - \varphi\left(\sqrt{2\beta}c + \frac{\alpha}{\sqrt{2\beta}}\right)\right)}{\Phi\left(\sqrt{2\beta}c + \frac{\alpha}{\sqrt{2\beta}}\right) - \Phi\left(\frac{\alpha}{\sqrt{2\beta}}\right)} \tag{4.2}$$

thereby yielding a closed form expression for the blocking probability.

4.3 In case of BPP arrivals

For the BPP approximation no closed form expression for the carried traffic is available. Instead, it is possible to use the call congestion (= probability of an arbitrary call is blocked) of the truncated system which by the arrival theorem can be found directly from the time congestion (= the probability of a full system at an arbitrary point in time) of the system with one fewer customer. However, the call congestion can differ significantly from the traffic congestion since the truncation immediately effects the arrival process (states above the truncation level disappear).

4.4 In case of renewal input and exponential holding times

In case of the GI/M/c queues it can be seen that the call congestion equals to the traffic congestion as shown in [MBla94]. For the GI/M/c system the occupancy distribution can be computed by the method of Takács (Chapter 3 and 4 in [Tak62]) and we also used that way.

5 NUMERICAL RESULTS

We have found the exact blocking probability of the GI/M/c case for four different interarrival time distributions, the deterministic, the Erlang-4, a hyperexponential with squared coefficient of variation $c_a^2 = 4$ and a hyperexponential with $c_a^2 = 20$. In all cases the offered traffic is kept equal to 10 and the number of servers is 15. Then we have evaluated four different approximative blocking probabilities.

In the column named "truncated", we have used the occupancy distribution in the infinite system at arbitrary time, and found the traffic congestion by truncation and renormalizing it.

In the column "BPP-app", we have used the BPP occupancy distribution, normalized it and used the relation $P_{\text{call cong}} = P_{\text{time cong}}\left(1 + \frac{c}{M}(Z-1)\right)$ from formula 16 in [Del81] to get the blocking probability from the time congestion.

In the column "tr. normal" we have used the traffic congestion given from (4.1) and (4.2).

Finally in column "Hayward" we have applied the classical Hayward approximation in which the blocking probability in a system with c servers and an offered traffic A with peakedness Z is approximated by the blocking probability in a system with c/Z servers and offered Poisson traffic A/Z.

Table 1: Approximate and exact blocking probabilities (offered traffic = 10, c = 15)

	exact	truncated	BPP-app	tr. normal	Hayward
$c_a^2 = 0$	0.67%	0.60%	0.40%	0.78%	0.90%
$c_a^2 = \frac{1}{4}$	1.31%	1.24%	1.03%	1.42%	1.55%
$c_a^2 = 4$	13.5%	15.0%	12.1%	15.2%	11.6%
$c_a^2 = 20$	25.1%	44.4%	32.2%	42.5%	33.9%

Except for the peaky arrival process with $c_a^2 = 20$ all the approximations performs reasonable well. The BPP approximation underestimates in the cases with a regular arrival process while the truncated normal distribution overestimates for peaky arrival processes.

6 CONCLUSION

It is possible to obtain reasonable approximations for the occupancy distribution in the G/G/∞ queue by approximating it with either a BPP distribution or a normal distribution restricted to the positive line with the same mean and variance. It is also to be concluded that the concept of generalized peakedness offers a very simple way of obtaining the variance of the exact occupancy distribution for a large subset of the G/G/∞ queues. Approximate formulas for the blocking probability in the corresponding finite capacity system has been presented and evaluated with satisfactory result in the case with exponential holding time.

The presented approximation procedures can be easily applied in a heterogeneous environment in which connections with different bandwidth requirements and different holding time distributions are offered to the same network. An obvious future activity is to evaluate the accuracy in these cases. Due to the large dimension of the state space involved this is a challenging task.

The application of the approximations on the network dimensioning problem is another interesting subject for future research.

REFERENCES

[Ben65] V.E. Benes, "Mathematical Theory of Interconnecting Networks", Academic Press, 1965.

[BHJ48] E. Brockmeyer, H.L. Hallstrom, A. Jensen, "The Life and Works of A.K. Erlang", Academy of Technical Sciences, Copenhagen, 1948.

[COST92] COST 224 (ed. J. W. Roberts), "Performance evaluation and design of multiservice networks", Final Report, Luxemburg, 1992.

[Del81] L.E.N. Delbrouck, "A Unified Approximate Evaluation of Congestion Functions for Smooth and Peaky Traffic", IEEE Trans. on Comm. Vol 29, Feb.1981, pp. 85-91.

[Eck83] A.E.Eckberg, "Generalized Peakedness of Teletraffic Processes", ITC-10 Session 4.4b, Montreal, 1983.

[Han83] S. Hansen, "Phase Type Distributions in Queueing Theory", Ph.D Thesis no. 41, IMSOR, Denmark, 1983.

[Har92] P.G. Harrison, "Performance Modelling of Communications Networks and Computer Architectures", Addison Wesley 1992.

[Kel79] F.P. Kelly, "Reversibility and Stochastic Networks", Wiley 1979.

[Kou86] D.D. Kourvatsos, "Maximum Entropy and the G/G/1/N queue", Acta Informatica, Vol. 23, pp. 545-465, 1986.

[MBl94] S. Molnár, S. Blaabjerg, "On some simple blocking approximations for the G/G/c queue", COST 242 TD(94)001, Barcelona, Spain, Feb. 10-11, 1994.

[Neu79] M.F. Neuts, "A Versatile Markovian Point Process", J. Appl. Prob. **16**, 1979, pp. 764-779

[Neu81] M.F. Neuts, "Matrix Geometric Solutions in Stochastic Models", John Hopkins University Press, 1981.

[RNe80] V. Ramaswami, M.F. Neuts, "Some Explicit Formulas and Computational Methods for the Infinite-Server Queues with Phase-Type Arrivals", J. Appl. Prob. **17**, 498-514, 1980.

[Sys60] R. Syski, "Introduction to Congestion Theory in Telephone Systems", Oliver and Boyd 1960.

[Tak62] L. Takács, "Introduction to the Theory of Queues", Oxford University Press, 1962