J. Appl. Prob. **44,** 1088–1102 (2007)
Printed in England

ASYMPTOTIC BLOCKING PROBABILITIES IN LOSS NETWORKS WITH SUBEXPONENTIAL DEMANDS

YINGDONG LU * ** AND
ANA RADOVANOVIĆ,* *** *IBM T.J. Watson Research Center*

Abstract

The analysis of stochastic loss networks has long been of interest in computer and communications networks and is becoming important in the areas of service and information systems. In traditional settings computing the well-known Erlang formula for blocking probabilities in these systems becomes intractable for larger resource capacities. Using compound point processes to capture stochastic variability in the request process, we generalize existing models in this framework and derive simple asymptotic expressions for the blocking probabilities. In addition, we extend our model to incorporate reserving resources in advance. Although asymptotic, our experiments show an excellent match between derived formulae and simulation results even for relatively small resource capacities and relatively large values of the blocking probabilities.

Keywords: Loss network; subexponential distribution

2000 Mathematics Subject Classification: Primary 60K25
Secondary 60J05; 60K05; 60K10

1. Introduction

The problem of satisfying a stream of customer (user) requirements from resources of finite capacities for some random processing time has long been present in many areas such as telephone and communication networks, inventory control (rental industry), and, recently, workforce management. For all of these applications, the system dynamics can be described as follows. Requests for resources arrive according to some point process in time. If there are enough available (nonengaged) resources to satisfy their requirements at the moment of arrival, required resources are committed for some random time that represents their processing duration (holding time) after which they are released and become available to accommodate future requests. If there is an insufficient amount of available resources at the moment of a request's arrival, the request is lost. The previously described system is usually referred to as a *loss network*, and one of the commonly analyzed performance metrics is the blocking probability, i.e. the probability that an incoming request is lost owing to an insufficient amount of available resources to satisfy its requirements.

Loss networks with fixed resource requirements have been intensively analyzed in the context of circuit-switched networks. Let requests require resources of $K < \infty$ different types for some random generally distributed processing time with finite mean. Furthermore, assume that

Received 15 September 2006; revision received 12 September 2007.
* Postal address: Mathematical Sciences Department, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA.
** Email address: yingdong@us.ibm.com
*** Email address: aradovan@us.ibm.com

requests belong to M different classes characterized by their resource requirements, processing durations, and arrival rates. Then, assuming that requests of different types arrive according to mutually independent Poisson processes, by the PASTA (Poisson arrivals see time averages) property [25], the blocking probability of an incoming request of type $1 \leq l \leq M$ is equal to the sum of probabilities of blocking states for an l-type request and is computed using the generalized Erlang formula (see, e.g. [15]), i.e.

$$1 - G(\boldsymbol{C})^{-1} G(\boldsymbol{C} - \boldsymbol{A}\boldsymbol{e}_l),$$

where

$$G(\boldsymbol{C}) = \left(\sum_{\boldsymbol{n} \in \mathcal{S}(\boldsymbol{C})} \prod_{l=1}^{M} \frac{\rho_l^{n_l}}{n_l!} \right)$$

and

$$\mathcal{S}(\boldsymbol{C}) := \{\boldsymbol{n} \in \mathbb{Z}_+^M : \boldsymbol{A}\boldsymbol{n} \leq \boldsymbol{C}\}, \tag{1}$$

where $\boldsymbol{n} = (n_1, \ldots, n_M)$ and $\boldsymbol{C} = (C_1, \ldots, C_K)$. In the previous expressions C_k, $1 \leq k \leq K$, is the capacity of resource type k, $\boldsymbol{A} = [A_{kl}]$ is a $K \times M$ matrix, where A_{kl} represents the amount of resources of type $1 \leq k \leq K$ required by a request of type $1 \leq l \leq M$, and ρ_l, $1 \leq l \leq M$, represents traffic intensities of l-type requests (computed as $\rho_l = \lambda_l/\mu_l$, where λ_l denotes the arrival rate of l-type requests and $1/\mu_l$ is the corresponding mean processing time). Furthermore, $\boldsymbol{e}_l$ is an M-dimensional vector with the lth component equal to 1 and the rest equal to 0. In the case of a single resource type and a single request class with exponentially distributed processing times, the blocking probability was first expressed by Erlang [7]. Later on, it was shown that Erlang's formula holds under more general assumptions on the call holding-time distributions (see [21]) and holds in the case of Poisson arrivals with retrials (see [4]). It is noteworthy to point out the difference between the Erlang loss network and a queue with finite buffer. The two systems follow very different dynamics resulting in a different behavior and, therefore, analysis; see, e.g. [1] and [11].

It is easy to see that the cardinality of the state space $\mathcal{S}(\boldsymbol{C})$ in (1) increases exponentially in the norm of vector $\boldsymbol{C}$, i.e. $|\boldsymbol{C}| \equiv \sum_{i=1}^{K} |C_i|$. In [18] it was shown that the calculation of $G(\boldsymbol{C})$ is a $\sharp P$-complete problem, which belongs to a class of problems that are at least as hard as NP-complete problems. To this end, many approximation techniques for evaluating blocking probabilities in large loss networks have been proposed. One of the most popular ones is known as the Erlang fixed-point method. The main idea of this approximation is to assume that deficiencies of different resource types occur independently. The application of the Erlang fixed-point method can be traced back as early as the 50s; see, e.g. [24]. In [14] Kelly studied the performance of the Erlang fixed-point method and established its relationship with a nonlinear optimization problem. He also proved uniqueness of the fixed point and its asymptotic exactness when resource capacities and arrival rates grow with the same rate; see [15]. Some of the related practical aspects of Kelly's analysis were investigated in [23]. The Erlang fixed-point method is further refined in [26]. There are also many other types of approximations such as the recursive algorithm investigated in [13], or the unified approach based on large deviations for all (light, critical, and heavy) traffic regimes investigated in [8]. Overall, except for the bounds in [8], these methods make use of the structural properties of the Erlang formula and, hence, largely rely upon the Poisson assumption for call arrivals. Another restriction of the above models is that the amount of resource requirements are assumed to be fixed; in fact, it is assumed that they are (0, 1) parameters in most of the cases considered. Meanwhile, in many

https://doi.org/10.1239/jap/1197908827 Published online by Cambridge University Press

applications we see that resource requirements could be highly variable and their distributions possibly long tailed; for specific examples, see [10], [12], and [16]. Furthermore, more recently, loss network models have been applied in the context of workforce management applications (see [20]), where the requests' behavior is even more volatile and extreme.

In this paper we analyze loss networks that have renewal arrivals and random resource requirements. In particular, we assume that request arrivals follow a compound renewal process, with the corresponding holding times being arbitrarily distributed with finite mean. In addition, we assume that holding times corresponding to different arrivals are mutually independent and independent of the arrival points as well. In order to cope with variability in resource requirements, we model them as subexponential random variables. We obtain a simple and explicit asymptotic expression for the blocking probabilities when the capacities of resources grow. For the case of a single resource loss network, we show that the stationary blocking probability is approximately equal to the tail of the resource requirement distribution. In addition, we extend our results to allow advance reservations of resources. Finally, we investigate general (multiple resources and arbitrary topology) loss networks and show that the asymptotic blocking probability behaves as the tail of the heaviest-tailed resource requirement. Although asymptotic, our numerical experiments show an excellent accuracy of the derived formulae even for relatively small capacities and relatively large values of the blocking probabilities, suggesting wide applicability of the obtained results.

Our paper is organized as follows. In Section 2 we introduce our model in the context of a single resource type. Then, in Subsection 2.1 we state and prove our main result in Theorem 1, while in Subsection 2.2 we extend it to the case of advance reservations. Further extension to the analysis of the stationary blocking probability in the case of general loss networks is stated and proved in Theorem 2 of Section 3. Our simulation experiments for some specific cases of arrival processes and resource requirements are presented in Section 4. Finally, we conclude our paper in Section 5. A discussion and the proof of existence of the stationary blocking probability is presented in the Appendix of [19].

2. Systems with one resource type

Let requests for resources from a common resource pool of capacity $C < \infty$ arrive at time points $\{\tau_n, -\infty < n < \infty\}$, which represent a renewal process with rate $0 < \lambda < \infty$, i.e. $\mathrm{E}[\tau_n - \tau_{n-1}] = 1/\lambda$. At each point τ_n, B_n amount of resources is requested. If available capacity is less than B_n, this request is rejected (blocked); otherwise, it is accepted and B_n amount of resources will be occupied for the length of time θ_n. Sequences $\{B_n\}$ and $\{\theta_n\}$ of independent and identically distributed (i.i.d.) random variables (RVs) are assumed to be mutually independent and independent of the arrival points $\{\tau_n\}$; furthermore, $\mathrm{E}\,\theta_n < \infty$ for all n. Let B and θ denote RVs that represent $\{B_n\}$ and $\{\theta_n\}$, respectively, i.e. $\mathrm{P}[B > x] = \mathrm{P}[B_n > x]$ and $\mathrm{P}[\theta > y] = \mathrm{P}[\theta_n > y]$ for any $n \in \mathbb{Z}$, $x \geq 0$, and $y \geq 0$.

In this paper we assume that B is a subexponential RV defined as follows; see, e.g. [9].

Definition 1. Let $\{X_i\}$ be a sequence of positive i.i.d. RVs with distribution function F such that $F(x) < 1$ for all $x > 0$. Denote by

$$\bar{F}(x) = 1 - F(x), \qquad x \geq 1,$$

the tail of F and by

$$\bar{F}^{n*} = 1 - F^{n*}(x) = \mathrm{P}[X_1 + \cdots + X_n > x]$$

https://doi.org/10.1239/jap/1197908827 Published online by Cambridge University Press

the tail of the n-fold convolution of F. The distribution function F is subexponential, denoted as $F \in \mathcal{S}$, if one of the following equivalent conditions holds:

- $$\lim_{x\to\infty} \frac{\bar{F}^{n*}(x)}{\bar{F}(x)} = n \quad \text{for some (all) } n \geq 2,$$
- $$\lim_{x\to\infty} \frac{\mathrm{P}[X_1 + \cdots + X_n > x]}{\mathrm{P}[\max(X_1, \ldots, X_n) > x]} = 1 \quad \text{for some (all) } n \geq 2.$$

For a brief introduction to subexponential distributions, the reader is referred to a recent survey [9]. This class of distributions is fairly large and well-known examples include regularly varying (in particular Pareto), some Weibull, log-normal, and 'almost' exponential distributions.

Next, let $\mathcal{N}_n^{(C)}$ denote the set of indices $i < n$ of resource requirements that arrive prior to τ_n, are accepted, and are *still active* by time τ_n. Furthermore, define

$$N_n^{(C)} \triangleq |\mathcal{N}_n^{(C)}|$$

to be a cardinality of the set $\mathcal{N}_n^{(C)}$. Thus, the total amount of resources $Q_n^{(C)}$ that an arrival at time τ_n finds engaged can be expressed as

$$Q_n^{(C)} = \sum_{i\in\mathcal{N}_n^{(C)}} B_i.$$

In this paper our goal is to estimate the stationary blocking probability, i.e. to estimate

$$\mathrm{P}[Q_n^{(C)} + B_n > C] \tag{2}$$

for large C. It can be shown that for the model introduced above there exists a unique stationary distribution for $Q_n^{(C)}$ and, therefore, the quantity in (2) is well defined. The proof of this result is based on constructing a Markov chain with general state space, of which $Q_n^{(C)}$ is a functional. Then, by using a discrete version of Theorem 1 of [21], we show that there exists a unique stationary distribution for the constructed Markov chain (and, therefore, $Q_n^{(C)}$) which is ergodic. This proof is not the main focus of this paper and an interested reader is referred to the Appendix of [19] for its details.

In this paper we use the following standard notation. For any two real functions $a(t)$ and $b(t)$, and fixed $t_0 \in \mathbb{R} \cup \{\infty\}$, let $a(t) \sim b(t)$ as $t \to t_0$ denote $\lim_{t\to t_0}[a(t)/b(t)] = 1$.

2.1. The blocking probability in a system with one resource type

In this section we estimate the stationary blocking probability $\mathrm{P}[Q_n^{(C)} + B_n > C]$ in a loss network with a single resource pool when its capacity C grows large.

Theorem 1. *Let $\{B_n, -\infty < n < \infty\}$ denote a sequence of subexponential RVs with finite mean. Then, the stationary blocking probability satisfies*

$$\mathrm{P}[Q_n^{(C)} + B_n > C] \sim \mathrm{P}[B > C] \quad \textit{as } C \to \infty.$$

Proof. First, observe that a request will be lost if it requires more than the total capacity C and, therefore,

$$\mathrm{P}[Q_n^{(C)} + B_n > C] \geq \mathrm{P}[B > C] \quad \text{for all } C > 0. \tag{3}$$

https://doi.org/10.1239/jap/1197908827 Published online by Cambridge University Press

In order to prove the asymptotic upper bound for $\mathrm{P}[Q_n^{(C)} + B_n > C]$, we start by conditioning on the size of B_n as

$$\begin{aligned}\mathrm{P}[Q_n^{(C)} + B_n > C] &= \mathrm{P}[Q_n^{(C)} + B_n > C,\ B_n > C] + \mathrm{P}[Q_n^{(C)} + B_n > C,\ B_n \le C] \\ &\triangleq I_1 + I_2. \end{aligned} \tag{4}$$

Note that I_1 is upper bounded by $\mathrm{P}[B > C]$. Next we prove that $I_2 = o(\mathrm{P}[B > C])$ as $C \to \infty$. In view of the definition of $\mathcal{N}_n^{(C)}$ from above,

$$I_2 = \mathrm{P}\left[\sum_{i \in \mathcal{N}_n^{(C)}} B_i + B_n > C,\ B_n \le C\right]. \tag{5}$$

Observe that, for $i \in \mathcal{N}_n^{(C)}$, the B_is are mutually dependent which makes direct analysis of the expression in (5) complex. For this reason we sample the original process of arrivals at the points τ_i, at which the requested amount of resources B_i is smaller than or equal to C, and observe another system of unlimited capacity with the sampled arrivals. Let $\mathcal{N}_{n,s}$ denote a set of request indices $i < n$ that belong to the sampled process and are still active at time τ_n, i.e.

$$\mathcal{N}_{s,n} = \{i < n \mid B_i \le C,\ \theta_i > \tau_n - \tau_i\}.$$

Note that the sampled process is also a renewal process with rate $\lambda\,\mathrm{P}[B \le C]/\,\mathrm{P}[B > C]$, and that the resource requirements B_i, $i \in \mathcal{N}_{s,n}$, are mutually independent. Furthermore, since $\mathcal{N}_n^{(C)} \subset \mathcal{N}_{s,n}$, we can upper bound I_2 in (5) by the probability that the total amount of required resources in a new system exceeds capacity C, i.e.

$$I_2 \le \mathrm{P}\left[\sum_{i \in \mathcal{N}_{s,n}} B_i + B_n > C,\ B_n \le C\right]. \tag{6}$$

Now, in view of the results derived in [6], for every integer n and for i.i.d. subexponential RVs $B_1, \ldots, B_n$,

$$\mathrm{P}\left[\sum_{i=1}^{n} B_i > C\right] \sim \mathrm{P}[\max(B_1, B_2, \ldots, B_n) > C] \quad \text{as } C \to \infty,$$

implying the asymptotic relation

$$\mathrm{P}\left[\sum_{i=1}^{n} B_i > C,\ B_i \le C \text{ for every } 1 \le i \le n\right] = o(\mathrm{P}[B > C]) \quad \text{as } C \to \infty.$$

In order to show that n can be replaced by $N_{s,n}$ in the above inequality, we need to integrate it with respect to the density of $N_{s,n}$, i.e.

$$\begin{aligned}&\mathrm{P}\left[\sum_{i \in \mathcal{N}_{s,n} \cup \{n\}} B_i > C,\ B_i \le C \text{ for every } i \in \mathcal{N}_{s,n} \cup \{n\}\right] \\ &\quad = \sum_{k=0}^{\infty} \mathrm{P}[N_{s,n} = k]\,\mathrm{P}\left[\sum_{i=1}^{k+1} B_i > C,\ B_i \le C \text{ for every } i = 1, \ldots, k+1\right].\end{aligned}$$

https://doi.org/10.1239/jap/1197908827 Published online by Cambridge University Press

Note that on the left-hand side of the previous equation index i can take negative values. Next, owing to the lemma stated by Kesten (see [3, Lemma 7]), for any $\varepsilon > 0$, there exists a positive constant $K(\varepsilon)$ such that

$$\frac{\mathrm{P}\left[\sum_{i=1}^{k} B_i > C,\ B_i \le C \text{ for every } 1 \le i \le k\right]}{\mathrm{P}[B > C]} \le \frac{\mathrm{P}\left[\sum_{i=1}^{k} B_i > C\right]}{\mathrm{P}[B > C]} \le K(\varepsilon)(1+\varepsilon)^k$$

for any integer k and for all capacity values $C < \infty$. Then, since the probability generating function $\mathrm{E}\, z^{N_{s,n}}$ is finite for any $z \in \mathbb{C}$ (see [17, Theorem 5] and [22, Theorem 1] for the detailed proof), we have

$$\sum_{k=0}^{\infty} \mathrm{P}[N_{s,n} = k](1+\varepsilon)^k < \infty.$$

Therefore, by applying the dominated convergence theorem we conclude that

$$\begin{aligned}
&\lim_{C\to\infty} \frac{\mathrm{P}\left[\sum_{i\in N_{s,n}} B_i + B_n > C,\ B_i \le C \text{ for every } i \in \mathcal{N}_{s,n} \cup \{n\}\right]}{\mathrm{P}[B > C]} \\
&\quad = \lim_{C\to\infty} \sum_{k=0}^{\infty} \frac{\mathrm{P}[N_{s,n} = k]\,\mathrm{P}\left[\sum_{i=1}^{k+1} B_i > C,\ B_i \le C \text{ for every } 1 \le i \le k+1\right]}{\mathrm{P}[B > C]} \\
&\quad = 0,
\end{aligned}$$

which, in conjunction with (3) and (4), completes the proof.

Remark. It may appear surprising that the performance of the loss network from above does not depend on engagement durations, as long as they have finite mean. In addition, the result is quite general and provides the asymptotic result for a large (subexponential) class of possible resource requirement distributions.

2.2. Advance reservations

Using the result of Theorem 1 and observations from the previous remark, we extend the loss networks model to allow requests to become effective with some delay with respect to the moments of their arrivals. In particular, a request that arrives at time τ_n and requires B_n amount of resources for some random time θ_n starting from the moment $\tau_n + D_n$ is accepted if previously admitted resource requirements allow that; otherwise, it is rejected. In other words, a request arriving at τ_n is lost if at any moment of time in the interval $(\tau_n + D_n, \tau_n + D_n + \theta_n)$ the total amount of active requirements requested prior to τ_n exceeds $C - B_n$. First, note that $B_n > C$ implies the loss of the nth request and, therefore, it is straightforward to conclude that the blocking probability in the system with advance reservations can be lower bounded by $\mathrm{P}[B > C]$.

Next we discuss the idea behind proving the upper bound on the blocking probabilities. By applying sample path arguments we can show that, at any moment of time, the amount of active resources in the previously described system with advance reservations can be bounded from above by the amount of active resources in another system of unlimited capacity, without advance reservations, with resource holding times $D_n + \theta_n$ for every n, and with requests for resources being sampled from the original process $\{B_n\}$ whenever the corresponding requirements are less than or equal to C. Equivalently, the blocking probability in the system with

https://doi.org/10.1239/jap/1197908827 Published online by Cambridge University Press

advance reservations can be bounded from above by

$$\mathrm{P}\Bigg[\sum_{i\in\mathcal{N}_{s,n}^{(C)}(\theta+D)} B_i + B_n > C\Bigg],$$

where $\mathcal{N}_{s,n}^{(C)}(\theta+D)$ is a set of request indices $i < n$ that are active at time τ_n, whose requirements are less than or equal to C and holding times last throughout the interval $(\tau_i, \tau_i + D_i + \theta_i)$, assuming that there is an unlimited resource capacity.

Finally, using the previous discussion, the properties of $\{B_n\}$, $\{\theta_n\}$, and $\{\tau_n\}$, introduced at the beginning of this section, assuming that the reservation times $\{D_n\}$, $\mathrm{E}\,D_n < \infty$, are i.i.d. and independent from $\{B_n\}$, $\{\theta_n\}$, and $\{\tau_n\}$, and applying the identical arguments used in the proof of Theorem 1, we obtain the following result.

Corollary 1. *The blocking probability in the system with advance reservations approaches* $\mathrm{P}[B > C]$ *as* $C \to \infty$.

3. Acquiring resources of different types (loss networks case)

Assume that there are $K \in \mathbb{N}$ resource types with capacities $C_1, \ldots, C_K$. Again, requests arrive at $\{\tau_n, -\infty < n < \infty\}$, which represent a renewal process with rate

$$0 < \lambda = \frac{1}{\mathrm{E}[\tau_1 - \tau_0]} < \infty.$$

There are $M < \infty$ request types and, given an arrival, the request is of type l, $1 \le l \le M$, with probability p_l, where $p_1 + \cdots + p_M = 1$, independent of $\{\tau_n\}$. We will use RVs $J_n \in \{1, 2, \ldots, M\}$ to denote the type of request arriving at τ_n. Furthermore, let $B_n^{(J_n,1)}, \ldots, B_n^{(J_n,K)}$ represent the amounts of required resources of each type at time τ_n, and let $\theta_n^{(J_n)}$, $\mathrm{E}\,\theta_n^{(J_n)} < \infty$, denote the corresponding random duration. We assume that the sequences $\{(B_n^{(J_n,1)}, \ldots, B_n^{(J_n,K)})\}$ and $\{\theta_n^{(J_n)}\}$ are mutually independent and independent of $\{\tau_n\}$. Given the event $\{J_n = l\}$, resource requirements $B_n^{(l,i)}$, $1 \le i \le K$, are mutually independent, nonnegative RVs drawn from distributions $F_{l,i}$, $1 \le i \le K$; if a request does not require resources of type i then $B_n^{(l,i)} = 0$ almost surely (a.s.), $-\infty < n < \infty$. Only if there is enough capacity available, will the request arriving at time τ_n be accepted and all of the engaged resources occupied for the duration of $\theta_n^{(J_n)}$; otherwise, the request is rejected.

Our goal is to estimate the blocking probability in the system described above. Let $Q_n^{(1)}, \ldots, Q_n^{(K)}$ denote the amounts of resources of each type that a request arriving at time τ_n finds engaged. Note that $Q_n^{(i)}$, $1 \le i \le K$, are mutually dependent and, as pointed out in the introduction, it is hard to compute the blocking probability of this system explicitly. Applying the same arguments used for the case of a single resource type (see the Appendix of [19]), we can show that the stationary distribution of $Q_n^{(i)}$, $1 \le i \le K$, exists. The probability that the request arriving at time τ_n is blocked equals

$$\mathrm{P}\Bigg[\bigcup_{1\le i\le K} \{Q_n^{(i)} + B_n^{(J_n,i)} > C_i\}\Bigg], \tag{7}$$

and again our goal is to estimate its value as $\min_i C_i$ grows large.

Asymptotic estimates derived in this section hold under the following assumptions.

https://doi.org/10.1239/jap/1197908827 Published online by Cambridge University Press

Assumption 1. *For each resource type $1 \le i \le K$, let $\mathcal{L}_i$ and $\mathcal{H}_i$ be two disjoint sets of request types ($|\mathcal{L}_i \cup \mathcal{H}_i| = M$) satisfying the following.*

- *At least one resource type is accessed by subexponentially distributed resource requirements, implying that $|\mathcal{H}_i| > 0$ for some $1 \le i \le K$.*
- *For every $l \in \mathcal{H}_i \ne \varnothing$, there exists a subexponential distribution $F_i \in \mathcal{S}$ such that $\bar{F}_{l,i}(x) \sim c_{l,i}\bar{F}_i(x)$ as $x \to \infty$ with $c_{l,i} > 0$.*
- *There exists a subexponential random variable $L \in \mathcal{S}$ that satisfies*

$$\mathrm{P}[L > x] \ge \max_{1 \le i \le K, l \in \mathcal{L}_i} \mathrm{P}[B_n^{(J_n,i)} > x \mid J_n = l] \quad \text{for all } x > 0$$

and $\mathrm{P}[L > x] = o(\bar{F}_i(x))$ as $x \to \infty$ for all $i \in \{j \mid \mathcal{H}_j \ne \varnothing\}$.

Remark. In Assumption 1 we require the resource requirement distributions to be asymptotically comparable. For each $1 \le i \le K$, $\mathcal{H}_i$ contains tail-dominant subexponential distributions that are asymptotically proportional to each other. Conversely, the only assumption imposed on the distributions in $\mathcal{L}_i$, $1 \le i \le K$, is that there is a subexponential tail that asymptotically dominates them. This asymptotic tail comparability is necessary for our main result to hold. In particular, these conditions are extensively used in (13)–(18) of the proof of Theorem 2, below.

Next we prove the following lemma, which investigates summations of RVs with different tail distributions.

Lemma 1. *Suppose that X_i, $1 \le i \le n$, are independent RVs with corresponding tail distributions $\bar{F}_i(x)$, $1 \le i \le n$. If there exists $F \in \mathcal{S}$ such that $\bar{F}_i(x) \sim c_i\bar{F}(x)$ as $x \to \infty$ with $c_i \ge 0$, $1 \le i \le n$, and $\sum_{i=1}^n c_i > 0$, then the following asymptotic relation holds:*

$$\mathrm{P}\left[\sum_{i=1}^n X_i > x,\ X_i \le x,\ 1 \le i \le n\right] = o(\bar{F}(x)) \quad \text{as } x \to \infty. \tag{8}$$

Proof. Note that

$$\mathrm{P}\left[\sum_{i=1}^n X_i > x\right] = \mathrm{P}\left[\sum_{i=1}^n X_i > x,\ X_i \le x,\ 1 \le i \le n\right] + \mathrm{P}\left[\sum_{i=1}^n X_i > x,\ \bigcup_{i=1}^n \{X_i > x\}\right].$$

Then, the previous expression, the fact that $\bigcup_{i=1}^n \{X_i > x\} \subset \{\sum_{i=1}^n X_i > x\}$, the independence of the X_is, as well as Lemmas 4.2 and 4.5 of [2], imply (8).

First we estimate the asymptotic lower bound for the expression in (7). Using our model assumptions, $\{B_n^{(J_n,i)} > C_i\} \subset \{Q_n^{(i)} + B_n^{(J_n,i)} > C_i\}$ and independence, we obtain

$$\mathrm{P}\left[\bigcup_{1 \le i \le K} \{Q_n^{(i)} + B_n^{(J_n,i)} > C_i\}\right] \ge \mathrm{P}\left[\bigcup_{1 \le i \le K} \{B_n^{(J_n,i)} > C_i\}\right] \sim \sum_{i=1}^K \sum_{l \in \mathcal{H}_i} p_l \bar{F}_{l,i}(C_i) \tag{9}$$

as $\min_i C_i \to \infty$.

https://doi.org/10.1239/jap/1197908827 Published online by Cambridge University Press

Next we estimate the asymptotic upper bound for the expression in (7). Using the union bound, we obtain

$$\mathrm{P}\Big[\bigcup_{1\le i\le K}\{Q_n^{(i)}+B_n^{(J_n,i)}>C_i\}\Big]\le\sum_{i=1}^{K}\mathrm{P}[Q_n^{(i)}+B_n^{(J_n,i)}>C_i].$$

Similarly as in (6) of Theorem 1, for each resource $1\le i\le K$,

$$\mathrm{P}[Q_n^{(i)}+B_n^{(J_n,i)}>C_i]\le\mathrm{P}\Big[\sum_{l\in\mathcal{L}_i}\sum_{j\in\mathcal{N}_{s,n}^{(l,C_i)}}B_j^{(l,i)}+\sum_{l\in\mathcal{H}_i}\sum_{j\in\mathcal{N}_{s,n}^{(l,C_i)}}B_j^{(l,i)}+B_n^{(J_n,i)}>C_i\Big],\quad(10)$$

where $\mathcal{N}_{s,n}^{(l,C_i)}$, $1\le l\le M$, are sets of indices $j<n$ defined as

$$\mathcal{N}_{s,n}^{(l,C_i)}\triangleq\{j<n\mid J_j=l,\ B_j^{(l,i)}\le C_i,\ \theta_j^{(l)}>\tau_n-\tau_j\}.$$

In the previous expressions we bounded the amount of allocated resources that are active at time τ_n by the corresponding quantity in another system of infinite capacity, where the corresponding request process is sampled from the original $\{B_n^{(J_n,i)}\}$, $1\le i\le K$, whenever the corresponding requirements are less than or equal to C_i, $1\le i\le K$.

In the rest of the proof we derive an asymptotic estimate for the expression in (10). After conditioning on $\{N_{s,n}^{(1,C_i)}=n_1,\dots,N_{s,n}^{(M,C_i)}=n_M\}$, $(N_{s,n}^{(l,C_i)}\triangleq|\mathcal{N}_{s,n}^{(l,C_i)}|,\ 1\le l\le M)$, we obtain

$$\begin{aligned}&\mathrm{P}[Q_n^{(i)}+B_n^{(J_n,i)}>C_i]\\&\quad\le\sum_{0\le n_1,\dots,n_M<\infty}\mathrm{P}[N_{s,n}^{(1,C_i)}=n_1,\dots,N_{s,n}^{(M,C_i)}=n_M]\\&\qquad\times\mathrm{P}\Big[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l}B_{(j)}^{(l,i)}+\sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l}B_{(j)}^{(l,i)}+B_n^{(J_n,i)}>C_i,\ B_{(j)}^{(l,i)}\le C_i,\\&\qquad\qquad 1\le j\le n_l,\,1\le l\le M\Big],\end{aligned}\quad(11)$$

where $B_{(j)}^{(l,i)}\overset{\mathrm{D}}{=}B_k^{(l,i)}$, $k\in\mathcal{N}_{s,n}^{(l,C_i)}$, $j=1,\dots,n_l$, are independent replicas of requests in $\mathcal{N}_{s,n}^{(l,C_i)}$ (where '$\overset{\mathrm{D}}{=}$' denotes equality in distribution). Next, after conditioning on $\{J_n=m\}$, $m=1,\dots,M$, and then on $B_n^{(m,i)}$ being smaller or larger than C_i, we can further upper bound the conditional blocking probability in (11) as

$$\begin{aligned}&\mathrm{P}\Big[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l}B_{(j)}^{(l,i)}+\sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l}B_{(j)}^{(l,i)}+B_n^{(J_n,i)}>C_i,\ B_{(j)}^{(l,i)}\le C_i,\ 1\le j\le n_l,\ 1\le l\le M\Big]\\&\quad\le\sum_{m=1}^{M}p_m\,\mathrm{P}\Big[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l}B_{(j)}^{(l,i)}+\sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l}B_{(j)}^{(l,i)}+B_n^{(m,i)}>C_i,\ B_{(j)}^{(l,i)}\le C_i,\\&\qquad\qquad 1\le j\le n_l,\ 1\le l\le M,\ B_n^{(m,i)}\le C_i\Big]\\&\quad+\sum_{m=1}^{M}p_m\,\mathrm{P}[B_n^{(m,i)}>C_i].\end{aligned}\quad(12)$$

https://doi.org/10.1239/jap/1197908827 Published online by Cambridge University Press

Thus, the probabilities in the first term on the right-hand side of (12) can be expressed in the form

$$\mathrm{P}\Bigg[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l'} B_{(j)}^{(l,i)} + \sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l'} B_{(j)}^{(l,i)} > C_i,\ B_{(j)}^{(l,i)} \le C_i,\ 1\le j\le n_l',\ 1\le l\le M\Bigg], \tag{13}$$

where $n_l' = n_l$ for $l \neq m$ and $n_l' = n_l + 1$ for $l = m$.

Next, in order to estimate the asymptotic upper bound of the term in (13), Assumption 1 enables us to distinguish between the following two cases: (i) $\mathcal{H}_i = \varnothing$ or $\sum_{l\in\mathcal{H}_i} n_l' = 0$, and (ii) $\mathcal{H}_i \neq \varnothing$ and $\sum_{l\in\mathcal{H}_i} n_l' > 0$.

Case (i): If $\mathcal{H}_i = \varnothing$ or $\sum_{l\in\mathcal{H}_i} n_l' = 0$, we find that the probability in (12) can be upper bounded as

$$\mathrm{P}\Bigg[\sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l'} B_{(j)}^{(l,i)} > C_i\Bigg] \le \mathrm{P}\Bigg[\sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l'} L_{(j)}^{(l,i)} > C_i\Bigg],$$

where we have used Assumption 1 and introduced $L_{(j)}^{(l,i)}$ to be independent RVs equal in distribution to L. Hence, since the $L_{(j)}^{(l,i)}$ are subexponential, we obtain

$$\lim_{C_i\to\infty} \frac{\mathrm{P}\big[\sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l'} B_{(j)}^{(l,i)} > C_i\big]}{\mathrm{P}[L > C_i]} \le \sum_{l\in\mathcal{L}_i} n_l'. \tag{14}$$

Case (ii): If $\mathcal{H}_i \neq \varnothing$ and $\sum_{l\in\mathcal{H}_i} n_l' > 0$, using Assumption 1 and Lemma 1, we derive the following asymptotic upper bound:

$$\mathrm{P}\Bigg[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l'} B_{(j)}^{(l,i)} + \sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l'} B_{(j)}^{(l,i)} > C_i,\ B_{(j)}^{(l,i)} \le C_i,\ 1\le j\le n_l',\ 1\le l\le M\Bigg] = o(\bar{F}_i(C_i)) \tag{15}$$

as $C_i \to \infty$.

Thus, in (13)–(15) we obtained upper bounds and their asymptotic estimates for the conditional blocking probabilities in the first term of (12) that hold for any finite nonnegative integers $n_1, \ldots, n_M$. Thus, in view of (11), in order to estimate an asymptotic upper bound of $\mathrm{P}[Q_n^{(i)} + B_n^{(J_n,i)} > C_i]$, we need to integrate the probabilities in (13) with respect to the densities of the RVs $N_{s,n}^{(l,C_i)}$, $l = 1, \ldots, M$. In this regard, note that in the case in which $\mathcal{H}_i \neq \varnothing$, by Assumption 1, the term in (13) can be upper bounded as

$$\begin{aligned}&\mathrm{P}\Bigg[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l'} B_{(j)}^{(l,i)} + \sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l'} B_{(j)}^{(l,i)} > C_i,\ B_{(j)}^{(l,i)} \le C_i,\ 1\le j\le n_l',\ 1\le l\le M\Bigg]\\ &\qquad\le \mathrm{P}\Bigg[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l'} B_{(j)}^{(l,i)} + \sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l'} L_{(j)}^{(l,i)} > C_i\Bigg],\end{aligned} \tag{16}$$

where, as before, the $L_{(j)}^{(l,i)}$ are independent RVs equal in distribution to L. Furthermore, since $\mathrm{P}[L > x] = o(\bar{F}_i(x))$ as $x \to \infty$, there exists a large enough finite integer H such that

https://doi.org/10.1239/jap/1197908827 Published online by Cambridge University Press

$\mathrm{P}[L > x] \le H\bar{F}_i(x)$ for all $x \ge 0$. Therefore, for any $x \ge 0$, we can write

$$\mathrm{P}[L > x] \le H\bar{F}_i(x) = \mathrm{P}\Big[\bigcup_{1\le r\le H}\{\hat{B}_r^{(i)} > x\}\Big] \le \mathrm{P}\Big[\sum_{r=1}^{H}\hat{B}_r^{(i)} > x\Big], \tag{17}$$

where $\hat{B}_r^{(i)}$, $1 \le r \le H$, are independent RVs having cumulative distribution function F_i. Now, in view of (17), each of the RVs $L_{(j)}^{(l,i)}$ in (16) can be stochastically upper bounded by a random variable that is equal in distribution to $\sum_{r=1}^{H}\hat{B}_r^{(i)}$. Thus, if we introduce Y_j, $j \ge 1$, to be independent RVs equal in distribution to $\sum_{r=1}^{H}\hat{B}_r^{(i)}$, we obtain

$$\mathrm{P}\Big[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l'}B_{(j)}^{(l,i)} + \sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l'}L_{(j)}^{(l,i)} > C_i\Big] \le \mathrm{P}\Big[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l'}B_{(j)}^{(l,i)} + \sum_{j=1}^{\sum_{l\in\mathcal{L}_i}n_l'}Y_j > C_i\Big],$$

which in conjunction with point (b) of Lemma 4.2 of [2] implies that, for any $\varepsilon > 0$, there exists a finite constant K_ε such that

$$\begin{aligned}\mathrm{P}\Big[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l'}B_{(j)}^{(l,i)} + \sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l'}L_{(j)}^{(l,i)} > C_i\Big] &\le \mathrm{P}\Big[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l'}B_{(j)}^{(l,i)} + \sum_{j=1}^{\sum_{l\in\mathcal{L}_i}n_l'}Y_j > C_i\Big]\\ &\le K_\varepsilon(1+\varepsilon)^{\sum_{l\in\mathcal{H}_i}n_l'+\sum_{l\in\mathcal{L}_i}n_l'}\bar{F}_i(C_i)\end{aligned} \tag{18}$$

for any $C_i < \infty$. Similarly, in cases where $\mathcal{H}_i = \varnothing$, we could apply the stochastic dominance

$$B_{(j)}^{(l,i)} \overset{d}{\le} L_{(j)}^{(l,i)}, \qquad l \in \mathcal{L}_i,$$

where the $L_{(j)}^{(l,i)}$ are, as before, independent, subexponential RVs equal in distribution to L. Then, by Kesten's lemma (see [3, Lemma 7]), the analogous bound to the one in (18) follows.

Finally, since (18) bounds uniformly probabilities in (13) for all $C_i < \infty$ and n_l', $1 \le l \le M$, in conjunction with (12), (11), $N_{s,n}^{(l,C_i)} \le N_n^{(l,\infty)}$ a.s., and the existence of $\mathrm{E}\,z^{N_n^{(l,\infty)}}$ for all $z \in \mathbb{C}$, $1 \le l \le M$, (see [17, Theorem 5] and [22, Theorem 1]), we can apply the dominated convergence theorem and conclude that

$$\lim_{C_i\to\infty}\frac{\mathrm{P}[Q_n^{(i)} + B_n^{(J_n,i)} > C_i]}{\sum_{l\in\mathcal{H}_i}p_l\bar{F}_{l,i}(C_i)} \le \mathbf{1}[\mathcal{H}_i \ne \varnothing].$$

Next, by adding asymptotic estimates for all $1 \le i \le K$, in conjunction with (9), we complete the proof of the following result.

Theorem 2. *For the request model introduced in this section, under the conditions imposed by Assumption 1, the stationary blocking probability for general loss networks satisfies*

$$\mathrm{P}\Big[\bigcup_{1\le i\le K}\{Q_n^{(i)} + B_n^{(J_n,i)} > C_i\}\Big] \sim \sum_{i=1}^{K}\sum_{l\in\mathcal{H}_i}p_l c_{l,i}\bar{F}_i(C_i) \quad \textit{as } \min_i C_i \to \infty.$$

https://doi.org/10.1239/jap/1197908827 Published online by Cambridge University Press

4. Numerical examples

In this section, with two simulation experiments, we demonstrate the accuracy of our asymptotic formulae, proved in Theorems 1 and 2. Our goal is to show that even though our results are asymptotic, the derived estimates match experiments with high accuracy even for systems with finite support demand distributions and moderately large capacities.

In each experiment, in order for the system to reach stationarity, we let the first 10^8 arrivals be a warm-up time. By repeating many experiments we observed that longer warm-up times did not lead to improved results. Then we counted the number of blocked requests among the next 10^9 arrivals. In both of the experiments below, measurements were conducted for capacities $C = 500 + 100j$, $0 \le j \le 9$, where the starting value of $C = 500$ was set to be slightly larger than the effective system's load $\lambda \, \mathrm{E}[\theta_n] \, \mathrm{E}[B_n]$. In Figures 1 and 2 simulation results are represented by circular data points, while our approximations, estimates obtained in Theorems 1 and 2, are represented by solid lines. In order to emphasize the difference between the simulations and approximation formulae, we present base 10 logarithms of the obtained values.

Example 1. Consider the case of a single resource type of capacity C. Let requests for resources arrive at Poisson time points with rate $\lambda = 1$. In addition, assume that engagement durations are exponentially distributed with mean $1/\mu = 1$. Next, let request requirements B_n be drawn from a finite support distribution, where $\mathrm{P}[B_n = i] = 0.3/i^{1.5}$, $1 \le i \le 1999$, and $\mathrm{P}[B_n = 2000] = 1 - \mathrm{P}[B_n < 2000]$ (a power law distribution). The effective load in this example is $\lambda \, \mathrm{E}[\theta_n] \, \mathrm{E}[B_n] \approx 485.8$. Experimental results are presented in Figure 1. Even though we start measuring rejections at capacities that are slightly larger than the mean requirement value, our approximation $\mathrm{P}[B_n > C]$ is very close to the experimental results. In particular, the relative approximation error is less than 1% for $C = 500$ and, for capacity values larger than or equal to $C = 1400$, this error is less than 0.3%.

Example 2. In this example we consider the case of two resource and two request types. Furthermore, we assume that the resource capacities are the same, i.e. $C = C_1 = C_2$. The

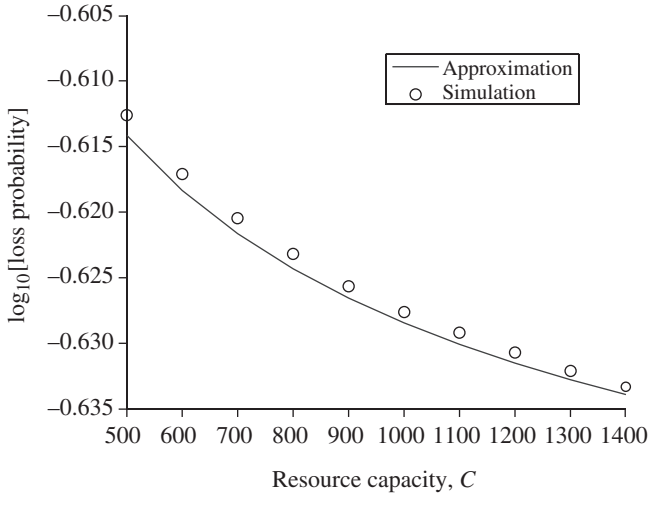

FIGURE 1: Illustration for Example 1.

https://doi.org/10.1239/jap/1197908827 Published online by Cambridge University Press

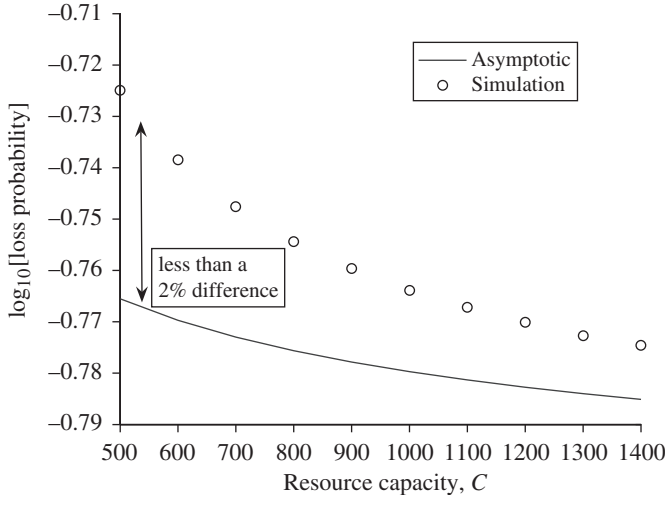

FIGURE 2: Illustration for Example 2.

frequencies of requests of types 1 and 2 are $p_1 = 0.3$ and $p_2 = 0.7$, respectively. Assume that the arrival points are separated by a fixed, unit length of time, i.e. $\tau_n - \tau_{n-1} = 1$ for all n. Type 1 request durations satisfy $\theta_i^{(1)} \sim \exp(4)$ and type 2 request holding times are drawn from the uniform distribution on $[0, 40]$, i.e. $\theta_i^{(2)} \sim \text{Unif}([0, 40])$. Resource requirements corresponding to engagements of type 1 are distributed as

$$\mathrm{P}[B^{1,1} = 1] = 0.8, \qquad \mathrm{P}[B^{1,1} = i] = 0.15\mathrm{e}^{-\sqrt{i}}, \quad 2 \le i \le 1999,$$

and

$$\mathrm{P}[B^{1,1} = 2000] = 1 - \mathrm{P}[B^{1,1} < 2000] \quad \text{for type 1 resources,}$$
$$\text{and} \quad \mathrm{P}[B^{1,2} = 50] = 1 \quad \text{for type 2 resources.}$$

Requests of type 2 require resources according to

$$\mathrm{P}[B^{2,1} = i] = \mathrm{geom}^{i-1}(1 - \mathrm{geom}), \qquad 1 \le i \le 1999$$
$$\text{and} \quad \mathrm{P}[B^{2,1} = 2000] = 1 - \mathrm{P}[B^{2,1} < 2000],$$

where geom $= 0.6$ for type 1 resources and

$$\mathrm{P}[B^{2,2} = i] = 0.3/i^{1.5}, \quad 1 \le i \le 1999, \qquad \mathrm{P}[B^{2,2} = 2000] = 1 - \mathrm{P}[B^{2,2} < 2000]$$

for type 2 resources. Our asymptotic results suggest that the blocking probability should be characterized by the heaviest-tailed demand distributions. The results of this experiment are presented in Figure 2. As in the previous case, we obtain a very accurate agreement between our approximation and the simulation. The relative approximation error in this case does not exceed 2% and becomes smaller as the resource capacities grow.

Remark. (i) We would like to point out that the accuracy of experimental results directly depends on the approximation errors (6) and (15), depending on the simulated scenarios. These

https://doi.org/10.1239/jap/1197908827 Published online by Cambridge University Press

errors highly depend on the tail properties of the resource requirement distributions. More specifically, under fairly general assumptions, the heavier the dominant tail of the resource requirement distribution is, the smaller the relative approximation error is. For detailed explanations, the reader is referred to [5, Section 1.3.2].

(ii) Note that our main results estimate the stationary blocking probability and, as we commented earlier, are indifferent to distributional properties of holding times. For this reason, as long as we can claim that the measurements are conducted in stationarity, the transience should not affect experimental results.

5. Concluding remarks

In this paper we have considered loss networks with reusable resources and finite resource capacities and estimated the probability that a request is rejected due to an insufficient amount of resources at points of their arrivals. Assuming a renewal process of request arrivals, subexponential resource requirements, and generally distributed activity durations, we have shown that the asymptotic blocking probability for a wide class of analyzed systems can be fully estimated using a resource requirement distribution, independent of the other system's properties. In particular, we have shown that the blocking probability behaves as the asymptotically dominant tail of the resource requirement distribution.

The model we have studied can be applied to a wide range of applications. Historically, loss networks (in particular Erlang loss networks) are widely used for modeling communication networks. Later, through the development of new service applications such as workforce management with similar modeling properties, the importance of accurately estimating blocking probabilities of general loss networks has become significant. In this regard, we have investigated loss networks with various request types and possibly highly variable random amounts of required resources. In addition, we have researched the possibility of incorporating random-advance reservations for incoming requests. These results should be of great interest to an emerging research community. Although our results are intended mainly for qualitative purposes, numerical examples demonstrate an excellent match between derived formulae and simulated systems performance, hence, strongly suggesting their application.

Acknowledgement

The authors would like to thank Professor Predrag Jelenković for valuable suggestions related to the possible generalizations of this work.

References

[1] Asmussen, S. and Pihlgård, M. (2007). Loss rates for Lévy processes with two reflecting barriers. *Math. Operat. Res.* **32,** 308–321.
[2] Asmussen, S., Henriksen, L. F. and Klüppelberg, C. (1994). Large claims approximations for risk processes in a Markovian environment. *Stoch. Process. Appl.* **54,** 29–43.
[3] Athreya, K. B. and Ney, P. E. (1972). *Branching Processes*. Springer, New York.
[4] Bonald, T. (2006). The Erlang model with non-Poisson call arrivals. In *Proc. Joint Internat. Conf. Measurement Modeling Comput. Systems*, ACM Press, New York, pp. 276–286.
[5] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events. For Insurance and Finance*. Springer, Berlin.
[6] Embrechts, P. and Goldie, C. M. (1980). On closure and factorization properties of subexponential and related distributions. *J. Austral. Math. Soc. Ser. A* **29,** 243–256.
[7] Erlang, A. K. (1917). Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Elektrotkeknikeren* **13,** 5–13.

[8] Gazdzicki, P., Lambadaris, I. and Mazumdar, R. R. (1993). Blocking probabilities for large multirate Erlang loss systems. *Adv. Appl. Prob.* **25,** 997–1009.
[9] Goldie, C. M. and Klüppelberg, C. (1998). Subexponential distributions. In *A Practical Guide to Heavy Tails: Statistical Techniques for Analysing Heavy Tailed Distributions*, eds M. T. R. Adler and R. Feldman, Birkhäuser, Boston, MA, pp. 435–459.
[10] Heyman, D. P. and Lakshman, T. V. (1996). Source models for VBR broadcast-video traffic. *IEEE/ACM Trans. Networking* **4,** 40–48.
[11] Jelenković, P. R. (1999). Subexponential loss rates in a GI/GI/1 queue with applications. *Queueing Systems* **33,** 91–123.
[12] Jelenković, P. R., Lazar, A. A. and Semret, N. (1997). The effect of multiple time scales and subexponentiality of MPEG video streams on queueing behavior. *IEEE J. Selected Areas Commun.* **15,** 1052–1071.
[13] Kaufman, J. S. (1981). Blocking in a shared resources environment. *IEEE Trans. Commun.* **29,** 1474–1481.
[14] Kelly, F. P. (1986). Blocking probabilities in large circuit-switched networks. *Adv. App. Prob.* **18,** 473–505.
[15] Kelly, F. P. (1991). Loss networks. *Ann. Appl. Prob.* **1,** 319–378.
[16] Krishnan, K. R. and Meempat, G. (1997). Long-range dependence in VBR video streams and atm traffic engineering. *Performance Evaluation* **30,** 46–56.
[17] Liu, L., Kashyap, B. R. K. and Templeton, J. G. C. (1990). On the $GI^X/G/\infty$ system. *J. Appl. Prob.* **27,** 671–683.
[18] Louth, G., Mitzenmacher, M. and Kelly, F. (1994). Computational complexity of loss networks. *Theoret. Comput. Sci.* **125,** 45–59.
[19] Lu, Y. and Radovanović, A. (2007). Asymptotic blocking probabilities in loss networks with subexponential demands. Preprint. Available at http://arxiv.org/abs/0708.4059.
[20] Lu, Y., Radovanović, A. and Squillante, M. (2006). Workforce management through stochastic network models. In *Proc. IEEE Conf. Service Operat. Logistics* (Shanghai, June 2006).
[21] Sevastyanov, B. A. (1957). An ergodic theorem for Markov processes and its application to telephone systems with refusals. *Theory Prob. Appl.* **2,** 104–112.
[22] Takacs, L. (1980). Queues with infinitely many servers. *RAIRO Rech. Opér.* **14,** 109–113.
[23] Whitt, W. (1985). Blocking when service is required from several facilities simultaneously. *AT&T Tech. J.* **64,** 1807–1856.
[24] Wilkinson, R. I. (1956). Theory of toll traffic engineering in the USA. *Bell Syst. Tech. J.* **35,** 421–513.
[25] Wolff, R. W. (1989). *Stochastic Modeling and Theory of Queues*. Prentice Hall, Englewood Cliffs, NJ.
[26] Zachary, S. (1991). On blocking in loss networks. *Adv. Appl. Prob.* **23,** 355–372.