Mathematical Problems in Engineering, 2002, Vol. 8(4–5), pp. 323–348

Taylor & Francis
Taylor & Francis Group

Asymptotic Expansions for Large Closed and Loss Queueing Networks

YAAKOV KOGAN

AT&T Labs, Middletown, NJ 07748, USA

(Received 30 January 2002; In final form 7 May 2002)

Loss and closed queueing network models have long been of interest to telephone and computer engineers and becoming increasingly important as models of data transmission networks. This paper describes a uniform approach that has been developed during the last decade for asymptotic analysis of large capacity networks with product form of the stationary probability distribution. Such a distribution has an explicit form up to the normalization constant, or the partition function. The approach is based on representing the partition function as a contour integral in complex space and evaluating the integral using the saddle point method and theory of residues. This paper provides an introduction to the area and a review of recent work.

Key words: Asymptotic expansions; Partition function; Loss networks; Closed queuing networks

1 INTRODUCTION

Many problems in design of large computer systems, voice and data network with randomly fluctuating demand require analysis of queueing network models consisting of many service stations. In this paper we describe work on asymptotic analysis of two particular classes of networks, called loss networks (LN) and closed queueing networks (CQN). Our aim is to present a unified approach to asymptotic analysis of these networks based on integral representations in complex space. This paper provides an introduction to the area and a review of recent work.

1.1 The Historical Context

A loss network is a generalization of the famous model of a telephone system published by the Danish mathematician Erlang in 1917 (see [10], p. 139). In the Erlang model, calls arrive at a link as a Poisson process of rate v. The link comprises of M circuits and a call is blocked and lost if all M circuits are busy. Otherwise the call is accepted and holds a single circuit for a random period of time. Call holding periods are independent of each other and of arrival times and are identically distributed with unit mean. Then Erlang's loss formula

$$E(v, M) = \frac{1}{G_L(M)} \frac{v^M}{M!} = 1 - \frac{G_L(M-1)}{G_L(M)}. \tag{1}$$

ISSN 1024-123X print; ISSN 1563-5147 online © 2002 Taylor & Francis Ltd

gives the proportion of calls that are lost, where $G_L(M)$ is the normalization constant (or partition function)

$$G_L(M) = \sum_{n=0}^{M} \frac{v^n}{n!}. \tag{2}$$

The stationary probability that n circuits are busy is given by

$$\pi(n) = \frac{1}{G_L(M)} \frac{v^n}{n!}. \tag{3}$$

By replacing the Poisson arrival stream in the Erlang model by the finite number N sources of calls, we obtain the Engset model [56]. A new call from the same source will be generated after an exponentially distributed idle period with parameter λ. The idle period starts either at completion of service if the previous call was served, or at the time of blocking if the previous call was blocked. Then the stationary probability that n circuits are busy is given by

$$\pi(n) = \frac{1}{G_L(N, M)} \binom{N}{n} \lambda^n,$$

where the normalization constant, or partition function

$$G_L(N, M) = \sum_{n=1}^{M} \binom{M}{n} \lambda^n \tag{4}$$

The loss probability is defined by the last expression in the right-hand side of (1) as before but with $G_L(M)$ replaced by $G_L(N, M)$ given by (4).

Loss networks are generalizations of the Erlang and Engset models. First, one can introduce different services. Calls of service $j, j = 1, \dots, C$ are generated by its own Poisson arrival stream or finite source of size N_j, they require b_j circuits and hold them for the holding period of call with mean $1/\mu_j$. If a service j request does not find b_j free circuits then it is lost. The state of the system is described by a vector $(n_1, \dots, n_C)$, where n_j is the number of service j calls being served. In [11] another generalization of the Engset model is introduced where sources require random number of circuits. Second, in addition to different services, a network with more than one link is considered [27]. Link l comprises M_l circuits, $l = 1, \dots, L$. Let $\mathcal{R}_j$ be the set of all routes for calls of service $j, j = 1, \dots, C$. A route $r \in \mathcal{R}_j$ identifies the subset of links $\mathcal{L}(r) \in \{1, \dots, L\}$ from which simultaneous service is required. A call on route r requires b_{lr} circuits from link l. A call requesting route r is blocked and lost if on any link $l \in \mathcal{L}(r)$ there are fewer than b_{lr} circuits free. Otherwise the call is connected and simultaneously holds all the necessary circuits on route r for the holding period of the call. Let n_r be the number of calls in progress on route r. The state of the network is described by a vector $\mathbf{n}' = (n_1, \dots, n_R)$ where R is total number of routes and the prime denotes transpose.

We describe next the class of closed queueing networks. Closed queueing networks are generalizations of another famous model originally studied by Khinchine [28] and Palm [50]. This model known in literature as the *machine servicing* model [19] or the *machine interference* model [15] in which a server (repairman) is assigned the responsibility of maintaining a group of N machines. Each machine is in one of two states: either "up" (running) or "down" (requiring repair service). When a machine breaks down, it joins the queue for repair. The repair of the first machine in the queue starts as soon as the server is free.

A relatively recent interest in this model is explained by the fact that it can be viewed as a general model of a multiaccess system [33]. In the simplest case

(a) on any one machine, breaks occur completely randomly in running time, at the same rate λ for all machines and independently for all machines;
(b) the distribution of repair time is exponential with mean $1/\mu$. The repair times for different breaks are independent, and the repair times are independent of the number of machines awaiting for service.

The ratio $\rho = \lambda/\mu$ is called the servicing intensity. Under these assumption the stationary probability that n machines are broken can be written as

$$P(n) = P(0)\frac{N!}{(N-n)!}\rho^n \tag{5}$$

and

$$G_C(N) = \frac{1}{P(0)} = N!\sum_{n=0}^{N}\frac{\rho^n}{(N-n)!} \tag{6}$$

is the normalization constant (or partition function) for the machine interference model.

A single-chain CQN with N customers is obtained from the machine interference model by adding more servers and defining the routing probability matrix. In the direct generalization of machine interference model a broken machine is directed for service to server i with service rate $\mu_i, i, i = 1, \dots, K$ with probability p_i, where $\sum_i p_i = 1$. The state of the network is described by a vector $(n_1, \dots, n_K)$ where n_i is the number of machines (called customers in a general case) including those in service at server i. A more general multichain CQN is obtained from the machine interference model by having J types of machines (customers). There are N_j machines with breakdown rate λ_j in group j, $j = 1, \dots, J$. The state of the multichain network is described by a matrix $\mathbf{n}$ with elements $n_{ji}, j = 1, \dots, J$; $i = 1, \dots, K$ representing the number of type j customers (including those in service) at server i. Closed queueing networks have been originally motivated by job-shop type systems [23]. The advent of multiprogramming computers and computer networks sparked off new interest to them [26], with the result that studies of CQN models have multiplied in the last 30 years.

1.2 Large Capacity Networks

A large class of loss and closed queueing networks has the so-called "product-form" solution: the stationary distribution of the network state decomposes completely into a product of individual node functions and the normalization constant (see [26, 27]). We are interested in asymptotic behavior of these product-form networks when their capacity defined by parameters M_l and N_r increases. To make the problem nontrivial we assume that offered traffics (parametrized by Poisson arrival rates ν_r for route r or by the number of request sources N_j) and service rates μ_i also increase with ratios M_l/ν_r (or M_l/N_j) and N_r/μ_i respectively held fixed.

There are four main approaches to the asymptotic analysis of the product-form stationary distributions with finite number of states. A direct approach is based on observation that for a large network the product-form distribution has the following asymptotic representation:

$$P(\mathbf{n}) \sim C \exp\{NF(\mathbf{x})\},$$

where N is a large parameter (*e.g.*, the total number of customers in the network) and $\mathbf{x} = \mathbf{n}/N$. Then the problem of finding the most likely state $\mathbf{n}$ is reduced to maximizing the function $F(\mathbf{x})$ under natural constraints, and a maximizing value of $\mathbf{x}$ can be found by Lagrangian methods. If the maximum point $\mathbf{x}^*$ is unique then simple approximations (*e.g.*, by Gaussian and Poisson distributions) for $P(\mathbf{n})$ the can be derived. This approach has been introduced by Pittel [53] in 1979 in the context of CQN and later Kelly (see survey [27]) applied it to LN. More accurate asymptotic approximations can be obtained using three other approaches based on singular perturbation methods and integral representations in real and complex space. The application of singular perturbation methods to queueing networks has been developed by Knessl *et al.* [29, 30] and Kness and Tier [31]. These methods are applied to the forward Kolmogorov equation for the probability distribution or recursions for the partition function, and, in general, their application does not require the product form solution. In contrast with the singular perturbation techniques, the methods based on integral representations can be applied to a relatively narrow subset of product-form networks, where the probability distribution or the partition function can be expressed through Laplace or contour integrals with explicit integrands. The main advantage of such explicit representations is significant simplification in derivations of asymptotic expansions.

Integral representations in real space provide an easy way for asymptotic analysis of Erlang and machine interference models. These representations are based on the Euler formula $n! = \int_0^\infty t^n e^{-t}\, \mathrm{d}t = \Gamma(n+1)$. An integral representation for $E(v, M)$, ascribed to Fortet [56], has the following form [25]:

$$E(v, M)^{-1} = v \int_0^\infty e^{-vy}(1+y)^M \,\mathrm{d}y = vI(M, v). \tag{7}$$

Assume that $v = cM$, where c is fixed while $M \to \infty$. Then the integral

$$I(M, v) = I(M, c) = \int_0^\infty e^{-M\phi(y)} \,\mathrm{d}y, \tag{8}$$

where $\phi(y) = cy - \ln(1+y)$, can be evaluated by Laplace's method [17]. We see that $\phi'' = 1/(1+y)^2 > 0$. Hence $\phi(y)$ has only one minimum on $(-\infty, \infty)$ and the minimum point $y^* = 1/c - 1$ is the solution of equation $\phi'(y) = 0$. Let y_0 be the minimum point of the function $\phi(y)$ on the positive semiaxis $[0, \infty)$. Then

$$y_0 = \begin{cases} y^* & \text{if } c \le 1 \text{ and } y^* > 0 \text{ if } c < 1 \\ 0 & \text{if } c \ge 1 \end{cases}$$

The first order approximation for $I(M, c)$ is obtained by expanding $\phi(y) - \phi(y_0)$ in the Taylor series at point y_0 up to the first non-zero term and then performing the integration taking into account that the main contribution to the integral comes from the vicinity around y_0. Depending on values of parameter c, there are three following asymptotic approximations for underloaded, critically loaded and overloaded regimes corresponding to $c < 1, c = 1$ and $c > 1$ respectively.

(a) If $c < 1$ then

$$I(M,c) = \frac{1}{c}\sqrt{\frac{2\pi}{M}}e^{M(c-1-\ln c)}(1 + O(M^{-1})) \tag{9}$$

and by substituting (9) in (7) with $v = cM$ we have

$$E(v,M) = \frac{1}{\sqrt{2\pi M}}e^{M(1-c+\ln c)}(1 + O(M^{-1})). \tag{10}$$

(Note that $1 - c + \ln c = \phi(y^*) < 0$ as $\phi(0) = 0$). This approximation can be directly obtained from (1) by applying Stirling's formula for $M!$ and approximating $G_L(M)$ by e^v (see (6.18) in [56]).

(b) If $c = 1$ then

$$I(M,c) = \sqrt{\frac{\pi}{2M}}(1 + O(M^{-1/2}) \tag{11}$$

and

$$E(v,M) = \sqrt{\frac{2}{\pi M}}(1 + O(M^{-1/2})). \tag{12}$$

(c) If $c > 1$ then

$$I(M,c) = \frac{M^{-1}}{c-1} + O(M^{-2}) \tag{13}$$

and

$$E(v,M) = 1 - \frac{1}{c} + O(M^{-1}). \tag{14}$$

Approximations for $E(v,M)$ have a long history (see [56]). The following normal approximation was used by Erlang [10]:

$$E(v,M) \approx \frac{1}{\sqrt{v}}\frac{e^{-h^2/2}}{\int_{-\infty}^{h} e^{-x^2/2}\,\mathrm{d}x},$$

where $h = (v - M)/\sqrt{v}$. With our assumption $v = cM$, this approximation provides correct asymptotics only for $c \geq 1$. In 1966, Borovkov (see Theorem 15 in Chapter 7 in [9]) derived asymptotic approximations for a generalization of the Erlang model, where interarrival times are i.i.d. random variables with a general distribution. In 1974, Jagerman [25] obtained asymptotic expansions for $E(v,M)^{-1}$ when $c > 1$ and $c = 1$ using integral representation (7) and a theorem on Abelian asymptotics for Laplace transforms. Finally in 1992, Pinsky [51] derived a new "simple" approximation:

$$E(v,M) \approx \exp\left(M\ln\frac{vs}{M} + M - vs\right)\sqrt{\frac{s + vs(1-s)^2}{M}}, \tag{15}$$

where

$$s = \frac{M + v + 1 - \sqrt{(M + v + 1)^2 - 4vM}}{2v}.$$

Denote the right hand side of (15) by $S(v, M)$ and assume that $v = cM$. Then it is easy to see that

$$\lim_{M\to\infty} \frac{E(v, M)}{S(v, M)} = \begin{cases} (2\pi)^{-1/2} & \text{if } c < 1 \\ \left(\frac{2}{\pi}\right)^{-1/2} & \text{if } c = 1 \cdot \\ \infty & \text{if } c > 1 \end{cases}$$

Remarkably for the machine interference model

$$\frac{1}{P(0)} = G_C(N) = \int_0^\infty e^{-t}(1 + \rho t)^N \,\mathrm{d}t = E\left(\frac{1}{\rho, N}\right) = \left(\frac{N}{r}\right) I\left(\frac{N, 1}{r}\right), \tag{16}$$

where the last equality is obtained by assuming $r = \rho N$ and changing variables $y = rt/N$. Hence we have the following simple relation between the repairman utilization $U = 1 - P(0)$ and Erlang's loss function $E(v, M)$:

$$U = 1 - E\left(\frac{1}{\rho}, N\right). \tag{17}$$

Substituting in (17) approximations (10), (12) and (14) we see that utilization is exponentially close to 1 if $r > 1$ (heavy usage), $U = 1 - \sqrt{2/\pi}N^{-1} + O(N^{-1})$ if $r = 1$ (moderate usage) and $U = r + O(N^{-1})$ if $r < 1$ (normal usage). The asymptotic expansions for $G_C(N)$ have been derived in 1971 by Ferdinand [20], who represented the partition function through gamma and incomplete gamma functions and used their asymptotic expansions in [32]. Apparently, authors of the early papers on asymptotic expansions for the Erlang and the machine interference models have been not familiar with Laplace's method. (Approximations (10) and (14) have been obtained by Laplace's method only in 1994 [48].) Neither of them has been aware about simple relation (16) between the two models.

Integral representation (7) is unique in the class of loss networks as its generalizations, except for one special case [44], are not known. In contrast, integral representation (16) can be generalized to single- and multi-chain generalizations of the machine interference model which have been described at the end of Section 1.1. In 1981, McKenna *et al.* [42] used Laplace's method and provided complete asymptotic analysis for a multichain generalizations of the machine interference model with one single server. For a CQN with K single-server nodes the integral representation becomes K-dimensional and its asymptotic expansion has been derived in [43] only in normal usage when the minimum of the function $\phi(\mathbf{x})$ (K-dimensional analog of $\phi(x)$) is at $\mathbf{x} = \mathbf{0}$.

Finally, the last approach to the asymptotic analysis of the product-form stationary distributions with finite number of states utilizes the fact that the generating function of the partition function (later generating partition function) has an explicit expression for many models. This makes it plausible to recover the partition function using the inverse Cauchy formula, which provides an integral representation in complex space. For large capacity networks the Cauchy integral can be transformed into the integral over the saddle point contour and evaluated asymptotically using the classical saddle point theory. In general, the generating partition function has poles, and the above transformation of the original contour into the saddle point contour requires calculation of residues for the poles inside the saddle point contour. This results in three different approximations for the partition function depending on whether the poles of the generating partition function are outside, inside or close to the saddle point contour. The advantage of the method based on integral representations in complex space has been initially demonstrated by the author in 1989–90 [34, 39] for relatively simple classes of LN and CQN. In subsequent author's paper [35] and papers coauthored with Birman [6, 7, 36], Berger [2–4], Choudhury and Susskind [11], Hofri [24], Shenfild [38] and Yakovlev [40], this method has been applied to a wide variety of LN and CQN models, where other methods lead to more complicated derivations or not applicable at all. Integral representations in complex space also stimulated development of new methods for the exact computation of the normalization constant [12, 13] and refined asymptotic expansions for partition functions and probability distributions [2–4, 47].

The outline of the paper is as follows. In Section 2 we present results on asymptotic expansions for one-dimensional partition functions and review their application to bottleneck analysis of single-chain closed queueing networks (including models of multiprocessor systems) and loss systems with a single link. In Section 3 we generalize some of the results of Section 2 for multichain closed queueing networks and loss networks with several links. Section 4 is motivated by the problem of dimensioning bandwidth for high-speed data transmission networks. Two models are considered. The first model is described by the generalized Erlang or Engset model. Asymptotic expansion for the probability distribution of busy circuits is derived in underloaded regime using uniform asymptotic expansion for the partition function. The second model is described by a CQN with multiple customer types that consists of one IS (infinite server) station and many PS (processor-sharing) stations. Asymptotic expansion for the probability distribution is derived for the total number of customers at the saturated PS station.

2 ASYMPTOTICS OF ONE-DIMENSIONAL PARTITION FUNCTIONS AND THEIR APPLICATION

2.1 Integral Representation and the Saddle-Point Method

For single-chain CQN and LN with one link the partition function depends on a single integer which is the number of customers and the number of links respectively. Then the generating partition function (GPF) defined as

$$\mathcal{G}(z) = \sum_{n=0}^{\infty} z^n G(n)$$

is one-dimensional complex function. The partition function can be recovered from the generating function either by differentiation as

$$G(N) = \frac{1}{N!}\frac{\mathrm{d}^N}{\mathrm{d}z^N}\mathcal{G}(z) \tag{18}$$

or by the application of the inverse Cauchy formula [14]

$$G(N) = \frac{1}{2\pi j}\oint_{C_1} \frac{\mathcal{G}(z)}{z^{N+1}}\,\mathrm{d}z, \tag{19}$$

where j is the imaginary unit and C_1 is any contour around the origin which does not contain any singularities of the generating function $\mathcal{G}(z)$.

Representation (18) is used for deriving recursive computational algorithms for the normalization constant [33, 52]. Such derivations, which are based on the product form nature of the stationary distribution, do not require the generating function in explicit form. Representation (19) was originally used by the author [34–36, 39] to find exact and asymptotic approximation for the normalization constant. The same representation has been used later in [12, 13] to develop efficient exact computational algorithms for the normalization constant. The use of representation (19) requires an explicit derivation of the GPF. In fact, such explicit formulas can be obtained under some mild assumptions.

In order to asymptotically evaluate the integral (19) using the saddle point method we represent the integrand as $c(N)q(t)\exp\{Np(t)\}$, where $c(N)$ is a constant whose calculation does not require summation, while the functions $p(t)$ and $q(t)$ do not depend on N. Denote

$$I(N) = \frac{1}{2\pi j}\oint_C q(t)\exp\{-Np(t)\}\,\mathrm{d}t. \tag{20}$$

where C is any contour around the origin, inside which the function $tq(t)$ is analytic. Note that $G(N) = c(N)I(N)$. The asymptotic approximation to the contour integral in (20) can be obtained by the saddle-point method if the following two conditions are satisfied.

1. There is a unique positive solution t_0 of equation $p'(t) = 0$ on the real axis, where the prime denotes derivative.
2. Let $t = ae^{j\omega}$, $0 \leq \omega < 2\pi$. Then

$$\min_a \max_{|t|=a} \operatorname{Re} p(t) = p(t_0) = \max_{|u|=t_0} \operatorname{Re} p(t).$$

This implies [18] that $|t| = t_0$ is a saddle-point contour, and $t = t_0$ is the only saddle point on it.

If $tq(t)$ is analytic inside the circle $|t| = t_0$ then by the saddle-point method [18]

$$\frac{1}{2\pi j}\oint_{|t|=t_0} q(t)\exp\{Np(t)\}\,\mathrm{d}t = \frac{\exp\{Np(t_0)\}}{\sqrt{2\pi Np''(t_0)}}\left[q(t_0) + O\left(\frac{1}{N}\right)\right]. \tag{21}$$

Let $\beta_1 < \cdots < \beta_k$ be the positive poles of $q(t)$ inside the saddle point contour. Then the asymptotic approximation for $I(N)$ has the following form (cf. [6, 36, 58]):

$$I(N) = \frac{\exp\{Np(t_0)\}}{\sqrt{2\pi Np''(t_0)}}[q(t_0) + O(N^{-1})], \quad \text{if } t_0 < \beta_2 \tag{22}$$

$$= -\sum_{i=1}^{k} q_{-i}\exp\{Np(\beta_i)\} + \frac{\exp\{Np(t_0)\}}{\sqrt{2\pi Np''(t_0)}}[q(t_0) + O(N^{-1})], \text{ if } t_0 > \beta_1, \tag{23}$$

where $q_{-i} = \mathrm{Res}_{t=\beta_i}\{q(t)\}$ is the residue of $q(t)$ at $t = \beta_i$. If β_1 is a simple pole and the only pole of $q(t)$ in the vicinity of the saddle point t_0 then

$$I(N) = e^{Np(t_0)}\Bigg\{ - q_{-1}\frac{e^{Nb^2}}{2}\mathrm{erfc}(b\sqrt{N}) + \frac{1}{\sqrt{2\pi Np''(t_0)}}\left[q^*(t_0) - q_{-1}\frac{p'''(t_0)}{6p''(t_0)}\right] + O(N^{-3/2})\Bigg\}, \tag{24}$$

where

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-y^2}\,\mathrm{d}y,$$
$$b = \mathrm{sgn}\left(1 - \frac{t_0}{\beta_1}\right)\sqrt{p(\beta_1) - p(t_0)}$$

and

$$q^*(t) = q(t) - \frac{q-1}{t-\beta_1} = \sum_{k=1}^{\infty}\frac{1}{k!}h^{(k)}(\beta_1)(t-\beta_1)^{k-1},$$

where $h(t) = q(t)(t - \beta_1)$. The uniform asymptotic expansion formula (24) covers the case, when $t_0 = \beta_1$. This critical point separates the regions with different asymptotic expansions.

2.2 Bottleneck Analysis in Single-chain Closed Queueing Networks

We start with a simple CQN consisting of N customers and $K + 1$ service stations with exponential service times. The state of the network is described by a vector $(n_0, n_1, \ldots, n_K)$ where n_i is the number of customers (including those in service) at station, i, $i = 0, 1, \ldots, K$. Station 0 has an infinite server (the number of servers coincides with the number of customers N) with service rate $\mu_0 = \lambda n_0$. The other K stations have single servers with fixed rates μ_i, $1 \le i \le K$. Let $(p_{i,j})$, $i, j = 0, \ldots, K$, be the routing probability matrix for the network, *i.e.*, a customer completing service at service station i is routed to service station j with probability $p_{i,j}$. The stationary probability distribution of queue length at the single servers can be written as:

$$P(n_1, \ldots, n_K) = \frac{1}{G(N)}\frac{1}{(N - \sum_{i=1}^K n_i)!}\prod_{i=1}^K \rho_i^{n_i} \tag{25}$$

where $\rho_i = \lambda p_i/\mu_i$, $\{p_i\colon 0 \le i \le K\}$ is the unique solution to:

$$p_0 = 1, \qquad p_i = \sum_{j=0}^{M} p_j p_{j,i} \quad 1 \le i \le M$$

and the partition function

$$G(N) = \sum_{n_1+\cdots+n_M \le N}\frac{1}{(N - \sum_{i=1}^K n_i)!}\prod_{i=1}^K \rho_i^{n_i}.$$

(For $K = 1$ distribution (25) coincides with (5) up to the normalization constant.) The respective GPF

$$\mathcal{G}(z) = e^z \prod_{i=1}^{K} \frac{1}{1 - \rho_i z}. \tag{26}$$

The utilization of single server i is denoted by $U_i(N)$ and it can be expressed through the partition function as

$$U_i(N) = \rho_i \frac{G(N-1)}{G(N)}, \quad 1 \leq i \leq K. \tag{27}$$

Assume K is fixed while $N \gg 1$. To avoid trivial results it is also assumed that the parameters ρ_i are of the order N^{-1}, *i.e.* $\rho_i = r_i/N$ for some constants r_i. Then, after the change of variables $t = z/N$ we have $c(N) = N^{-N}, p(t) = t - \ln t$, and $q(t) = t^{-1}(1 - r_1 t)^{-1} \cdots (1 - r_K t)^{-1}$. (Without IS station $p(t) \equiv 0$ and $G(N)$ is expressed explicitly through the residues of $tq(t)$ [22].) In this case the saddle point $t_0 = 1$ and $\beta_1 = 1/r_1$ assuming $r_1 > r_2 \cdots > r_K$. If $r_1 < 1$ then (22) and (27) imply that $U_i = r_i + O(1/N)$ that coincides with the server utilization in $M/M/1$ system with traffic intensity $r_i, i = 1, \ldots, K$ (normal usage). For $r_1 > 1$ the main contribution in (23) is provided by the first term $r_1^N e^{N/r_1}$ with an exponentially small remainder. Then Eq. (27) implies that $U_1(N)$ can be approximated by (17) with $\rho = \rho_1$ and it is exponentially close to 1 (heavy usage), while for $2 \leq i \leq K$, $U_i(N) = r_i/r_1 + O(1)$ that coincides with the server utilization in $M/M/1$ system with traffic intensity r_i/r_1 (normal usage). Based on this result the single-server station 1 is referred to as *bottleneck*.

In general, station $i, 1 \leq i \leq K$ may have several identical servers or a limited queue dependent (LQD) server [36]. In such a case the term $\rho_i^{n_i}$ in (25) is replaced by $\rho_i^{n_i} / \prod_{k=1}^{n_i} s_i(k)$, where function $s_i(k)$ is constant for $n \geq L_i$, where $L_i < N$ is a fixed integer. For a CQN with LQD servers the GPF

$$\mathcal{G}(z) = e^z \prod_{t=1}^{K} \frac{A_i(\rho_i z)}{1 - (\rho_i / s_i(L_i))z}, \tag{28}$$

where $A_i(z)$ is a polynomial of order $L_i - 1$ defined in [36]. In this case the saddle point $t_0 = 1$ as before while $\beta_1 = \max_i s_i(L_i)/r_i$ and the bottleneck analysis is similar to the previous case.

The bottleneck analysis can be generalized to CQN, where K is also large such that $K/N = \gamma$, for some constant γ. We consider three examples of such networks. In the first CQN, there are Q LQD servers and T 'large' groups of identical LQD servers. Then after the change of variables $t = z/N$ we have $c(N) = N^{-N}$,

$$p(t) = t - \ln t - \sum_{j=1}^{T} \gamma_j [\ln(1 - \alpha_j t) - \ln A_j(r_j t)], \tag{29}$$

and

$$q(t) = \frac{1}{t} \prod_{i=1}^{Q} \frac{A_i(r_i t)}{1 - \alpha_i t}, \tag{30}$$

where $\alpha_i = r_i/s_i(L_i)$, $1 \leq i \leq Q$ and $\gamma_j = M_j/N$ where M_j is the number of identical stations in group j. In this case t_0 is a positive root of a polynomial equation and $\beta_1 = \max_{1 \leq i \leq Q} 1/\alpha_i$.

In the second CQN, there are $K + 1$ single servers. The first K single servers have a special form of intensities which can be written as [31, 35]

$$\rho_i = \frac{1}{N} r\left(\frac{i}{N}\right), \tag{31}$$

where $r(s)$ is a piecewise smooth function in $[0, \gamma]$ and $\gamma = K/N$. The intensity for the last single server is $\rho_{K+1} = r_{K+l}/N$, where r_{K+1} is a constant. In this case it is convenient to rewrite the the product in (26) as

$$\prod_{i=1}^{K}\left(1 - \frac{z}{N} r\left(\frac{i}{N}\right)\right)^{-1} = \exp\left\{-\sum_{i=1}^{K} \ln\left(1 - \frac{z}{N} r\left(\frac{i}{N}\right)\right)\right\}. \tag{32}$$

By the Euler–MacLauren formula we can approximate the sum in (32) by an integral and obtain the GPF

$$\mathcal{G}(z) = \frac{1}{1 - (r_{K+1}/N)z}\left(\frac{1 - (z/N)r(0)}{1 - (z/N)r(\gamma)}\right)^{1/2} \exp\left\{z - \int_0^{\gamma} \ln\left(1 - \left(\frac{z}{N}\right)r(s)\right) \mathrm{d}s\right\}. \tag{33}$$

Then, after the change of variables $t = z/N$ we obtain $c(N) = N^{-N}$,

$$p(t) = t - \ln t - \int_0^{\gamma} \ln(1 - \mathrm{tr}(s))\, \mathrm{d}s$$

and

$$q(t) = \frac{1}{t(1 - r_{K+1}t)}\left(\frac{1 - \mathrm{tr}(0)}{1 - \mathrm{tr}(\gamma)}\right)^{1/2}.$$

In this case $\beta_1 = 1/r_{K+1}$ and the saddle point t_0 is a single root of the equation

$$p'(t) = 1 - \frac{1}{t} + \int_0^{\gamma} \frac{r(s)\, \mathrm{d}s}{1 - \mathrm{tr}(s)}$$

in the interval $(0, a)$, $a = \min\{1, \min_{0 \leq s \leq \gamma} 1/r(s)\}$. For related results and generalizations see [41, 16]. The third CQN is related to the multiple-bus multiprocessor model [24]. In a particular case of the multiprocessor model [37] with N processors, K memory modules and crossbar interconnection network, we have $\rho_i \equiv \rho = \lambda/K\mu$ and $\mathcal{G}(z) = e^z/(1 - \rho z)^K$. (The behavior of a processor consists of cycles of computation followed by a request to one of K memory modules.) The state of this model is defined by the total number of processors l queued for memory modules, and the stationary distribution

$$P(l) = P(0)(N)_l\binom{K - 1 + l}{l}\rho^l, \tag{34}$$

where $(N)_l = N!/(N-l)!$ A multiple-bus interconnection network with $B < \min(N, K)$ buses studied in [24] is more cost-effective than the simpler crossbar organization. In this case contention may occur for buses as well as memories. The stationary distribution for CQN model of the multiple-bus multiprocessor system has the following form:

$$P(l) = P(0)(N)_l A(l)\rho^l, \tag{35}$$

where

$$A(0) = 1, \quad A(l) = \frac{K}{B}A(l-1) + \sum_{j=0}^{B-1}\left(1 - \frac{j}{B}\right)\binom{K}{j}\binom{l-1}{j-1}. \tag{36}$$

The asymptotic analysis of distribution (35) is reduced to that of the combinatorial factor $A(l)$ whose generating function has the following explicit expression:

$$\mathcal{A}(z) = \sum_{l\geq 0} A(l)z^l = (1-\gamma z)^{-1}\sum_{j=0}^{B-1}\left(1 - \frac{j}{B}\right)\binom{K}{j}\left(\frac{z}{1-z}\right)^j, \quad \gamma = \frac{K}{B}. \tag{37}$$

The complexity of $\mathcal{A}(z)$ makes asymptotic analysis of $A(l)$ significantly less trivial than in the previous cases. However, the nature of results is similar as the following approximations show Let $l_0 = KB/(K-B)$ and $x_0 = l_0/N$. Then application of the saddle-point method to the integral

$$\frac{1}{2\pi j}\oint_C \mathcal{A}(z)\,\mathrm{d}z = A(l)$$

leads to the following approximations [24]:

$$A(l) \approx \begin{cases} A_0(l) & \text{if } x_0 \geq 1 \\ A_0(l) & \text{if } x_0 \leq 1 \text{ and } l \leq l_0 \\ A_b(l) & \text{if } x_0 < 1 \text{ and } l > l_0, \end{cases}$$

where

$$A_0(l) = \binom{K-1+l}{l}, \quad A_b(l) = \left(\frac{\gamma-1}{2\pi K}\right)^{1/2}\left(\frac{\gamma}{\gamma-1}\right)^K \gamma^l$$

and $-A_b(l)$ is an approximation for the residue $\operatorname{Res}\{\mathcal{A}(z); z = \gamma^{-1}\}$. These approximations have an interesting and important interpretation: when $x_0 \geq 1$, as N increases, the system becomes asymptotically equivalent to the crossbar multiprocessor system, *i.e.*, *bus-sufficient*. For $x_0 < 1$ and $l > l_0$, the residue of $\mathcal{A}(z)$, being closer to the origin than the saddle point, dominates and the bus interconnection network becomes a bottleneck.

2.3 Generalized Erlang and Engset Models

For the generalized Erlang model the stationary distribution

$$\pi(n_1, \dots, n_C) = \frac{1}{G(M)} \prod_{j=1}^{C} \frac{v_j^{n_j}}{n_j!}, \quad \sum_{j=1}^{C} b_j n_j \le M, \tag{38}$$

where $v_j = \lambda_j / \mu_j$, λ_j is mean arrival rate of service j calls and the partition function

$$G(M) = \sum_{\sum_{j=1}^{C} b_j n_j \le M} \prod_{j=1}^{C} \frac{v_j^{n_j}}{n_j!}. \tag{39}$$

The loss probability of service j is

$$L_j = \frac{G(M - b_j)}{G(M)}.$$

The respective GPF is [55]

$$\mathcal{G}(z) = \frac{\exp[\sum_{j=1}^{C} v_j z^{b_j}]}{1 - z}.$$

Assuming $M = N$, $v_j = Nc_j$, where $c_j, j = 1, \dots, C$ are constant, while $N \gg 1$ we have $c(N) = 1$,

$$p(t) = \sum_{j=1}^{C} c_j t^{b_j} - \ln t \tag{40}$$

and $q(t) = [t(1 - t)]^{-1}$.

In the generalized Engset model [11], calls of service $j, j = 1, \dots, C$ are generated by a finite source of size $N_j, N = (N_1, \dots, N_C)$. Service j call requires b_{kj} circuits, or resource units with probability p_{kj}, where $k \in K_j$ and $\sum_{k \in K_j} p_{kj} = 1$. Let M be the total number of available resource units. If the requested number b_{kj} of resource units is available then the call will hold them for the holding period with mean $1/\mu_{kj}$. Otherwise the call will be blocked and lost. Let $1/\lambda_j$ be the mean intergeneration time for service j calls and $\rho_{kj} = (\lambda_j / \mu_{kj}) p_{kj}$. For each $k \in K_j$ denote by n_{kj} the number of type j requests holding b_{kj} resource units and $\mathbf{n} = \{n_{kj}, k \in K_j, j = 1, \dots, C\}$. If holding times are exponentially distributed then the stationary distribution

$$\pi(\mathbf{n}) = \frac{1}{G(\mathbf{N}, M)} \prod_{j=1}^{C} \left(\frac{N_j!}{\prod_{k \in K_j} n_{kj}! (N_j - \sum_{k \in K_j} n_{kj})!} \right) \prod_{k \in K_j} \rho_{kj}^{n_{kj}} = \frac{1}{G(\mathbf{N}, M)} \hat{\pi}(\mathbf{n}),$$

where

$$G(\mathbf{N}, M) = \sum_{\mathbf{n}} \hat{\pi}(\mathbf{n})$$

and the summation is only over permissible values of **n** given by

$$\sum_{j=1}^{C}\sum_{k\in K_j} b_{kj}n_{kj} \le M, \qquad \sum_{k\in K_j} n_{kj} \le N_j.$$

The loss probability of service j is

$$L_j = 1 - \sum_{k\in K_j} p_{kj}\frac{G(\mathbf{N}-\mathbf{e}_j, M-b_k)}{G(\mathbf{N}, M)},$$

where $\mathbf{e}_j$ is the vector with elements $e_{ji} = \delta_{ji}$, where δ_{ji} denotes the Kronecker delta. The respective GPF is [11]

$$\mathcal{G}(z) = \frac{\prod_{j=1}^{C}(1+\sum_{k\in K_j}\rho_{kj}z^{b_{jk}})^{N_j}}{1-z}. \tag{41}$$

Assuming $M = N, N_j = N\alpha_j$, where $\alpha_j, j = 1, \ldots, C$ are constant, while $N \gg 1$ we have $c(N) = 1$,

$$p(t) = \sum_{j=1}^{C}\alpha_j \ln\left(1+\sum_{k\in K_j}\rho_{kj}z^{b_{jk}}\right) - \ln t \tag{42}$$

and the same $q(t) = [t(1-t)]^{-1}$ as in the Erlang model.

Under some natural conditions [11, 38, 47] there is a unique positive solution t_0 of $p'(t) = 0$, where $p(t)$ is given by Eq. (42) or (40). Moreover, it is easy to prove that $|t| = t_0$ is a saddle-point contour, and $t = t_0$ is the only saddle point in it. Now asymptotics of the partition functions for generalized Erlang and Engset models can be obtained using expansions (22), (23) and (24) with $\beta_1 = 1$ and $\beta_i = 0$ for $i > 1$. For the Erlang and Engset models the equation $p'(t) = 0$ is linear. In particular, for the Erlang model $t_0 = 1/c$ (the subscript 1 is omitted) and we obtain the same three approximations as in Section 1.1 (see (9)–(14)) for underloaded ($t_0 > 1$), critically loaded ($t_0 = 1$) and overloaded ($t_0 < 1$) regimes. The same classification applies to generalized Erlang and Engset models [11, 38, 45–47]. It is interesting to note that for both LN and CQN the three asymptotic regimes are defined by comparing some number with 1. For LN this number is the saddle point of $p(t)$ while for CQN it is the closest to the origin positive pole of $q(t)$.

3 ASYMPTOTICS OF MULTIDIMENSIONAL PARTITION FUNCTIONS AND THEIR APPLICATION

For multichain CQN and LN with several links the partition function depends on several integer parameters whose number equals to the number of customer types and the number of links respectively. Then the generating partition function defined as

$$\mathcal{G}(z_1, \ldots, z_m) = \sum_{n_1=0}^{\infty}\cdots\sum_{n_m=0}^{\infty} z_1^{n_1}\cdots z_m^{n_m}G(n_1, \ldots, n_m),$$

where $m = J$ or $m = L$, is a multidimensional complex function. In this section, we first provide explicit expressions for the GPF for both classes of networks and then generalize some of the results in the previous section using the integral representation for the partition function in multidimensional complex space. Similar to the one-dimensional case, the key step in asymptotic evaluation of the multidimensional contour integral is its transformation to the saddle-point contour integral, which has an explicit asymptotic expression. This approach requires one to find the saddle-point contour, move the initial integration contour to the saddle-point contour, pick up the residues between these two contours, and, finally, find the term providing the main contribution to the asymptotic expansion of the partition function. However, in contrast to the one-dimensional case, implementation of these steps is generally far from trivial.

3.1 Explicit Expressions for GPF

First, consider a CQN with J customer types and $K+1$ service stations, one of which is infinite server (IS) and K others are processor-sharing (PS) stations. We assume that customers of each type visit all stations. It is convenient to number the IS-station by 0. Let $\mathbf{n}$ denote a $J \times K$ matrix whose element n_{ji} represents the number of type j customers at PS-station i. The population of customers of type j is a constant N_j, $1 \le j \le J$. The state space is the set $\mathcal{S}$ of matrices $\mathbf{n}$ which have integer components, and satisfy the population constraints

$$\mathcal{S} = \left\{ \mathbf{n} | 0 \le n_{ji}, \sum_i n_{ji} \le N_j, 1 \le j \le J, 1 \le i \le K \right\}.$$

Then the product form solution has the form [26]:

$$P(\mathbf{n}) = \frac{1}{G} \prod_{j=1}^{J} \frac{1}{(N_j - \sum_i n_{ji})!} \prod_{i=1}^{K} n_i! \frac{\rho_{ji}^{n_{ji}}}{n_{ji}!} = \frac{1}{G} \hat{P}(\mathbf{n}), \tag{43}$$

where $n_i = \sum_j n_{ji}$ and the normalization constant

$$G = G(N_1, \ldots, N_J) = \sum_{\mathbf{n} \in \mathcal{S}} \hat{P}(\mathbf{n}).$$

Moreover,

$$\rho_{ji} = \frac{e_{ji} \lambda_j}{\mu_{ji}},$$

where e_{ji} is the relative visiting rate of type j customers to PS-station i as compared to the IS-station, $1/\lambda_j$ is the mean service time of a type j customer at the IS-station, $1/\mu_{ji}$ is the mean service time of an isolated type j customer at PS-station i. Using the definition of the multidimensional GPF and performing the summation we have

$$\mathcal{G}(z_1, \ldots, z_J) = e^{z_1 + \cdots + z_J} \prod_{k=1}^{K} \left(1 - \sum_{j=1}^{J} \rho_{ji} z_j \right)^{-1}. \tag{44}$$

Next, consider a special multichain CQN that has been introduced in [6] in the context of modeling large multiprogramming systems. This CQN consists of J dedicated single servers (one for each chain) and T groups of stations with identical single servers inside each group. After service completion at the single server in chain j a customer visits a station in group l with probability p_{jl}/K_l, where K_l is the number of stations in group l and $\sum_{l=1}^{T} p_{jl} = 1$ for each j. Let μ_j is service rate of the single server in chain j, and θ_l is service rate of a single server in group l. Then the respective GPF is

$$\mathcal{G}(z_1, \ldots, z_J) = \prod_{j=1}^{J} \left(\frac{1 - z}{\mu_j} \right)^{-1} \prod_{l=1}^{T} \left(1 - \frac{1}{\theta_l K_l} \sum_{j=1}^{2} p_{jl} t_j \right)^{-K_l} . \tag{45}$$

Finally, consider the LN with independent Poisson arrival streams of rate λ_r and generally distributed holding time with the mean $1/\mu_r$ or route r. Its stationary probability distribution has the form [27]

$$\pi(\mathbf{n}) = \frac{1}{G} \prod_{r=1}^{R} \frac{\nu^{n_r}}{n_r!},$$

where $\nu_r = \lambda_r/\mu_r$ and the normalization constant

$$G = G(\mathbf{M}) = \sum_{B\mathbf{n} \leq \mathbf{M}} \prod_{r=1}^{R} \frac{\nu^{n_r}}{n_r!},$$

Here B is the matrix with elements b_{lr} and $\mathbf{M} = (M_1, \ldots, M_L)$. The respective GPF is [48]

$$\mathcal{G}(z_1, \ldots, z_L) = \frac{\exp(\sum_{r=1}^{R} \nu_r \prod_{l=1}^{L} z_l^{b_{lr}})}{\prod_{l=1}^{L} (1 - z_l)} . \tag{46}$$

Whittle [57] obtained a result equivalent to (46) but with an additional factor in the numerator, which is redundant. The multidimensional generalization of (41) is given in [11]. Let L_r be the stationary probability that a call requesting route r is lost. Then

$$L_r = 1 - \frac{G(\mathbf{M} - B\mathbf{e}_r)}{G(\mathbf{M})},$$

where $\mathbf{e}_r$ is the vector with elements $e_{rs} = \delta_{rs}$, where δ denotes the Kronecker delta.

3.2 Asymptotic Approximations

As before, the partition function can be recovered from the generating function by the application of the inverse Cauchy formula [54]

$$G(n_1, \ldots, n_m) = \frac{1}{(2\pi j)^m} \oint_{\gamma_1^1} \cdots \oint_{\gamma_m^1} \frac{\mathcal{G}(z_1, \ldots, z_m)}{z_1^{n_1+1} \cdots z_m^{n_m+1}} \mathrm{d}z_1 \cdots \mathrm{d}z_m, \tag{47}$$

where $\gamma_k^1 = \{z_k = \varepsilon e^{j\omega_k}, \omega_k \in [0, 2\pi]\}$, $\varepsilon \ll 1, k = 1, \ldots, m$ and j is the imaginary unit. Note that this integral is just an application of one-dimensional Cauchy integral m times. Bertozzi and McKenna [5] tried to generalize Gordon's [22] result and obtain an explicit expression for the partition function through the residues of the GPF for a multichain CQN consisting only of FCFS or PS nodes. However, a multidimensional residue cannot always be calculated explicitly, and therefore Gordon's result does not have a direct analog in multichain networks. Explicit expression for the partition function is not available even for the simplest two-chain CQN which consists only of two FCFS or PS nodes (see [5, Section 4.5]). There are four main differences between asymptotic evaluation of the contour integral in one-dimensional and multidimensional cases. First, the saddle-point contour is no longer a circle, but a two-dimensional surface that must lie in the domain of analyticity of the generating function, whose singularity set is defined by two-dimensional varieties. It is worth noting that this condition is quite independent from that defining the generic properties of the saddle-point contour. Second, in general, the geometric relation between the two contours in multi-dimensional case is not as evident, and one has to apply the theory of homology. Third, the residue calculation is non-trivial, even in the two-dimensional case. Only some of the residues can be explicitly calculated, while others are reduced to one-dimensional integrals and then evaluated by the saddle-point method. Fourth, the identification of the main contribution in the asymptotic expansion is also more complex.

In this section, we provide explicit asymptotic expressions for partition functions with GPF given by Eqs. (45), (46) and (44). Asymptotic results for a CQN consisting of only PS nodes can be obtained as a byproduct of our more general results. For simplicity of notation we formulate the results in two-dimensional case. Similar to one-dimensional case we introduce a large parameter N and represent the integrand in (47) for $m = 2$ as $w(N)q(t_1, t_2)$ $\exp\{Np(t_1, t_2)\}$, where $w(N)$ is a constant whose calculation does not require summation, while the functions $p(t_1, t_2)$ and $q(t_1, t_2)$ do not depend on N. Denote

$$I(N) = -\frac{1}{4\pi^2}\oint_{\gamma_1}\oint_{\gamma_2} q(t_1, t_2)\exp\{-Np(t_1, t_2)\}\, \mathrm{d}t_1\, \mathrm{d}t_2, \tag{48}$$

where $\gamma_k = \{t_k = \varepsilon e^{j\omega_k}, \omega_k \in [0, 2\pi]\}$, $\varepsilon \ll 1, k = 1, 2$. Note that $G(N) = w(N)I(N)$. To evaluate the integral $I(N)$ by the saddle-point method we have to find a saddle-point contour. Let $\mathbf{t} = (t_1, t_2)$ and the notation $|\mathbf{t}| = \mathbf{c}$ stand for $|t_1| = c_1, |t_2| = c_2$. Assuming that the function $p(\mathbf{t})$ has a single maximum in real space at $\mathbf{t} = \mathbf{t}^0$ and using the fact that the series expansion for the respective generating function has positive coefficients (cf. [18]) one can prove that

$$\min_{\mathbf{c}} \max_{|\mathbf{t}|=\mathbf{c}} \mid \exp\{Np(\mathbf{t})\} = \exp\{Np(\mathbf{t}^0)\},$$

which means that $\gamma_k^0 = \{t_k = t_k^0 e^{j\omega_k}, \omega_k \in [0, 2\pi]\}$, $\varepsilon \ll 1, k = 1, 2$ is a saddle-point contour, and the saddle point $\mathbf{t}^0 = (t_1^0, t_2^0)$ satisfies the following system of equations:

$$\frac{\partial p(t_1, t_2)}{\partial t_k} = 0, \quad k = 1, 2. \tag{49}$$

Denote

$$I_0(N) = -\frac{1}{4\pi^2}\oint_{\gamma_1^0}\oint_{\gamma_2^0} q(t_1, t_2)\exp\{-Np(t_1, t_2)\}\, \mathrm{d}t_1\, \mathrm{d}t_2. \tag{50}$$

Then (see [18, 58])

$$I_0(N) = \frac{q(t_1^0, t_2^0)\exp\{Np(t_1^0, t_2^0)\}}{2\pi N\sqrt{\det p''_{t_1t_2}(t_1^0, t_2^0)}}\left(1 + O\left(\frac{1}{N}\right)\right), \tag{51}$$

where $p''_{t_1t_2}(t_1^0, t_2^0)$ is the Hessian matrix of $p(t_1, t_2)$ at $(t_1, t_2) = (t_1^0, t_2^0)$ and det stands for the determinant of the matrix.

First, consider a special case $q(t_1, t_2) = q_1(t_1)q_2(t_2)$, where transformation of the initial two-dimensional contour integral to the saddle-point contour integral is quite similar to the one-dimensional case. For simplicity, assume that $q_k(t_k)$ has only one simple positive pole at $t_k = \beta_k, k = 1, 2$. If $\beta_k > t_k^0, k = 1, 2$, then approximation (51) holds. Otherwise we move the initial contour to the saddle-point contour and calculate the residues as follows. We have (cf. [6, 11])

$$I(N) = -q_{-1}\frac{1}{2\pi j}\oint_{\gamma_2} q_2(t_2)\exp\{-Np(\beta_1, t_2)\}\,dt_2 + I_0(N) \text{ if } \beta_1 < t_1^0,\ \beta_2 > t_2^0 \tag{52}$$

$$= -q_{-2}\frac{1}{2\pi j}\oint_{\gamma_1} q_1(t_1)\exp\{-Np(t_1, \beta_2)\}\,dt_1 + I_0(N) \text{ if } \beta_1 > t_1^0,\ \beta_2 < t_2^0 \tag{53}$$

$$= q_{-1}q_{-2}\exp\{Np(\beta_1, \beta_2)\} + I_0(N) \text{ if } \beta_k < t_k^0,\ k = 1, 2, \tag{54}$$

where $q_{-k} = \text{Res}_{t=\beta_k}\{q_k(t)\}$ is the residue of $q_k(t)$ at $t = \beta_k, k = 1, 2$. The first term in (52), (53) and (54) is one- and two-dimensional residue respectively of the integrand at $t_k = \beta_k, k = 1, 2$. It is not difficult to prove that it provides the main contribution to $I(N)$ with an exponentially small remainder. Finally, contour integrals in (52) and (53) can be evaluated using relations (22)–(24).

Equations (52)–(54) can be applied to asymptotic analysis of the CQN with GPF (45) and the LN with GPF (46). In the CQN case we assume the following scaling:

$$\mu_j = N\mu_{0j}, \quad N_j = \alpha_j N, \quad K_l = \gamma_l N,$$

where $\mu_{0j}, \alpha_j, j = 1, 2$ and $\gamma_l, l = 1, \ldots, T$ are bounded while $N \to \infty$. Then after the change of variables $t_j = z_j/N$ in the initial integral representation for the partition function we have

$$w(N) = N^{-(\alpha_1+\alpha_2)N}, \qquad q(t_k) = \left[t_k\left(\frac{1 - t_k}{\mu_{0k}}\right)\right]^{-1}, \quad k = 1, 2$$

and

$$p(t_1, t_2) = -\sum_{j=1}^{2}\alpha_j \ln t_j - \sum_{l=1}^{T}\gamma_l \ln\left(1 - \frac{1}{\theta_l\gamma_l}\sum_{j=1}^{2}p_{jl}t_j\right).$$

In the LN case we assume

$$v_r = Nv_{0r}, \quad r = 1, \ldots, R \quad \text{and} \quad M_l = \alpha_l, \quad l = 1, 2,$$

where ν_{0r} and α_l are bounded while $N \to \infty$. Then

$$w(N) = 1, \quad q(t_k) = [t_k(1-t_k)]^{-1}, \quad k = 1,2, \quad p(t_1,t_2) = \sum_{r=1}^{R} \nu_{0r} t_1^{b_{1r}} t_2^{b_{2r}} - \sum_{l=1}^{2} \alpha_l \ln t_l.$$

Asymptotic approximations for the loss probabilities in a particular case $b_{11} = b_{21} = b_{22} = 1, b_{12} = 0$ are derived in [48]. For asymptotic analysis of the LN with finite sources see [11].

We conclude this section with asymptotic analysis of the partition function whose GPF is given by (44). We assume the following scaling:

$$N_k = \alpha_k N, \qquad \rho_{ik} = \frac{r_{ik}}{N}, \qquad \alpha_k > 0, \ r_{ik} > 0, \ i,k = 1,2, \ N \to \infty, \tag{55}$$

where α_k and r_{ik} are bounded. After the change of variables $t_k = z_k/N$ in the integral representation for $G(N_1, N_2)$ we have

$$w(N) = N^{-(\alpha_1+\alpha_2)N}, \quad p(t_1,t_2) = t_1 + t_2 - \alpha_1 \ln t_1 - \alpha_2 \ln t_2$$

and

$$q(t_1,t_2) = [t_1 t_2 (1 - r_{11}t_1 - r_{12}t_2)(1 - r_{21}t_1 - r_{22}t_2)]^{-1}.$$

In this case $q(t_1,t_2)$ does not have the form $q_1(t_1)q_2(t_2)$ and the analysis becomes much more complicated. Denote

$$\eta_i = r_{i1}\alpha_1 + r_{i2}\alpha_2, \qquad i = 1,2.$$

Equation (49) has a unique solution $(t_1^0, t_2^0) = (\alpha_1, \alpha_2)$ that provides the saddle point. If both η_1 and η_2 are less than 1 (normal usage) then the function $q(t_1,t_2)/(t_1t_2)$ does not have singularities inside the saddle-point contour and $I(N) = I_0(N)$ with the approximation given by (51). To evaluate the integral $I(N)$ when at least one of η_i is greater than 1 (heavy usage at least at one of the PS stations) we have to we move the initial integration contour to the saddle-point contour and, as before, to pick up all the residues of the integrand between the two contours. These residues are defined by the set of singularities of $q(t_1,t_2)$. More exactly, denote by S the set of singularities of the integrand. (The integrand is holomorphic in $\mathbf{C}^2 \backslash S$.) We have:

$$\begin{aligned} &S = \bigcup_{i=1}^{4} S_i, \quad S_1 = \{t_1 = 0\}, \quad S_2 = \{t_2 = 0\}, \\ &S_3 = \{1 - r_{11}t_1 - r_{12}t_2 = 0\}, \quad S_4 = \{1 - r_{21}t_1 - r_{22}t_2 = 0\}. \end{aligned} \tag{56}$$

The exact relation between the two integrals can be expressed in terms of some "basis" integrals, corresponding to the points of intersection of S_i and S_k, $i,k = 1,\dots,4$. It turns out that these "basis" integrals provide the main contribution to the asymptotics of $I(N)$, rather than the saddle-point integral. The dimension of the basis integrals is readily reduced at least by 1, and they are evaluated by one-dimensional saddle-point method or calculated explicitly.

Denote

$$p_i(t) = p\left(t, \frac{1 - r_{i1}t}{r_{i2}}\right), \quad i = 1, 2,$$
$$\Delta = r_{11}r_{22} - r_{12}r_{21},$$
$$q_i(t) = \frac{1}{t(1 - r_{i1}t)(r_{i2} - r_{3-i,2} + \Delta t)}, \quad i = 1, 2.$$

Let also t_{0i} be the least positive roots of the equation $p_i'(t) = 0, i = 1, 2$. Assuming $\Delta \neq 0$, denote

$$\beta_{10} = \frac{r_{22} - r_{12}}{\Delta}, \quad \beta_{20} = \frac{r_{11} - r_{21}}{\Delta}.$$

(β_{10}, β_{20}) is the intersection point of S_3 and S_4. Now we can formulate the main result in [40].

THEOREM 1 *If* (55) *holds and at least one of* η_i *is greater than* 1, *then the partition function has the following asymptotics* (*as* $N \to \infty$) :

1. *If* $\eta_i > 1$ *while* $\eta_{3-i} < 1$ *then*

$$I(N) = \frac{r_{i2}}{\sqrt{2\pi N p_i''(t_{0i})}} q_i(t_{0i}) e^{Np_i(t_{0i})}(1 + O(N^{-1})), \quad i = 1, 2. \tag{57}$$

2. *Let both* η_1 *and* η_2 *are greater than* 1.
   (a) *If in addition* $\Delta \neq 0$ *while* $p_1'(\beta_{10})p_2'(\beta_{20}) < 0$ *then*

$$I(N) = \frac{e^{Np(\beta_{10}, \beta_{20})}}{\beta_{10}\beta_{20}|\Delta|}(1 + O(N^{-1})). \tag{58}$$

   (b) *Otherwise asymptotics* (57) *holds.*

4 ASYMPTOTIC EXPANSIONS FOR PROBABILITY DISTRIBUTIONS

This section is motivated by the problem of dimensioning bandwidth for high-speed data transmission networks. Two models are considered. The first model is described by the generalized Erlang or Engset model. It can be applied to engineering of the router uplink whose utilization should be well below 100%. The second model is described by a CQN with multiple customer types that consists of one IS station and many PS stations [4]. It is applied to bandwidth engineering of peering links which are usually saturated. In prior work, [2], we motivated a CQN model and, using asymptotic approximations, determined dimensioning rules for the case of a single link and single type of connections. As in a classical situation of the sum of independent random variables (see *e.g.*, [19, Chapter XVI]) these asymptotic expansions provide the normal approximation and a correction to it which have an explicit expression up to a solution of polynomial equations.

4.1 Generalized Erlang and Engset Models

Let Q be the total number of busy circuits, or resource units in the generalized Erland or Engset model. Then

$$\Pr\{Q > m\} = 1 - \frac{G(m)}{G(M)}, \tag{59}$$

where $G(k)$ is the partition function with the total number of available resource units equal k. Under scaling assumption in Section 2.3 we have

$$G(m) = \frac{1}{2\pi j}\oint_C \frac{\exp\{Np(t,x)\}}{t(1-t)}\,\mathrm{d}t, \tag{60}$$

where $p(t,x)$ is obtained from Eq. (40) or (42) by substituting $-x\ln t$ instead of $-\ln t$ and $x = m/N$. Assume that the saddle point t_0 of the function $p(t,1) \equiv p(t)$ is greater than 1 that implies underloaded regime. Then one can prove (see *e.g.* [38]) that Q/M converges with probability 1 to

$$x^* = \begin{cases} \sum_{j=1}^{C} c_j b_j & \text{for generalized Erlang model} \\ \sum_{j=1}^{C} \dfrac{\alpha_j p_j b_j}{1+\rho_j} & \text{for generalized Engset model.} \end{cases}$$

Let $t_0(x)$ be a single positive root of equation $p_t'(t,x) = 0$. We assume that $x > x^*$ and x is close to x^*. This implies that $t_0(x)$ is close to 1 and $t_0(x) > 1$. Now asymptotic expansion for the probability distribution (59) can be obtained using for $G(m)$ the uniform asymptotic expansion (24) or more accurate but quite complicated expansion in [47]. The first term of the asymptotic expansion for $\Pr\{Q > m\}$ provides the normal approximation:

$$\Pr\{Q > m\} \approx 1 - \Phi(\sqrt{2N[p(1) - p(t_0(x), x)]}),$$

where $\Phi(x)$ is the standard normal distribution function with mean 0 and variance 1.

4.2 A Large Closed System with Multiple Customer Types

We consider a CQN that consists of one IS station with multiple customer types and one PS station [4]. This model can be applied to the dimensioning of bandwidth and of admission control for different data sources subject to feedback control in packet-switched communication networks when available bandwidth at the network nodes is shared between all active sources. Data sources are modeled by an IS station, network nodes are modeled by processor-sharing (PS) stations, and a 'customer' in the CQN represents an active data source. The distinguishing property of this application is that this CQN model is valid only if the PS station is saturated (heavy usage); see [2] for further details. The saturated station is defined asymptotically as the station where the number of customers grows proportionally to the total number of customers in the network as the latter increases with service rates at the PS station. The application includes the performance metric that the bandwidth received by an active data source at a given network node is greater than a target value with probability

$1 - \alpha$, where α is in the range of 0.001 to 0.1. As the network nodes of interest have a packet-based implementation of processor sharing, the performance metric can be restated as the number of active data sessions at a network node (the total number of customers at the PS station in the CQN model) is less than a target value with the given probability $1 - \alpha$. As the above probability will be calculated in the context of network planning and of network operations, the calculation will need to be done often and quickly.

Let Q_j be the random variable for the steady-state number of type-j customers at the PS node. The steady state probability distribution for the CQN is obtained from (43) with $K = 1$ and can be written in the following form:

$$\Pr\{Q_1 = n_1, \ldots, Q_J = n_J\} = \frac{1}{G} \prod_{j=1}^{J} \frac{N_j!}{(N_j - n_j)!} n! \frac{\rho_j^{n_j}}{n_j!},$$

where $n \equiv \sum_{j=1}^{J} n_j$.

Denote by $Q \equiv \sum_{j=1}^{J} Q_j$ the total number of customers at the PS node, and let

$$P(n) \equiv \Pr\{Q = n\} = \sum_{n_1 + \cdots + n_J = n} \Pr\{Q_1 = n_1, \ldots, Q_J = n_J\}$$

be its probability mass function. In general, the above sum does not seem to be reduced to a product of functions depending only on n and/or network parameters. However the exponential generating function $\mathcal{P}(z)$ for the sequence $P(n)$ has the following simple expression:

$$\mathcal{P}(z) = \sum_{n=0}^{N} P(n) \frac{z^n}{n!} = G^{-1} \prod_{j=1}^{J} (1 + \rho_j z)^{N_j}$$

which is easily derived from definitions of $P(n)$ and $\mathcal{P}(z)$. Using the Cauchy formula, we obtain for $P(n)$ the following integral representation in complex space:

$$P(n) = \frac{1}{G} n! \frac{1}{2\pi j} \oint_C \frac{\prod_{j=1}^{J} (1 + \rho_j z)^{N_j}}{z^{n+1}} \, \mathrm{d}z,$$

where C is any circular contour around $z = 0$.

We study the asymptotics of $P(n)$ under the following two assumptions.

1. The total number of customers in the network $N = \sum_{k=1}^{K} N_k$ is large, *i.e.* $N \gg 1$ and moreover

$$r_j = N\rho_j \quad \text{and} \quad \alpha_k = \frac{N_k}{N}, \tag{61}$$

where r_j and $\alpha_j, j = 1, \ldots, J$, remain bounded as $N \to \infty$.

2. The PS station is saturated which is expressed by the following heavy usage condition

$$\sum_{j=1}^{J} \alpha_j r_j > 1. \tag{62}$$

Now we can formulate the main results in [4].

PROPOSITION 2 *Let conditions* (61) *and* (62) *be satisfied and* $N \to \infty$ *while* $n = Nx$, *where both* x and $1 - x$ are $O(1)$. Then the probability distribution of the total number of customers at the PS station has the following asymptotic expansion $\Pr\{Q = n\}$

$$= \sqrt{\frac{\Delta}{2\pi N}} f\left(\frac{n}{N}\right) \exp\left\{-N\left(F\left(\frac{n}{N}\right) - F(x^*)\right)\right\}\left(1 + O\left(\frac{1}{N}\right)\right),$$

where x^* *is a single positive solution of equation*

$$\sum_{j=1}^{J} \frac{\alpha_j r_j}{1 + r_j x} = 1,$$

$$\Delta = \sum_{j=1}^{J} \frac{\alpha_j r_j^2}{(1 + r_j x^*)^2},$$

$$F(x) = x - x \ln x - S(u_o(x))$$

$$f(x) = \frac{1}{u_o(x)} \sqrt{\frac{x}{S''(u_o(x))}},$$

$$S(u) = \sum_{j=1}^{J} \alpha_j \ln(1 + r_j u) - x \ln u$$

and $u_o(x)$ *is a unique positive solution of equation*

$$S'(u) = \sum_{j=1}^{J} \frac{\alpha_j r_j}{1 + r_j u} - \frac{x}{u} = 0$$

on the real axis for each $x \in (0, 1)$.

COROLLARY 1 *The function* $F(x)$ *defines the logarithmic asymptotics of the probability distribution* $P(n) = \Pr\{Q = n\}$ *in the following sense*:

$$\lim_{N \to \infty} \frac{\ln P(n)}{N} = -(F(x) - F(x^*)).$$

Moreover, $F(x^*)$ *defines the logarithmic asymptotics of the normalization constant* $G = G(N)$:

$$\lim_{N \to \infty} \frac{\ln G(N)}{N} = -F(x^*).$$

COROLLARY 2 *The normalized total number of processor sharing customers Q/N converges to x^* in probability and*

$$\frac{Q - Nx^*}{\sqrt{N}}$$

is asymptotically normal with mean 0 *and variance*

$$\sigma^2 = \Delta^{-1} - x^*.$$

Asymptotic expansion for complementary probability distribution has the following form:

$$\Pr\{Q > m\} \sim \frac{1}{2}\mathrm{erfc}\left\{\sqrt{N(F(a) - F(x^*))}\right\} - \sqrt{\frac{\Delta}{2\pi N}} e^{\{N(F(x^*)-F(a))\}} \times \left(f(a)H(a) + \frac{(2\Delta)^{-1/2} - f(a)[F(a) - F(x^*)]^{1/2}[F'(a)]^{-1}}{[F(a) - F(x^*)]^{1/2}}\right),$$

where

$$H(a) = \frac{1}{F'(a)} - \frac{1}{\exp\{F'(a)\} - 1} = \frac{0.5 + \sum_{l=1}^{\infty}[F'(a)]^l/(l+2)!}{1 + \sum_{l=1}^{\infty}[F'(a)]^l/(l+1)!},$$

and

$$F'(a) = -\ln a + \ln u_o(a).$$

References

[1] Berger, A., Bregman, L. and Kogan, Y. (1999) Bottleneck analysis in multiclass closed queueing networks and its application, *Queueing Systems*, **31**, 217–237.

[2] Berger, A. and Kogan, Y. (2000) Dimensioning bandwidth for elastic traffic in high-speed data networks, *IEEE/ACM Transactions on Networks*, **8**, 643–654.

[3] Berger, A. and Kogan, Y. (1999) Multi-class elastic data traffic: Bandwidth engineering via asymptotic approximations, In: Key, P. and Smith, D. (Eds.), *Teletraffic Engineering in a Competitive World* (Vol. 3a). Amsterdam: Elsevier, pp. 77–86.

[4] Berger, A. and Kogan, Y. (2000) Distribution of processor-sharing customers for a large closed system with multiple classes, *SIAM J. Appl. Math.*, **60**, 1330–1339.

[5] Bertozzi, A. and McKenna, J. (1993) Multidimensional residues, generating functions, and their application to queueing networks, *SIAM Review*, **35**, 239–268.

[6] Birman, A. and Kogan, Y. (1992) Asymptotic evaluation of closed queueing networks with many stations, communications in statistics, *Stochastic Models*, **8**, 543–564.

[7] Birman, A. and Kogan, Y. (1996) Error bounds for asymptotic approximations of the partition function, *Queueing Systems*, **23**, 217–234.

[8] Bleistein, N. and Handelsman, R. A. (1986) *Asymptotic Expansions of Integrals*. Toronto: Dover.

[9] Borovkov, A. A. (1976) *Stochastic Processes in Queueing Theory*. New York: Springer-Verlag.

[10] Brockmeyer, E., Halstronand, H. L. and Jensen, A. (1948) *The Life and Works of A. K. Erlang*. Copenhagen: Academy of Technical Sciences.

[11] Choudhury, G. L., Kogan, Y. and Susskind, S. (1998) Exact and asymptotic solutions for models of new telecommunication services, *Ann. Oper. Res.*, **79**, 393–407.

[12] Choudhury, G. L., Leung, K. K. and Whitt, W. (1995) Calculating normalization constants of closed queueing networks by numerically inverting their generating functions, *J. ACM*, **42**, 935–970.

[13] Choudhury, G. L., Leung, K. K. and Whitt, W. (1995) An algorithm to compute blocking probabilities in multi-rate multi-class multi-resource loss models, *Advances in Applied Probability*, **27**, 1104–1143.

[14] Copson, E. T. (1935) *An Introduction to the Theory of Functions of a Complex Variable*. London: Oxford Univ. Press.

[15] Cox, D. R. and Smith, W. L. (1961) *Queues.* London: Methuen.

[16] Fayolle, G. and Lasgoutttes, J.-M. (1996) Asymptotics and scalings for large product-form networks via the Central Limit Theorem, *Markov Processes Relat. Fields*, **2**, 317–348.

[17] De Bruin, N. G. (1961) *Asymptotic Methods in Analysis* (2nd ed.). New York: Wiley.

[18] Fedoryuk, M. V. (1987) *Asymptotics: Integrals and Series.* Moscow: Nauka [in Russian].

[19] Feller, W. (1957) *An Introduction to Probability Theory and its Applications* (Vol. 1, 2nd ed.). New York: John Wiley & Sons.

[20] Ferdinand, A. E. (1971) An analysis of machine interference model, *IBM Syst. J.*, **10**, 129–142.

[21] Freidlin, M. and Wentzell, A. (1984) *Random Perturbation of Dynamical Systems.* New York: Springer.

[22] Gordon, J. J. (1990) The evaluation of normalizing constants in closed queueing networks, *Oper. Res.*, **38**, 863–869.

[23] Gordon, W. J. and Newell, G. F. (1967) Closed queueing systems with exponential servers, *Oper. Res.*, **15**, 254–265.

[24] Hofri, M. and Kogan, Y. (1994) Asymptotic analysis of product-form distributions related to large interconnection networks, *Theoretical Computer Science*, **125**, 61–90.

[25] Jagerman, D. L. (1974) Some properties of the Erlang loss function, *BSTJ*, **53**, 525–551.

[26] Kelly, F. P. (1993) *Reversibility and Stochastic Networks.* Chichester: Wiley.

[27] Kelly, F. P. (1993) Loss networks, *Annal. Appl. Prob.*, **1**, 319–378.

[28] Khinchine, A. Ja. (1993) Uber die mit tlere Dauer des Stillistandes von Mashinen, *Mat. Sbornik*, **40**, 119–123 (Russian; German summary).

[29] Knessl, C., Matkowsky, B. J., Schuss, Z. and Tier, C. (1986) On the performance of state-dependent single-server queues, *SIAM J. Appl. Math.*, **46**, 657–697.

[30] Knessl, C., Matkowsky, B. J., Schuss, Z. and Tier, C. (1987) Asymptotic expansions for a closed multiple access system, *SIAM J. Comput.*, **16**, 378–398.

[31] Knessl, C. and Tier, C. (1990) Asymptotic expansions for large closed queueing networks, *J. ACM*, **37**, 144–174.

[32] Knuth, D. E. (1969) *The Art of Computer Programming* (Vol. 1). Reading MA: Addison-Wesley.

[33] Kobayashi, H. (1978) *Modeling and Analysis: An Introduction to System Performance Evaluation Methodology.* Reading MA: Addison-Wesley.

[34] Kogan, Y. (1989) Exact analysis for a class of simple, circuit-switched networks with blocking, *Adv. Appl. Prob.*, **21**, 952–955.

[35] Kogan, Y. (1992) Another approach to asymptotic expansions for large closed queueing networks, *Operations Research Letters*, **11**, 317–321.

[36] Kogan, Y. and Birman, A. (1992) Asymptotic analysis of closed queueing networks with bottlenecks, IFIP Transactions C-5, performance of distributed systems and integrated communication networks, In: Hasegawa, T., Takagi, H. and Takahashi, Y. (Eds.), *Proceedings of the IFIP WG 7.3 International Conference on the Performance of Distributed Systems and Integrated Communication Networks*, Kyoto, Japan, 10–12 Sept., 1991, North-Holland, Amsterdam, 1992, pp. 265–280.

[37] Kogan, Ya. A. and Boguslavsky, L. B. (1985) Performance analysis of memory interference in multiprocessors with private cache memories, *Performance Evaluation*, **5**, 97–104.

[38] Kogan, Y. and Shenfild, M. (1994) Asymptotic solution of generalized multiclass Engset model, In: Labetoulle, J. and Roberts, J. W. (Eds.), *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks, Proc. ITC 14* (Vol. 1b). Amsterdam: Elsevier, pp. 1239–1249.

[39] Kogan, Y. A. and Signaevsky, V. A. (1990) *Performance Evaluation of High-Speed Computers.* Moscow: Nauka [in Russian].

[40] Kogan, Y. and Yakovlev, A. (1996) Asymptotic analysis for closed multichain queueing networks with bottlenecks, *Queueing Systems*, **23**, 235–258.

[41] Malyshev, V. A. and Yakovlev, Y. V. (1993) Condensation in large closed Jackson networks, *Rapports de Recherche de l'INRIA-Rocquencourt*, **1854**.

[42] McKenna, J., Mitra, D. and Ramakrishnan, K. G. (1981) A class of closed Markovian queueing networks: Integral representations, asymptotic expansions and generalizations, *Bell Syst. Tech. J.*, **60**, 599–641.

[43] McKenna, J. and Mitra, D. (1982) Integral representation and asymptotic expansions for closed Markovian queueing networks: Normal usage, *Bell Syst. Tech. J.*, **61**, 661–683.

[44] Mitra, D. (1987) Asymptotic analysis and computational methods for a class of simple, circuit-switched networks with blocking, *Adv. Appl. Prob.*, **19**, 219–239.

[45] Mitra, D. and Morrison, J. A. (1994) Erlang capacity of a shared resource, In: Labetoulle, J. and Roberts, J. W. (Eds.), *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks, Proc. ITC 14* (Vol. 1b). Amsterdam: Elsevier, pp. 875–885.

[46] Mitra, D. and Morrison, J. A. (1994) Erlang capacity and uniform approximations for shared unbuffered resources, *IEEE/ACM Transactions on Networking*, **2**, 558–570.

[47] Morrison, J. A., Ramakrishnan, K. G. and Mitra, D. (1998) Refined asymptotic approximations to loss probabilities and their sensitivities in shared unbuffered resources, *SIAM J. Appl. Math.*, **59**, 494–513.

[48] Morrison, J. A. (1994) Loss probabilities in a simple circuit-switched network, *Adv. Appl. Prob.*, **26**, 456–473.

[49] Olver, F. W. J. (1974) *Asymptotics and Special Function.* New York: Academic Press.

[50] Palm, C. (1943) Intensitatschwankungen im Fernsprechverkehr, *Ericson Techniks*, **6**, 1–189.

[51] Pinsky, E. (1992) A simple approximation for the Erlang loss function, *Performance Evaluation*, **15**, 155–161.

[52] Pinsky, E. and Conway, A. (1991) Exact computation of blocking probabilities in state-dependent multifacility blocking models, In: Hasegawa, T., Takagi, H. and Takahashi, Y. (Eds.), *Proc. Intern. Conf. Performance of Distributed Systems and Integrated Communication Networks.* Kyoto, pp. 353–362.

[53] Pittel, B. (1979) Closed exponential networks of queues with saturation: The Jackson-type stationary distribution and its asymptotic analysis, *Math. Oper. Res.*, **6**, 357–378.

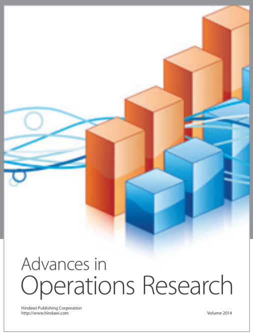
[54] Range, R. M. (1986) *Holomorphic Functions and Integral Representations in Several Complex Variables.* New York: Springer-Verlag.

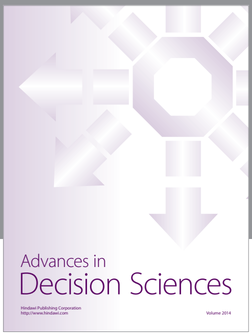
[55] Simonian, A. (1992) Analyse Asymptotique des Taux de Blocage pour un Traffic Multidebit, *Ann. Telecommun.*, **47**, 56–63.

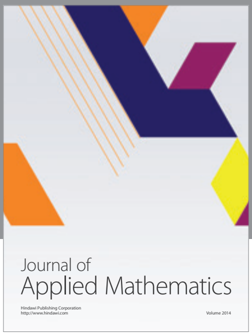
[56] Syski, R. (1986) *Introduction to Congestion Theory in Telephone Traffic* (2nd ed.). Amsterdam: North-Holland.

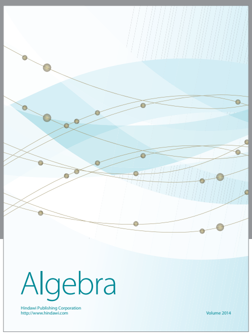
[57] Whittle, P. (1988) Approximation in large-scale circuit-switched networks, *Prob. Eng. Inform. Sci.*, **2**, 279–291.

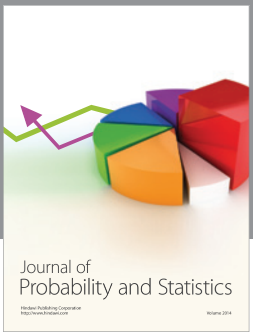
[58] Wong, R. (1986) *Asymptotic Approximations of Integrals.* Boston: Academic Press.


Advances in
Operations Research

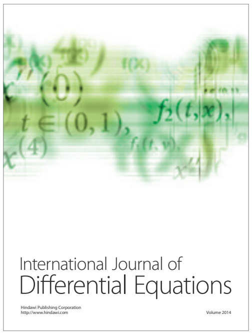

Advances in
Decision Sciences

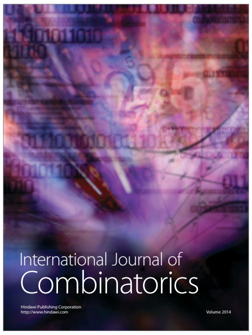

Journal of
Applied Mathematics

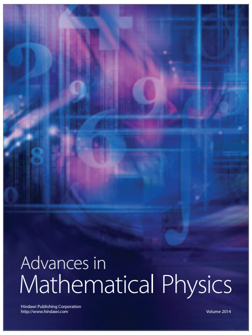

Algebra

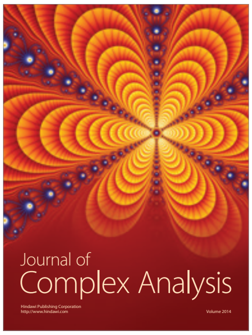

Journal of
Probability and Statistics

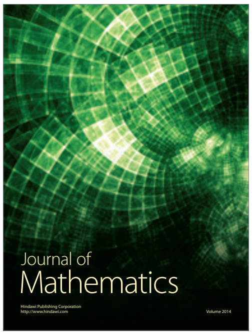
The Scientific
World Journal

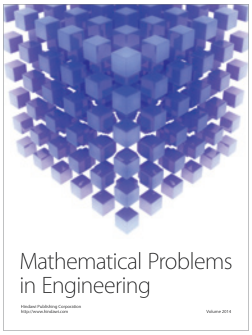

International Journal of
Differential Equations

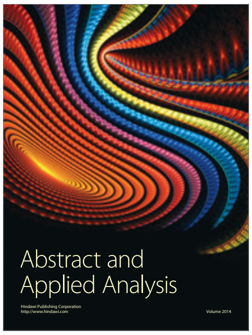

International Journal of
Combinatorics

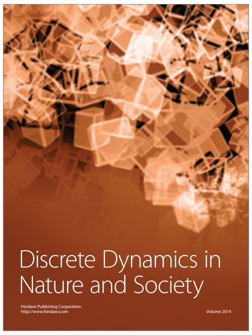

Advances in
Mathematical Physics

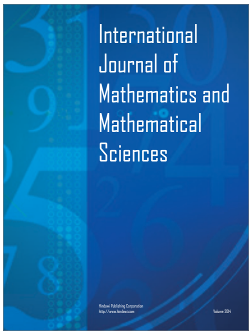

Journal of
Complex Analysis


Journal of
Mathematics


Mathematical Problems
in Engineering

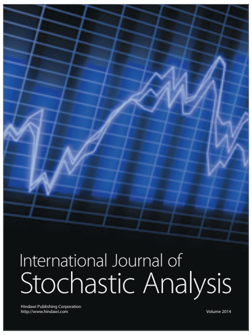

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


International Journal of
Stochastic Analysis

Journal of
Optimization