The Performance of Single Resource Loss Systems in Multiservice Networks

N. G. Bean [a] *, R. J. Gibbens [b†] and S. Zachary [c]

[a]Teletraffic Research Centre, Department of Applied Mathematics, University of Adelaide, South Australia 5005, Australia

[b]Statistical Laboratory, University of Cambridge, 16 Mill Lane, Cambridge, CB2 1SB

[c]Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Riccarton, Edinburgh, EH14 4AS

Recent developments in communication networks have led to much interest in systems where traffic of widely differing characteristics is integrated together. In earlier work the authors develop an analysis of single resource loss systems under the assumption of heavy traffic. In this paper we discuss the analysis with special emphasis on its practical implementation for solving real world examples that arise in the study of multiservice networks. The assumption of heavy traffic also holds in this paper, but there is good reason to expect that results are also accurate when the resource is near to critical loading.

1. Introduction

Recent developments in communication networks have led to much interest in systems where traffic of widely differing characteristics is integrated together. In Bean *et al.* [1] we develop an analysis of single resource loss systems under the assumption of heavy traffic. In this paper we discuss the analysis with special emphasis on its practical implementation for solving real world examples that arise in the study of multiservice networks. The extension of this approach to networks is possible by means of the reduced load approximation (see Whitt [11]). The assumption of heavy traffic also holds in this paper, but there is good reason to expect results that are also accurate when the resource is near to critical loading.

Formally we study a resource of integer capacity C offered a finite number of traffic streams indexed in a set J. Calls of type $j \in J$ arrive as a Poisson stream of rate ν_j and have exponential holding times of mean μ_j^{-1}; each such call requires integer capacity e_j, and is accepted if and only if the resulting free capacity of the resource is at least r_j; otherwise the call is lost. The parameters r_j are referred to as *trunk reservation parameters* and provide an important mechanism for controlling the behaviour of the system.

*Work carried out in the Statistical Laboratory, University of Cambridge supported by the George Murray Scholarship, Cambridge Commonwealth Trust and the SERC under Grant GR/F 94194.
†Supported by the SERC under Grant GR/F 94194.

All arrival streams and holding times are independent. In order to ensure irreducibility of all the stochastic processes involved, we assume, without loss of generality, that the capacities e_j, $j \in J$, have greatest common divisor equal to 1.

Various authors have considered forms of this model. Kaufman [6] considers the model without trunk reservation and develops an elegant and efficient recursion technique for exactly determining the blocking probabilities (see also Dziong and Roberts [2] and Zachary [12] for generalizations to networks). Tran-Gia and Hübner [10] develop a simple approximate technique based on the Kaufman recursion for the model with trunk reservation. Gersht and Lee [3] present an approximate scheme, also based on the Kaufman recursion, to model the different call holding times of different traffic types and include a fixed point calculation to determine the *average* departure rate of a unit of bandwidth. Kelly [8] also addresses the model with trunk reservation but uses the Erlang fixed point approximation [7] with the additional step of calculating the average departure rate of a unit of bandwidth. However, no error calculation or asymptotic justification is presented for any of these approximate methods.

The capacity requirements e_j, $j \in J$, of calls offered to the resource correspond to the important concept of *effective bandwidths* that has arisen from many studies of multiservice networks [4,5,9]. The effective bandwidth is an accurate assessment of the capacity required by a traffic source at each resource of the network in order to guarantee constraints on cell loss or delay.

Of particular interest for loss systems is the determination of the *blocking probability*, that is, the equilibrium call rejection probability, associated with each call type. For the model considered in this paper it is possible in principle to determine blocking probabilities exactly, since they are functions of the equilibrium distribution of a Markov process. However the state space for this process is typically so large as to make this determination impossible in practice. In Section 2 we show how, by studying the stochastic process which describes merely the free capacity of the system at any time, good approximations to the blocking probabilities may be obtained. These approximations are asymptotically exact under the limiting scheme described there. In Section 3 we consider some numerical aspects of the determination of these approximations. In Section 4 we give some numerical examples.

2. Analysis of the Model

Let $\boldsymbol{N}(t) = (N_j(t),\ j \in J)$ where $N_j(t)$ is the number of calls of type j in progress at time t. Then $\boldsymbol{N}(\cdot)$ is a (vector) Markov process with state space $\mathcal{S} = \{\boldsymbol{n} \in \mathbb{Z}_+^{|J|} : \sum_{j\in J} e_j n_j \le C\}$ (where $\mathbb{Z}_+$ is the set of non-negative integers). Let $\pi^* = (\pi^*(\boldsymbol{n}),\ \boldsymbol{n} \in \mathcal{S})$ denote its equilibrium distribution.

For each time t, define also $M(t) = C - \sum_{j\in J} e_j N_j(t)$. Then $M(t)$ is the free capacity at time t and the process $M(\cdot)$ takes values in the state space $\mathcal{M} = \{0, 1, \ldots, C\}$. The equilibrium distribution $\bar{\pi} = (\bar{\pi}(m),\ m \in \mathcal{M})$ of this process is given by $\bar{\pi}(m) = \sum_{\boldsymbol{n}\in\mathcal{S}(m)} \pi^*(\boldsymbol{n})$, where $\mathcal{S}(m) = \{\boldsymbol{n} \in \mathcal{S} : \sum_{j\in J} e_j n_j + m = C\}$. Note that a knowledge of the one-dimensional distribution $\bar{\pi}$ is sufficient for the determination of blocking probabilities. We now show how $\bar{\pi}$ may be determined, at least to a good approximation, without the very much more difficult evaluation of the higher-dimensional distribution π^*.

Define

$$\bar{\theta}_j(m) = \frac{\sum_{\boldsymbol{n}\in\mathcal{S}(m)} \pi^*(\boldsymbol{n}) n_j}{\sum_{\boldsymbol{n}\in\mathcal{S}(m)} \pi^*(\boldsymbol{n})}, \qquad m \in \mathcal{M}, \quad j \in J. \tag{1}$$

Thus $\bar{\theta}_j(m)$ is the expected value of $N_j(t)$ under the equilibrium distribution π^*, conditioned on the event $N(t) \in \mathcal{S}(m)$. Define also $\bar{\boldsymbol{\theta}} = (\bar{\theta}_j(m),\ m \in \mathcal{M},\ j \in J)$. Then $(\bar{\boldsymbol{\theta}}, \bar{\pi})$ satisfies the system of equations (2)–(5) in $(\boldsymbol{\theta}, \pi)$, where $\boldsymbol{\theta} = (\theta_j(m),\ m \in \mathcal{M},\ j \in J)$, $\pi = (\pi(m),\ m \in \mathcal{M})$, $\boldsymbol{\theta} \geq 0$ and $\pi \geq 0$, given by

$$\pi(m)\left[\sum_{j: r_j + e_j \leq m} \nu_j + \sum_{j: m+e_j \in \mathcal{M}} \mu_j \theta_j(m)\right] = \sum_{\substack{j: r_j \leq m, \\ m+e_j \in \mathcal{M}}} \pi(m+e_j)\nu_j + \sum_{j: e_j \leq m} \pi(m-e_j)\mu_j\theta_j(m-e_j), \qquad m \in \mathcal{M}, \tag{2}$$

$$\sum_{m\in\mathcal{M}} \pi(m) = 1, \tag{3}$$

$$\sum_{j\in J} e_j \theta_j(m) + m = C, \qquad m \in \mathcal{M} \tag{4}$$

and

$$g_j(\boldsymbol{\theta}, \pi) = 0, \quad j \in J, \tag{5}$$

where, for each j, the function g_j is defined by

$$g_j(\boldsymbol{\theta}, \pi) = \nu_j \sum_{m \geq e_j + r_j} \pi(m) - \mu_j \sum_{m\in\mathcal{M}} \pi(m)\theta_j(m). \tag{6}$$

To see these results, observe first that the equations (2) follow easily, for each m, by summing the global balance equations for the equilibrium distribution π^* of the Markov process $\boldsymbol{N}(\cdot)$ over the states $\boldsymbol{n} \in \mathcal{S}(m)$ and using the definition (1). The equation (3) is trivially satisfied by $\pi = \bar{\pi}$. That $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$ satisfies the equations (4) follows immediately from (1), or from the interpretation of $\bar{\theta}_j(m)$ as an expected value. Finally, the equations (5) (with $(\boldsymbol{\theta}, \pi) = (\bar{\boldsymbol{\theta}}, \bar{\pi})$) may formally be derived from the global balance equations for π^*, or from the observation that, in equilibrium, the expected acceptance rate equals the expected departure rate for each call type j.

Note that the equations (2) and (3) may be regarded as the equations determining the unique equilibrium distribution π of a Markov process on $\mathcal{M}$ with transition rates given, for each $j \in J$, by

$$m \to \begin{cases} m - e_j, & \text{at rate} \quad \nu_j I(m \geq r_j + e_j), \\ m + e_j, & \text{at rate} \quad \mu_j \theta_j(m) I(m + e_j \in \mathcal{M}) \end{cases} \tag{7}$$

(where I denotes the indicator function). Thus, as usual, any one of the equations (2) may be omitted (being implied by the remainder). Similarly, if $(\boldsymbol{\theta}, \pi)$ satisfies the equations (2), then, by multiplying each of these equations by the corresponding value of m and summing over all $m \in \mathcal{M}$, it follows without too much difficulty that $\sum_{j \in J} e_j g_j(\boldsymbol{\theta}, \pi) = 0$. Thus any one of the equations (5) may also be omitted from the above system.

In general the equations (2)–(5) are insufficient to determine $(\bar{\boldsymbol{\theta}}, \bar{\pi})$. We may however use them to determine $(\bar{\boldsymbol{\theta}}, \bar{\pi})$ approximately by making appropriate assumptions, for each j, about the dependence of $\bar{\theta}_j(m)$ on m.

To motivate our approximation scheme we first consider some asymptotic theory. Bean *et al.* [1] show that, when C is large, and under the heavy traffic condition

$$\sum_{j \in J} e_j \frac{\nu_j}{\mu_j} > C, \tag{8}$$

the process $\boldsymbol{N}(\cdot)/C$ evolves approximately as a deterministic dynamical system $\boldsymbol{x}(\cdot)$, and that the process $M(\cdot)/C$ eventually remains close to 0. The fixed points of this dynamical system are in one-to-one correspondence with the solutions of the system of equations (2), (3) and (5) *modified* by replacing $\mathcal{M}$ throughout by $\mathbb{Z}_+$ and, for each $j \in J$, by replacing $\theta_j(m)$ by a positive constant θ_j (independent of m), where we additionally require that (in place of the equations (4))

$$\sum_{j \in J} e_j \theta_j = C. \tag{9}$$

For each solution $(\boldsymbol{\theta}, \pi)$ (where $\boldsymbol{\theta}$ is here interpreted as $(\theta_j,\ j \in J)$) of this modified system of equations, the corresponding fixed point of the dynamical system is given by $\boldsymbol{\theta}/C$.

Now suppose that the dynamical system $\boldsymbol{x}(\cdot)$ possesses a unique fixed point $\bar{\boldsymbol{x}}$ to which all of its trajectories converge (a condition satisfied in all the examples we have studied). Bean *et al.* [1] show that the component π of the corresponding unique solution $(\boldsymbol{\theta}, \pi)$ of the above modified system of equations then approximates the equilibrium distribution $\bar{\pi}$ of the process $M(\cdot)$. The intuitive explanation for this is that, since in equilibrium $\boldsymbol{N}(\cdot)/C$ remains close to $\bar{\boldsymbol{x}}$ and $M(\cdot)/C$ remains close to 0, for each j the constant θ_j becomes a reasonable approximation to $\bar{\theta}_j(m)$. (Further, under these conditions, the process $M(\cdot)$ behaves approximately as a Markov process with transition rates given by the set of equations (7) with θ_j replacing $\theta_j(m)$.) Under the limiting scheme of Kelly [7], in which the arrival rates ν_j and the capacity C are allowed to grow in proportion, and with the heavy traffic condition (8) continuing to hold, the solution π of the modified system of equations remains constant and is the limit (under weak convergence) of the exact equilibrium distribution $\bar{\pi}$.

If the dynamical system $\boldsymbol{x}(\cdot)$ does possess multiple fixed points, then the component π of each of the corresponding solutions of the above modified system of equations represents an approximate quasi-equilibrium distribution of the process $M(\cdot)$, that is, a distribution which behaves as an equilibrium distribution over a sustained period of time. However, we have no evidence to suggest that such multiple fixed points can occur.

While the above approximation to $\bar{\pi}$ becomes exact under the limiting scheme described, it is nevertheless too crude for most practical applications. We therefore seek to improve it by retaining the state space $\mathcal{M}$ and, for each j, approximating $\bar{\theta}_j(m)$ as a linear function of m. Although this linear approximation is less than ideal from a theoretical viewpoint, it nevertheless appears to work well in practice. We also expect that it will continue to work reasonably well even when the heavy traffic condition (8) is not satisfied, particularly if the resource is instead at or near to critical loading. Recalling that $\bar{\theta}_j(m)$ must also be always positive, we thus replace the equations (4) by

$$\theta_j(m) = a_j(C - m), \qquad j \in J, \quad m \in \mathcal{M} \tag{10}$$

where

$$a_j \geq 0, \quad \text{for each } j \in J, \quad \text{and} \quad \sum_{j \in J} e_j a_j = 1. \tag{11}$$

The equilibrium distribution $\bar{\pi}$ is then approximated by the component π of the solution $(\boldsymbol{a}, \pi)$ (where $\boldsymbol{a} = (a_j,\ j \in J)$) of the system of equations (2), (3), (5), (10) and (11), provided that this solution is unique. Note that, after eliminating $\boldsymbol{\theta}$ and the redundant equations, the above system of equations consists of $|\mathcal{M}| + |J|$ equations in $|\mathcal{M}| + |J|$ unknowns. Multiple solutions, were they to occur, would again correspond to quasi-equilibrium distributions of the process $M(\cdot)$.

3. Practical Discussion

In the previous section we describe a model for approximating $\bar{\pi}$ by treating $\bar{\theta}_j(m)$ as linear in m. In this section we consider the numerical determination of this approximation through the solution of the system of equations (2), (3), (5), (10) and (11). This problem may be regarded as that of finding the component $\boldsymbol{a}$ of the solution $(\boldsymbol{a}, \pi)$: for any given $\boldsymbol{a}$ in the $(|J| - 1)$-dimensional region defined by the conditions (11), the equations (10) determine a value of $\boldsymbol{\theta}$ and then the equations (2) and (3) determine a distribution π; we must therefore choose $\boldsymbol{a}$ such that $(\boldsymbol{\theta}, \pi)$ satisfies the equations (5) (where again any one member of this set may be omitted). In the special case where $|J| = 2$ a bi-section method can be used, but in general a modified version of the Newton-Raphson routine in which derivatives need not be supplied is highly effective.

A further consideration arises in the solution of the equations (2) and (3) (for given $\boldsymbol{\theta}$) at each iteration of any numerical procedure. For moderate values of C these may be solved by matrix inversion (omitting any one of the equations (2)). For large C this is impractical. In this case we may make a further appeal to the asymptotic theory of Bean *et al.* [1] which shows that, provided $\theta_j(m)/C$ varies slowly with m (as is the case here), the tail of the distribution π determined by the equations (2) and (3) is approximately geometric with a parameter p which can be directly determined in the manner described below. We may therefore choose a threshold value $m_0 \in \mathcal{M}$ and assume that

$$\pi(m) = \pi(m_0) p^{m - m_0}, \qquad \text{for all } m \geq m_0. \tag{12}$$

The parameter p should be taken as the unique real root between 0 and 1 of the polynomial $f : \mathbb{C} \to \mathbb{C}$ of degree $2\hat{e}$ (where $\hat{e} = \max_{j \in J} e_j$) given by

$$f(z) = z^{\hat{e}} \sum_{j \in J} (\nu_j + \mu_j \theta_j(m_0)) - \sum_{j \in J} \left(z^{\hat{e} + e_j} \nu_j + z^{\hat{e} - e_j} \mu_j \theta_j(m_0) \right). \tag{13}$$

The value of m_0 should be chosen so that $\sum_{m \geq m_0} \pi(m)$ is small. If m_0 and p are thus determined, the set of equations (2) need only be considered for $m = 0, 1, \ldots, m_0 - 1$. Substitution of the relation (12) into these equations and equation (3) yields $m_0 + 1$ linear equations in $\pi(0), \pi(1), \ldots, \pi(m_0)$, so determining the entire distribution π. An alternative approach to the problem caused by large C would be to assume that $\pi(m) = 0$ for all $m \geq m_0$. However, to achieve a similar degree of accuracy would require a much greater value of m_0 as $\sum_{m \geq m_0} \pi(m)$ would then need to be negligible.

4. Examples

In this section we use the algorithm of the previous section to calculate blocking probabilities for the call types given in Table 1 and the choice of system parameters given in Table 2, where the arrival rate of calls is in units of call per second rounded to two decimal places. We consider four choices of call characteristics, described by the effective bandwidth and mean call holding time, which might, for example, be associated with the applications shown.

Table 1
Parameter values used to describe each call

Application	Call Type	Effective Bandwidth (Mbps)	Holding Time (s)
Digitized Voice	I	0.04	180
Interactive Video Retrieval	II	0.50	1200
File Transfer	III	2.00	60
Distribution Video	IV	2.00	1800

Table 2
System parameter values

Figure	Type 1		Type 2		Capacity (Mbps)
	Call Type	Arrival Rate	Call Type	Arrival Rate	
1	II	0.24	IV	0.13	622
2	III	2.68	IV	0.08	622
3	I	32.38	II	0.70	622

For each of the examples defined in Table 2, we present the analytic results and compare them to simulations of the system, for which we also display the relevant 99% confidence intervals. In each example we consider a single resource and two types of offered traffic, and write $s = r_1 - r_2$ while holding at least one of r_1 or r_2 equal to zero. We study how well our model approximates the two blocking probabilities as s varies. (The choice of an optimal value of s will of course depend on the cost structure adopted, and perhaps other criteria.)

Figure 1 shows an example where the algorithm gives highly accurate estimates of the blocking probabilities for both types of call over a wide range of values for the trunk reservation parameters. In this example the link is critically loaded in that

$$\sum_{j \in J} \frac{e_j \nu_j}{\mu_j} = C. \tag{14}$$

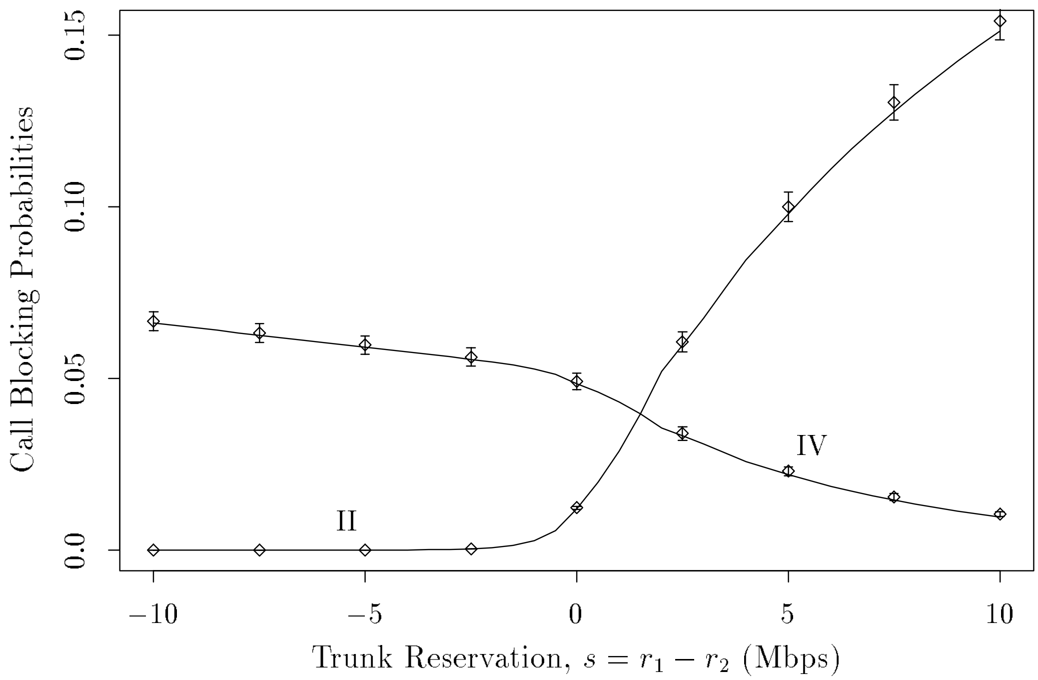

Figure 1. Analytical and simulation results with varying trunk reservation

This figure shows the typical behaviour of the blocking probabilities at such a link. The relationship between the blocking probabilities of the call types is governed by the relative values of $e_1 + r_1$ and $e_2 + r_2$. So, for example, when $e_1 + r_1 = e_2 + r_2$ the blocking probabilities are equal for both call types. The blocking probability for the calls of type 1 is almost zero when the trunk reservation parameters are such that s is large and negative and as s increases, the blocking probability decreases for calls of type 2 and increases for calls of type 1. Note that when there is no trunk reservation, that is $s = 0$, the calls with the higher effective bandwidth have a higher blocking probability than those of the other call type.

In Figure 2 the link is again critically loaded and has two call types of identical effective bandwidth but very different holding times. We again find that the algorithm provides accurate estimates for the blocking probabilities.

Finally, in Figure 3, we consider a more extreme case where both the effective bandwidths and holding times differ considerably between call types. The link satisfies the

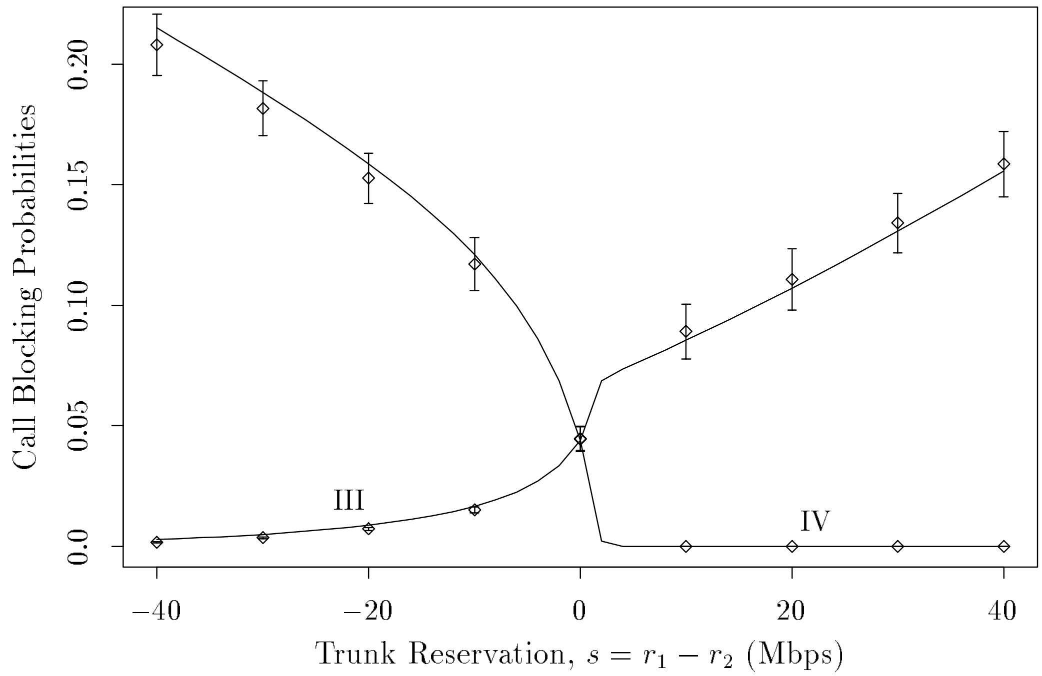

Figure 2. Analytical and simulation results with varying trunk reservation

heavy traffic condition (8), in fact

$$\sum_{j \in J} \frac{e_j \nu_j}{\mu_j} = C \times 1{\cdot}05, \tag{15}$$

corresponding to 5% overload. In this case the results are somewhat less accurate despite capturing all the qualitative features.

Our results show that the algorithm described in this paper can be used in a wide range of practical examples to give very accurate estimates for the blocking probabilities in multiservice networks. Further, the asymptotic theory of the preceding sections shows that any inaccuracies decrease as the system capacity increases.

To improve these results, for example where call types vary widely or where the load is light, would require further investigation into the form of $\theta_j(m)$ (see equations (10) and (11)).

REFERENCES

1. N. G. Bean, R. J. Gibbens, and S. Zachary. Analysis of Large Single Resource Loss Systems under Heavy Traffic, with Appplications to Integrated Networks. *To appear in Adv. Appl. Prob.*, March, 1995.
2. Z. Dziong and J. W. Roberts. Congestion Probabilities in a Circuit Switched Integrated Services Network. *Performance Evaluation*, 7:267–284, 1987.

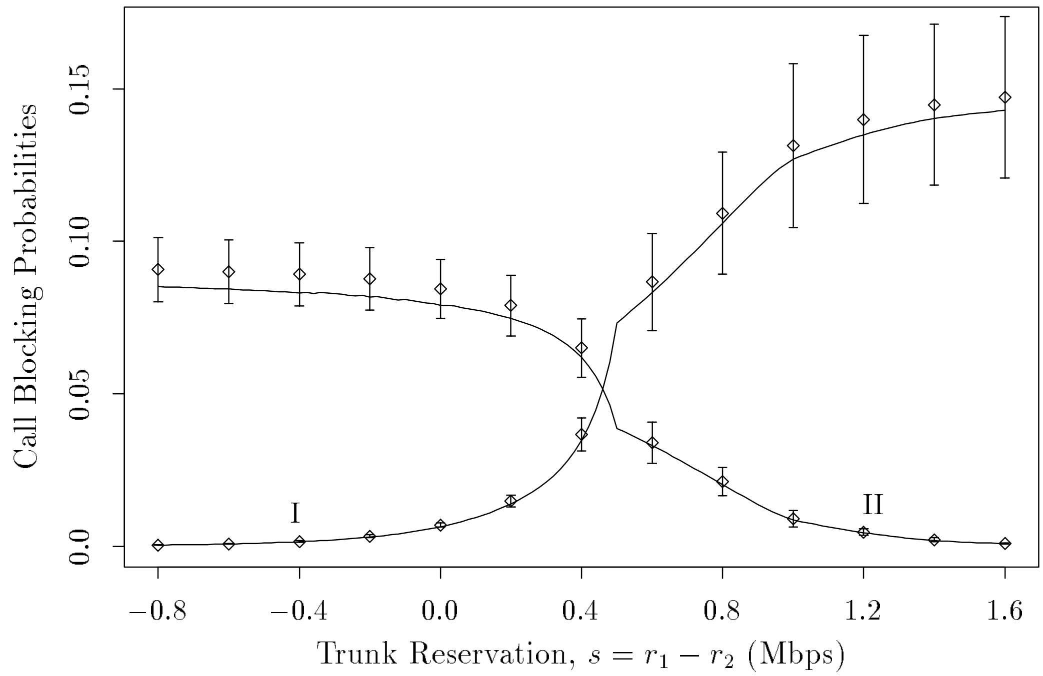

Figure 3. Analytical and simulation results with varying trunk reservation

3. A. Gersht and K. J. Lee. A Bandwidth Management Strategy in ATM Networks. Pre-print, GTE Laboratories, 40 Sylvan Road, Waltham, MA 02254, USA., 1990.
4. R. J. Gibbens and P. J. Hunt. Effective Bandwidths for the Multi-type UAS Channel. *Queueing Systems*, 9:17–28, September 1991.
5. J. Y. Hui. Resource Allocation for Broadband Networks. *IEEE J.S.A.C.*, SAC-6(9):1598–1608, December 1988.
6. J. S. Kaufman. Blocking in a Shared Resource Environment. *IEEE Trans. Comm.*, COM-29(10):1474–1481, October 1981.
7. F. P. Kelly. Blocking Probabilities in Large Circuit Switched Networks. *Adv. Appl. Prob.*, 18:473–505, 1986.
8. F. P. Kelly. Fixed Point Models of Loss Networks. *J. Austral. Math. Soc. Ser. B*, 31:204–218, 1989.
9. F. P. Kelly. Effective Bandwidths at Multi-Class Queues. *Queueing Systems*, 9:5–16, September 1991.
10. P. Tran-Gia and F. Hübner. An Analysis of Trunk Reservation and Grade of Service Balancing Mechanisms in Multiservice Broadband Networks. In *IFIP Workshop TC6, Modelling and Performance Evaluation of ATM Technology*, La Martinique, French Carribean Island, 25-27 January 1993.
11. W. Whitt. Blocking When Service Is Required From Several Facilities Simultaneously. *AT&T Technical Journal*, 64(8):1807–1856, October 1985.
12. S. Zachary. On Blocking in Loss Networks. *Adv. Appl. Prob.*, 23:355–372, 1991.