APPLICATIONS OF SINGULAR PERTURBATION METHODS IN QUEUEING

Charles Knessl and Charles Tier

Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
851 South Morgan St.
Chicago, IL 60607-7045

ABSTRACT

A survey is presented describing the application of singular perturbation techniques to queueing systems. The goal is to compute performance measures by constructing approximate solutions to specific problems involving either the Kolmogorov forward or backward equation which contain a small parameter. These techniques are particularly useful on problems for which exact solutions are not available. Four different classes of problems are surveyed: (i) state-dependent queues; (ii) systems with a processor-sharing server; (iii) queueing networks; (iv) time dependent behavior. For each class, an illustrative example is presented along with the direction of current research.

1 Introduction

In analyzing queueing models, one would like to compute certain performance measures, such as the steady state queue length distribution, transient queue length distribution, mean length of a busy period, unfinished work distribution, sojourn time distribution, time for the queue to reach some specified number, etc. For a specific model, these quantities may all be characterized as solutions to certain equations. Thus, computing the performance measures amounts to solving these equations together with appropriate boundary/initial conditions. Given a Markov process $X(t)$, the transition probability density $p(x,t|x_0,t_0)$=Pr$[X(t) = x|X(t_0) = x_0]$ satisfies the forward and backward Kolmogorov equations, which we write in an abstract form as

$$\frac{\partial p}{\partial t} = L_{x,t}p \quad (t > t_0), \quad p(x,t_0^+|x_0,t_0) = \delta(x-x_0) \tag{1.1}$$

$$-\frac{\partial p}{\partial t_0} = L^*_{x,t_0}p \quad (t_0 < t), \quad p(x,t_0|x_0,t_0^-) = \delta(x-x_0). \tag{1.2}$$

Here L is a linear operator that involves the variable x and time t, and L^* is its adjoint. The δ in (1.1)-(1.2) is the Kronecker delta if the state space is discrete and the Dirac delta if the state space is continuous. The precise forms of L and L^* depend on whether we are looking at a discrete model (such as the number $N(t)$ of customers in an $M/M/1$ queue) or a continuous model (such

This research was supported in part by NSF Grant DMS-93-00136 and DOE Grant DE-FG02-93ER25168

as the unfinished work $U(t)$ in an $M/G/1$ queue), or some combination of these. For example, in considering the joint queue length distribution in a network of Markovian queues, L is generally a multi-dimensional difference operator. If the model is space and time homogeneous, which occurs say if the arrival and service rates are constant, then L is a constant coefficient operator. However, the form of L is generally different near the boundaries of the state space, so that inherent to the problems (1.1)-(1.2) are complicated sets of boundary conditions. These make it difficult to obtain **simple, exact** solutions to (1.1)-(1.2) for all but the simplest of queueing models.

For the unfinished work process $U(t)$, the operator L is an integro-differential operator since the process has non-local transitions. The difficulty in solving (1.1)-(1.2) has led to the introduction of approximations. A popular tool in queueing theory is the use of diffusion approximations. Here one replaces the original process $X(t)$ by a diffusion process $\tilde{X}(t)$ (time may need to be scaled). Then, computing $\tilde{p}$, the transition density for $\tilde{X}$, involves solving (1.1)-(1.2) with L now being a partial differential operator of second order. This may itself be a difficult task since the boundary conditions associated with the approximate problem are frequently still complicated, especially for models in more than one space dimension.

We have thus far discussed (1.1)-(1.2) in terms of Markovian models. For models with general interarrival time and/or service time distributions, the processes of interest are no longer Markovian but may be imbedded in a (higher-dimensional) Markov process by using the method of supplementary variables (see [8]). For the new process, we can again obtain (1.1)-(1.2) except that L now depends on (x, y), with y being the vector of supplementary variables, and is usually more complicated than the L associated with exponential arrivals/service. Even for non-Markovian models, it is still generally easy to derive the appropriate problems (1.1)-(1.2); the difficult task is the solution of the equations.

In the queueing literature, exact solutions to (1.1)-(1.2) are generally available only for steady state $(t \to \infty)$ problems in one and two dimensions and for time-dependent problems in one dimension. For example, the steady state distribution of $N(t)$ for the $M/M/1$ queue has a very simple form, but the form for the transient distribution is a complicated expression involving an infinite sum of Bessel functions. For the $M/G/1$ queue, there is an explicit expression for the Laplace transform of the steady state distribution of $N(t)$ (or $U(t)$), though the transform cannot be (analytically) inverted for general service time densities. The time dependent distribution is very complicated. Its double transform (over space and time) can be characterized in terms of the solution to a functional equation, but there is no hope of inverting the transform. For the $GI/G/1$ model, solving for even the steady state distribution of $U(t)$ is equivalent to solving a Wiener-Hopf integral equation with a general kernel. Solution of this problem can be expressed in terms of two complex contour integrals, one for inverting a Laplace transform and the other for the analytical solution of the Wiener-Hopf problem. This shows that even for one-dimensional models, one cannot obtain simple analytic expressions for the various performance measures. For problems in more than one dimension, the situation is even worse. Jackson networks, or, more generally, "product-form" networks, are an important class of multi-dimensional models for which one can explicitly obtain the steady state queue length distribution. However, obtaining time-dependent information for these models is much harder. Even for Jackson networks, it is difficult to analyze the busy period and various other "first-passage time" problems. Solutions to non-product form networks are complicated even in two dimensions under Markovian assumptions. Using transforms and function-theoretic arguments, these models ([6]) may be reduced to solving certain classic problems in the theory of singular integral equations such as Dirichlet and Riemann-Hilbert problems. However, one is again left with inverting a two-dimensional transform, which itself is often characterized in a form that is not particularly explicit.

From the above (brief and incomplete) summary of exactly solvable models, it is clear that approximations must play a major role in the analysis of queues. Here we examine a set of methods called "asymptotic and singular perturbation techniques". Their role in queueing theory is basically twofold. First, they can be used to simplify exact solutions when these are available. Since the exact solutions discussed above are extremely complicated, it is useful to evaluate these expressions in certain limiting cases, in order to gain more insight into the qualitative structure of the particular model. Asymptotic formulas often clearly show the dependence of the solutions on the various variables/parameters in the problem, whereas the full exact expressions may be difficult to interpret in terms of the underlying model. Of course, an asymptotic formula can never contain as much quantitative (numerical) information as an exact answer, but it can provide reasonably accurate numerical results at a greatly reduced computational cost. Also, a queueing model is itself an approximation to a physical system. Thus, we believe that obtaining qualitative information is just as important as obtaining accurate numerical values.

A second, and we believe more important, aspect of using perturbation methods is to make progress on problems for which exact solutions are not available. Since it is likely that most models in queueing theory (and indeed in any other area of applied mathematics) will never be solved exactly, obtaining useful approximations is very important. What do we mean by "useful?" The two main criteria for the usefulness of an asymptotic approximation are (i) its numerical accuracy and (ii) its ability to make transparent qualitative properties of the solution. It is desirable to have both (i) and (ii), but it is much better to have either (i) or (ii) than to have nothing. The verification of (i) can only be obtained by comparing the approximation to exact results or to numerical approximations, assuming the latter are reliable. Deciding on (ii) is somewhat harder as what looks complicated to one person may look simple to another. We believe that if an asymptotic result can be expressed in terms of elementary functions (sines, exponentials, Gaussians) or well-studied special functions (Bessel, parabolic cylinder), it can be called "simple" enough to be useful. In deciding whether an asymptotic answer is simple, it is also appropriate to view this result in terms of the exact answer, and how complicated it is, or would be (if it could be obtained at all).

Using asymptotics to simplify exact expressions usually involves the approximate evaluation of sums and integrals, using ideas such as Laplace's method, the method of steepest descent (saddle point method), the Euler-MacLaurin formula, Poisson summation, integral representations of sums, etc. These have been used to good effect in queueing theory; examples are [49], where the authors obtained asymptotic expansions for product-form networks for large population sizes, and [54], where integral representations and subsequent asymptotics were used on processor-sharing models. Of course, these methods assume that one has an exact (and sufficiently explicit) representation for the quantity to be evaluated.

Many other asymptotic methods exist which may be applied directly to the equation(s) satisfied by the given performance measure. They do not rely on having an exact solution so that they are clearly applicable to a much wider class of problems. These methods include the WKB method, "boundary layer" techniques, the ray method, and the method of matched asymptotic expansions. They are called "singular perturbation" techniques (general references are [1,15]). To see what is meant by singular, consider a function $f(x;\varepsilon)$ which depends on the variable(s) x and an additional (small) parameter ε. If f is an analytic function of ε, i.e. if the series $\sum_{j=0}^{\infty} f_j(x)\varepsilon^j$ converges for $|\varepsilon|$ sufficiently small, then such a series is called a "regular" perturbation series. If either the series diverges (but is still asymptotic for $\varepsilon \to 0$) or the expansion involves a more complicated asymptotic sequence than powers of ε, then the perturbation series is called "singular". Most interesting problems in applied mathematics are of singular perturbation type, and this we believe is also true in queueing theory. An important example of a singular perturbation series is the WKB form $e^{-\phi(x)/\varepsilon} \sum_{j=0}^{\infty} A_j(x)\varepsilon^j$, which we shall show arises naturally even in very elementary queueing models. Such a series is generally divergent, but even the leading term $(A_0(x)e^{-\phi(x)/\varepsilon})$ is usually an excellent approximation to the quantity that is to be computed.

Another important feature of singular perturbation methods is that such problems tend to contain several scales, which must be treated separately. This leads to several different asymptotic expansions which must be related to one another, and this is usually done by the "asymptotic matching principle". For example, if we consider the m-server $M/M/m$ queue in the limit $m = \varepsilon^{-1} \to \infty$, it is necessary to construct different asymptotic expansions for $N(t) = O(m)$ and $N(t) = O(1)$, which corresponds respectively to having a finite fraction of the servers occupied and to having just a few occupied servers.

The methods discussed above have usually been developed in the context of second order equations (ODEs and PDEs). Also, an individual method was usually introduced and developed in the context of a particular scientific application. The WKB method was first used as an approximation tool for solving the Schrödinger equation in quantum mechanics; the ray method was developed to solve problems in high frequency wave propagation; boundary layer ideas were first used in the study of viscous flow past obstacles. From our point of view, however, we consider these methods as mathematical techniques for approximately solving equations. They are useful for analyzing queueing models for which the operators L and L^* in (1.1)-(1.2) depend upon a small parameter; call it ε. The size of ε can be used to simplify the equations in a systematic way, and thus to obtain explicit, approximate formulas. Typically, ε measures the reciprocal of the size the system; e.g. $\varepsilon = m^{-1}$ where m is the number of servers in an m server queue, or ε may be the inverse of the customer population in a large, closed Jackson network. The most commonly used asymptotic approximations in queueing theory are light traffic and heavy traffic. For these, ε may be taken as the arrival rate and the difference between arrival and service rates, respectively.

For most models in queueing theory (1.1)-(1.2) involve difference equations, integral equations, delay equations and various combinations of these. This is because $N(t)$ takes on discrete values and $U(t)$ has jumps (non-local transitions) at arrival times. For these types of problems, singular perturbation methods are not as well developed as they are for ordinary and partial differential equations. Thus, developing methods for solving (1.1)-(1.2) will also enhance the scope of singular perturbation techniques, as they can now be applied to integral and other equations, and these arise naturally is many fields of science and engineering. Our primary goal is to compute solutions to specific problems that arise in queueing theory and in other stochastic models. However, the mathematical methodology should also be useful in other areas.

In the sections that follow, we will apply perturbation methods to some specific queueing models. We show that they are useful for several different classes of problems. In section 2, we consider queues which have state-dependent parameters. State-dependent (and time-dependent) queues lead to an operator L in (1.1)-(1.2) that is not "constant coefficient," and such problems are difficult (or impossible) to tackle using transform methods. In section 3, we consider several queues which have a processor-sharing server. We shall obtain approximations to the sojourn time through such systems. In section 4, we consider queuing networks. For product-form networks, there exist explicit expressions for the joint, steady state queue length distribution. We show how to compute this asymptotically, from a recursion which is satisfied by the normalization constant (partition function). We also show in section 4 how to use perturbation methods to analyze bottlenecks in networks. For the simplest Jackson network consisting of two $M/M/1$ queues in tandem, we compute the time until the network population becomes large, and then the time needed to settle back to its equilibrium state. For these first passage time problems, there seem to be no exact expressions available. In section 5, we show how to use the ray method to compute the time-dependent behavior of queueing models. As an example, we consider the Erlang loss model ($M/M/m/m$ queue).

2 State-dependent Queues

We shall first consider the classic repairman problem, which corresponds to the finite source (finite population) $M/M/1$ queue. Then we extend our results to systems which have a general service time distribution. We also consider an $M/G/1$ queue characterized by the unfinished work $U(t)$, and which has state-dependent arrivals and service.

2.1 Repairman problems

Denote the service rate by μ_0 and let $p_n(t) = \Pr[N(t) = n]$ with $p_n = p_n(\infty)$. If M is the total number of customers that are in the population, then this model corresponds to a queue with a state-dependent arrival rate equal to $\lambda(M - N(t))$. The steady state balance equations are

$$(\mu_0 + \lambda(M-n))p_n = \lambda(M+1-n)p_{n-1} + \mu_0 p_{n+1};\ 1 \le n \le M \tag{2.1}$$

$$\lambda M p_0 = \mu_0 p_1 \tag{2.2}$$

with $p_{M+1} \equiv 0$. Solving (2.1)-(2.2) and normalizing the probabilities, we easily obtain

$$p_n = p_0 \frac{M!}{(M-n)!}\left(\frac{\lambda}{\mu_0}\right)^n,\ p_0 = 1/\sum_{n=0}^{M}\frac{M!}{(M-n)!}\left(\frac{\lambda}{\mu_0}\right)^n. \tag{2.3}$$

Now consider the asymptotic limit $M \to \infty$, $\mu_0 = \mu M = O(M)$ which corresponds to systems with many customers and a fast server. Letting $\rho = \lambda/\mu = \lambda M/\mu_0 = O(1)$ and approximating $M!$ and $(M-n)!$ by Stirling's formula we obtain

$$p_n \sim \frac{p_0}{\sqrt{1-\xi}} e^{-M\phi(\xi)},\ \xi = \frac{n}{M} \tag{2.4}$$

$$\phi(\xi) = \xi \log\left(\frac{1}{\rho}\right) + \xi + (1-\xi)\log(1-\xi).$$

This approximation is valid for all $\xi \in [0, 1)$ as $M \to \infty$. It ceases to be valid when $\xi \approx 1$ (i.e. $M - n = O(1)$) but by then p_n is exponentially small. Next we use the integral representation

$$1/p_0 = \frac{M}{\rho}\int_0^\infty \exp\left(-M[\frac{x}{\rho} - \log(1+x)]\right) dx \tag{2.5}$$

to obtain the expansion for p_0. The integral in (2.5) is a Laplace type integral whose asymptotic expansion is easily obtained, and is different for $\rho > 1$, $\rho = 1$ and $\rho < 1$. We have

$$\begin{aligned} p_0 &\sim 1-\rho; \quad \rho < 1 \\ &\sim \sqrt{2/(\pi M)}; \quad \rho = 1 \\ &\sim \frac{1}{\sqrt{2\pi M}}\exp\left(M[1 - \frac{1}{\rho} + \log(\frac{1}{\rho})]\right); \quad \rho > 1. \end{aligned} \tag{2.6}$$

We could also obtain a result for $\rho \approx 1$, valid for $\rho - 1 = O(M^{-1/2})$, which will asymptotically match between the first and third formulas in (2.6). From (2.4) and (2.6) we observe that p_n is peaked at $\xi = 0$ if $\rho \leq 1$ and at $\xi = 1 - \rho^{-1}$ if $\rho \geq 1$. Also, the probability that the system is empty has the asymptotic orders of magnitude $O(1)$, $O(M^{-1/2})$, and $O(M^{-1/2}e^{-cM})$ $(-c = 1 - \rho^{-1} - \log \rho < 0)$ when $\rho < 1$, $\rho = 1$, and $\rho > 1$, respectively. This shows that the asymptotic structure of this problem is very sensitive to the value of ρ.

Now we present an alternate approach to the asymptotics, which uses only (2.1). We scale $n = M\xi$ with $p_n = P(\xi)$ to get

$$(1 + \rho(1 - \xi))P(\xi) = \rho(1 - \xi + \varepsilon)P(\xi - \varepsilon) + P(\xi + \varepsilon) \tag{2.7}$$

where $\varepsilon \equiv M^{-1}$. This is a difference equation with small differences which resembles a singularly perturbed ODE, as can be seen by expanding the right side of (2.7) for small ε. We seek solutions of (2.7) in the WKB form

$$P(\xi) = C(\varepsilon)e^{-\phi(\xi)/\varepsilon}[A_0(\xi) + \varepsilon A_1(\xi) + \cdots]. \tag{2.8}$$

The constant $C(\varepsilon)$ will be determined by normalization and we set $A_0(\xi) = A(\xi)$. Using (2.8) in (2.7) and expanding for small ε, we obtain, at the first two orders, the equations

$$1 + \rho(1 - \xi) = \rho(1 - \xi)e^{\phi'(\xi)} + e^{-\phi'(\xi)} \tag{2.9}$$

and

$$0 = \rho e^{\phi'(\xi)}[A(\xi) - (1 - \xi)(A'(\xi) + \frac{1}{2}\phi''(\xi)A(\xi))] + e^{-\phi'(\xi)}[A'(\xi) - \frac{1}{2}\phi''(\xi)A(\xi)]. \tag{2.10}$$

Equation (2.9) is a nonlinear ODE for $\phi(\xi)$, but it is very easy to solve since it is a quadratic equation for $e^{\phi'(\xi)}$. One root of this quadratic is clearly $\phi' = 0$ and the other is

$$\phi'(\xi) = -\log \rho - \log(1 - \xi)$$

which integrates to

$$\phi(\xi) = -\xi \log \rho + \xi + (1 - \xi)\log(1 - \xi) \tag{2.11}$$

where we have chosen $\phi(0) = 0$, since $\phi(0)$ can be incorporated into the constant $C(\varepsilon)$ in (2.8). With (2.11), (2.10) is a **linear, first order** ODE for $A(\xi)$. It is easily solved (up to a multiplicative constant) to yield

$$A(\xi) = (1 - \xi)^{-1/2}. \tag{2.12}$$

The solution $\phi' = 0$ must be rejected since it would lead to a non-integrable $A(\xi)$. Now we have the approximation $P(\xi) \sim C(\varepsilon)Ae^{-\phi/\varepsilon}$ and it remains only to determine the constant $C(\varepsilon)$, which can be done from the normalization

$$C(\varepsilon)\sum_{n=0}^{1/\varepsilon} e^{-\phi(\varepsilon n)/\varepsilon}[A(\varepsilon n) + \varepsilon A_1(\varepsilon n) + \cdots] = 1.$$

This sum may be approximated for $\varepsilon \to 0$ using Laplace's method for sums and the Euler-MacLaurin formula. Using (2.11)-(2.12) we would find that $C(\varepsilon)(\sim p_0)$ is again asymptotically given by (2.6) for the 3 cases of ρ.

Expressions (2.11)-(2.12) agree precisely with (2.4), which we obtained using the exact result and Stirling's formula. To obtain corrections to this leading order approximation we could use the full Stirling series to approximate $M!$ and $(M-n)!$ in (2.3), and then obtain the full asymptotic series for the integral in (2.5) using Laplace's method. Alternately, we could continue the expansion of (2.7) using (2.8). The correction terms $A_j(\xi)$, for $j \geq 1$, will satisfy linear ODEs of the same form as (2.10), and these are easy to solve.

What have we gained from our direct approach of using (2.1) (or (2.7)) instead of the full solution? Equation (2.1) is a linear second order difference equation whereas (2.9) is non-linear and first order and (2.10) is linear and first order. It turns out to be slightly easier to solve (2.9)-(2.10) than (2.1). The simplification is not so dramatic for this simple model, but the advantages of the direct approach will be much more apparent when we deal with general service time distributions, transient problems, and problems in more than one dimension. It turns out that it is frequently much easier to solve PDEs of the first order than it is to solve PDEs or difference equations of second order. This is true even if the first order problems are non-linear.

Next we consider the same model but with a general service time distribution, whose density we denote by $\tilde{b}(x)dx =$Pr[service time $\in (x, x+dx)$]. We allow $\tilde{b}$ to be a delta function so that this analysis also applies to $M/D/1$ models. Now $N(t)$ is no longer Markov, so we consider the process $(N(t), Y(t))$, where the supplementary variable $Y(t)$ measures the elapsed service time of the customer presently being served. We let

$$\begin{aligned} p_n(y,t)dy &= \text{Pr}[N(t)=n, Y(t) \in (y, y+dy)],\ n \geq 1 \\ p_0(t) &= \text{Pr}[N(t)=0] \end{aligned}$$

and denote the steady-state limits by $p_n(y)$, p_0; and the marginal (steady-state) queue length distribution by $p_n = \int_0^\infty p_n(y)dy$ $(n \geq 1)$. The balance equations are now

$$p_n'(y) = \lambda(M-n+1)p_{n-1}(y) - [\lambda(M-n) + \tilde{\mu}(y)]p_n(y);\ 2 \leq n \leq M \tag{2.13}$$

$$p_n(0) = \int_0^\infty p_{n+1}(y)\tilde{\mu}(y)dy;\quad 2 \leq n \leq M-1 \tag{2.14}$$

where $\tilde{\mu}(y) = \tilde{b}(y)/\int_y^\infty \tilde{b}(x)dx$ is the service rate, conditioned on the elapsed service time. The boundary conditions turn out to be

$$p_1'(y) = -[\lambda(M-1) + \tilde{\mu}(y)]p_1(y) \tag{2.15}$$

$$\lambda M p_0 = \int_0^\infty p_1(y)\tilde{\mu}(y)dy \tag{2.16}$$

$$
\begin{aligned}
p_1(0) &= \lambda M p_0 + \int_0^\infty p_2(y)\tilde{\mu}(y)dy && (2.17)\\
p_M(0) &= 0 && (2.18)
\end{aligned}
$$

$$
p_0 + \sum_{n=1}^{M} \int_0^\infty p_n(y)dy = 1. \qquad (2.19)
$$

To solve this system asymptotically, we again assume that M is large and that service times tend to be small. To make the latter more precise we write the service density in the scaled form $\tilde{b}(x) = Mb(Mx)$ so that the mean service time is $O(M^{-1})$. Setting $\tilde{\mu}(y) = M\mu(My)$, we introduce into (2.13)-(2.19) the scaled variables $\xi = n/M$, $\eta = My$, $\varepsilon = M^{-1}$ with $p_n(y)dy = P(\xi,\eta)d\eta$ and obtain the scaled problem

$$
\begin{aligned}
P_\eta(\xi,\eta) &= \lambda(1-\xi+\varepsilon)P(\xi-\varepsilon,\eta) - [\lambda(1-\xi)+\mu(\eta)]P(\xi,\eta) && (2.20)\\
P(\xi,0) &= \int_0^\infty P(\xi+\varepsilon,\eta)\mu(\eta)d\eta. && (2.21)
\end{aligned}
$$

For the moment we ignore the boundary conditions (2.15)-(2.19).

When service times were exponential, we needed two expansions for $P(\xi)$, valid on the respective scales $0 \le \xi < 1$ and $1-\xi = O(\varepsilon)$ ($M-n = O(1)$). For the present model it is necessary to consider 3 scales:
(i) $0 < \xi < 1$ (ii) $\xi = O(\varepsilon)$ ($n = O(1)$) (iii) $1-\xi = O(\varepsilon)$ ($M-n = O(1)$). The third scale is again unimportant, since the distribution is exponentially small there for any value of

$$
\rho \equiv \lambda M \int_0^\infty x\tilde{b}(x)dx = \lambda \int_0^\infty x b(x)dx \equiv \frac{\lambda}{\mu} = O(1).
$$

We proceed to asymptotically solve the problem on scales (i) and (ii), and then relate the two expansions by asymptotic matching. After we have covered the entire state space, there will remain one undetermined constant, and this is obtained by normalization (2.19).

When $0 < \xi < 1$ and $M \to \infty$, we set

$$
P(\xi,\eta) = e^{-\int_0^\eta \mu(z)dz} C(\varepsilon) \exp\Big\{-\Big[\frac{1}{\varepsilon}\phi_0(\xi,\eta) + \phi_1(\xi,\eta) + \varepsilon\phi_2(\xi,\eta) + \cdots\Big]\Big\}. \qquad (2.22)
$$

The first factor is included for convenience. The form (2.22) is asymptotically equivalent to (2.8) if we identify $A_0 = e^{-\phi_1}$, $A_1 = -A_0\phi_2$, etc. We shall only compute the leading term in (2.22), which means that we must compute ϕ_0 **and** ϕ_1. Using (2.22) in (2.20) and expanding for $\varepsilon \to 0$ we obtain at the first two orders

$$
\begin{aligned}
-\phi_{0,\eta} &= 0 && (2.23)\\
-\phi_{1,\eta} &= \lambda(1-\xi)(e^{\phi_{0,\xi}} - 1). && (2.24)
\end{aligned}
$$

Thus ϕ_0 is independent of η and we write $\phi_0 = \phi_0(\xi)$. Then the right side of (2.24) depends on ξ only, which we denote by $K(\xi)$, and then

$$\phi_1(\xi, \eta) = -\eta K(\xi) + L(\xi) \tag{2.25}$$

where K, L are as yet undetermined. Using (2.22) with (2.25) in (2.21), we obtain at leading order the following equation for K

$$1 + \frac{K(\xi)}{\lambda(1-\xi)} = \int_0^\infty e^{\eta K(\xi)} b(\eta) d\eta \tag{2.26}$$

which is a transcendental equation that involves the **scaled** service density $b(\cdot)$. Explicit solutions to (2.26) can be obtained for exponential, E_2, and H_2 servers. In the general case (2.26) is easily solved numerically, and a convexity argument shows that (2.26) has a unique non-zero solution.

To determine L (and hence ϕ_1), we must examine the third term in the expansion of (2.20) and then use the result in (2.21). Omitting the details, the final result is a linear ODE for L:

$$[1-\lambda(1-\xi)I_1(\xi)]L'(\xi) = \frac{1}{2}\phi_0''(\xi) + \lambda[1-\frac{1}{2}(1-\xi)\phi_0''(\xi)]I_1(\xi) - \frac{1}{2}K'(\xi)\lambda(1-\xi)I_2(\xi); \tag{2.27}$$

$$I_j(\xi) = \int_0^\infty \eta^j e^{\eta K(\xi)} b(\eta) d\eta.$$

Since ϕ_0, K, I_1, I_2 are known via (2.26), (2.27) is easily integrated. After some algebra, we find that the leading term in (2.22) is

$$\begin{aligned} P(\xi,\eta) &\sim e^{-\int_0^\eta \mu(z)dz} C(\varepsilon) e^{-\phi_0(\xi)/\varepsilon} e^{\eta K(\xi)} e^{-L(\xi)} \\ &= C(\varepsilon) e^{-\int_0^\eta \mu(z)dz} e^{\eta K(\xi)} \\ &\quad \times \left| \frac{(1-\xi)K'(\xi)}{\lambda(1-\xi)+K(\xi)} \right|^{1/2} \exp\left\{ -\frac{1}{\varepsilon}\int_0^\xi \log\left[1+\frac{K(z)}{\lambda(1-z)}\right] dz \right\}. \end{aligned} \tag{2.28}$$

This expansion is not valid for $\xi = O(\varepsilon)$ ($n = O(1)$) since it does not satisfy the boundary equations (2.15)-(2.18).

To obtain an appropriate expansion for $n = O(1)$, we go back to the discrete space variable n and consider (2.13)-(2.14) for $n = O(1)$. Since $M \to \infty$, the upper boundary disappears. If we expand for $n = O(1)$,

$$p_n(y) = D(\varepsilon)[Q_n(\eta) + \varepsilon Q_n^{(1)}(\eta) + \varepsilon^2 Q_n^{(2)}(\eta) + \cdots]$$

then at the leading order we get

$$Q_n'(\eta) = \lambda Q_{n-1}(\eta) - [\lambda + \mu(\eta)]Q_n(\eta);\ n \geq 2 \tag{2.29}$$

$$Q_n(0) = \int_0^\infty Q_{n+1}(\eta)\mu(\eta)d\eta;\ n \geq 2. \tag{2.30}$$

Now we must consider the boundary equations (2.15)-(2.17), which imply that

$$\begin{aligned} Q_1'(\eta) &= -[\lambda + \mu(\eta)]Q_1(\eta) \\ \lambda Q_0 &= \int_0^\infty Q_1(\eta)\mu(\eta)d\eta \\ Q_1(0) &= \lambda Q_0 + \int_0^\infty Q_2(\eta)\mu(\eta)d\eta. \end{aligned} \tag{2.31}$$

The problem (2.29)-(2.31) is simpler than the original problem (2.13)-(2.18) in two respects. First, the domain of (2.29)-(2.31) is $n \geq 0$, so that we have a problem on an infinite interval rather than the finite interval $0 \leq n \leq M$. Second, for $n = O(1)$, we can to leading order approximate quantities such as $\lambda(M-n) \approx \lambda M$ so that (2.29)-(2.31) is basically a "constant coefficient" problem in n, which is easy to solve using transforms (generating functions). If $\rho < 1$, then the equations (2.29)-(2.31) are precisely those satisfied by the steady-state probabilities in the standard (∞ population) $M/G/1$ queue. If $\rho > 1$, the $M/G/1$ model is transient so that $Q_n(\eta)$ can no longer be interpreted probabilistically. It is simply an approximation to the finite source model valid on the scale $n = O(1)$. The correction terms $Q_n^{(j)}(\eta)$ $(j \geq 1)$ will satisfy inhomogeneous versions of the problem (2.29)-(2.31), and these can also be solved using generating functions. For the leading term we obtain

$$Q_n(\eta) = \frac{\lambda Q_0}{2\pi i}\left\{\int_C \frac{(s-1)e^{\lambda(s-1)\eta}}{s^n[s - \widehat{b}(\lambda - \lambda s)]}ds\right\}e^{-\int_0^\eta \mu(z)dz}. \tag{2.32}$$

Here $\widehat{b}(s) = \int_0^\infty e^{-sz}b(z)dz$ and C is a small loop about $s = 0$ in the complex plane.

Now, the expansions for $n = O(1)$ and $0 < \xi = n/M < 1$ contain the hitherto undetermined constants $D(\varepsilon), C(\varepsilon)$. One of these can be determined by normalization, but we need one additional condition. This is obtained by requiring that the two expansions "asymptotically match." This means that they should agree on an intermediate scale where $\varepsilon \ll \xi \ll 1$, which corresponds to $n \to \infty$, $n/M \to 0$. Symbolically the matching condition may be written as

$$C(\varepsilon)\exp\Big[-\frac{1}{\varepsilon}\sum_{k=0}^\infty \varepsilon^k\phi_k(\xi,\eta)\Big]\Big|_{\xi\ll 1} \sim D(\varepsilon)\sum_{k=0}^\infty \varepsilon^k Q_n^{(k)}(\eta)\Big|_{n\gg 1} e^{\int_0^\eta \mu(z)dz} \tag{2.33}$$

and must hold to all orders in ε. The left side of (2.33) is evaluated by expanding (2.28) as $\xi \to 0$. To leading order this gives

$$C(\varepsilon)e^{\eta K(0)}\left|\frac{K'(0)}{\lambda + K(0)}\right|^{1/2}\left(1 + \frac{K(0)}{\lambda}\right)^{-n} \tag{2.34}$$

where we have used $\xi = \varepsilon n$. From (2.32), the large n behavior is determined by the singularity of the integrand that is closest to $s = 0$. This is a simple pole at $s = S^*$, which satisfies $S^* > 1$

($S^* < 1$) according as $\rho < 1$ ($\rho > 1$). Computing the residue at this pole we see that to leading order the right side of (2.33) becomes

$$D(\varepsilon)\lambda Q_0 \frac{(S^*-1)e^{\lambda(S^*-1)\eta}}{\lambda\int_0^\infty ze^{\lambda(S^*-1)z}b(z)dz - 1}(S^*)^{-n}. \tag{2.35}$$

From the definitions of $K(\xi)$ and S^*, we have

$$1 + \frac{K(0)}{\lambda} = S^*$$

so that the forms of (2.34) and (2.35) agree and the constants C, D are related by

$$C(\varepsilon) = D(\varepsilon)Q_0|K'(0)[\lambda + K(0)]|^{1/2}. \tag{2.36}$$

Our final step is to determine $C(\varepsilon)$ (or $D(\varepsilon)$) from normalization (2.19). This requires that we asymptotically evaluate the sum in (2.19) for $\varepsilon \to 0$ ($M \to \infty$). The expansion of the sum depends on whether $\rho < 1$ or $\rho > 1$. Below we summarize our final results for the marginal probabilities $p_n = \int_0^\infty p_n(y)dy$

(a) $0 < \xi < 1$

$$\begin{aligned}
p_n &= P(\xi) \sim \frac{C(\varepsilon)}{\lambda}\left|\frac{K'(\xi)}{\lambda(1-\xi)+K(\xi)}\right|^{1/2}\frac{e^{-M\phi(\xi)}}{\sqrt{1-\xi}};\\
\phi(\xi) &= \int_0^\xi \log\Big[1 + \frac{K(z)}{\lambda(1-z)}\Big]dz;\\
1 + \frac{K(\xi)}{\lambda(1-\xi)} &= \int_0^\infty e^{\eta K(\xi)}b(\eta)d\eta;\\
C(\varepsilon) &\sim \Big(1-\rho\Big)|K'(0)[\lambda+K(0)]|^{1/2}, \quad \Big(\rho = \frac{\lambda}{\mu} < 1\Big);\\
C(\varepsilon) &\sim \lambda\sqrt{\frac{2}{\pi M}} \quad (\rho = 1);\\
C(\varepsilon) &\sim \sqrt{\frac{\lambda\mu}{2\pi M}}\exp\Big[M\phi(1-\frac{1}{\rho})\Big] \quad (\rho > 1)
\end{aligned}$$

(b) $\xi = O(M^{-1})$ i.e. $n = O(1)$

$$\begin{aligned}
p_n &\sim \frac{1}{2\pi i}\Big\{\int_C \frac{-1+\widehat{b}(\lambda-\lambda s)}{s-\widehat{b}(\lambda-\lambda s)}s^{-n}ds\Big\}\frac{C(\varepsilon)}{|K'(0)[\lambda+K(0)]|^{1/2}}, \; n \geq 1\\
p_0 &\sim \frac{C(\varepsilon)}{|K'(0)[\lambda+K(0)]|^{1/2}}.
\end{aligned}$$

We remark that if $b(z) = \mu e^{-\mu z}$ then $K(\xi) = \mu - \lambda(1-\xi)$ and (a), (b) reduce to (2.4) with (2.6). When $\rho \geq 1$ we have

$$K(1-\rho^{-1}) = 0 \;\text{ and }\; K'(1-\rho^{-1}) = (2\lambda)/(\mu^2 m_2)$$

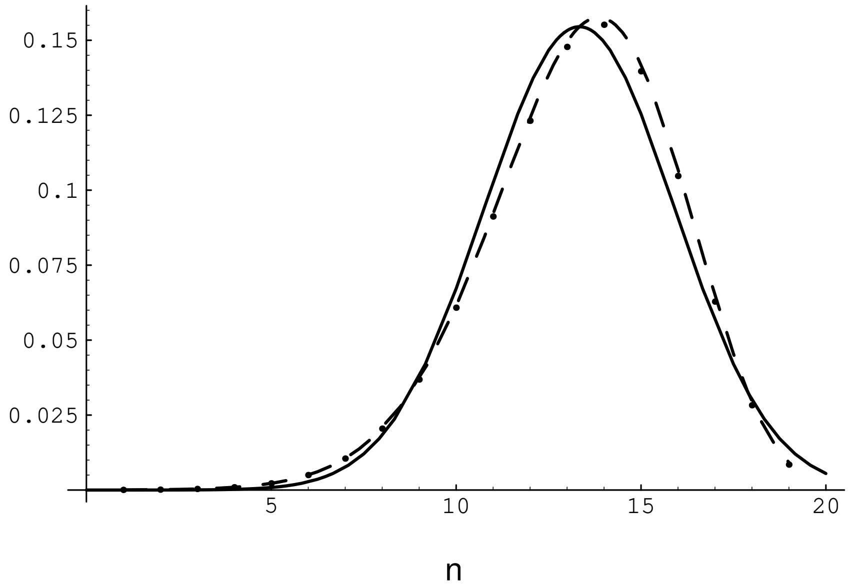

Figure 1: Graphs of $p_n^{EXA} = \cdots$, $p_n^{WKB} = - - -$, and $p_n^{DIFF} =$ —, with $\rho = 3.0$ and $M = 20$.

where m_2 is the second moment of the (scaled) service time density $b(\cdot)$. Thus, if we expand our approximation for $\rho > 1$ for ξ close to $\xi^* \equiv 1 - \rho^{-1}$, we obtain the Gaussian form

$$p_n \approx \sqrt{\frac{\lambda}{\pi\mu^3 m_2 M}}\, \exp\Big[-\frac{\lambda M}{\mu^3 m_2}(\xi - \xi^*)^2\Big] \tag{2.37}$$

which corresponds to the "diffusion approximation" to the process $N(t)$. Expression (2.37) is only valid for $\xi - \xi^* = O(M^{-1/2})$, whereas that in (a) is valid for all ξ except $\xi \approx 0$ (where (b) is valid) and $\xi \approx 1$. Thus the results obtained by the WKB approach are much more uniform than those from a diffusion approximation. They are also more accurate numerically. In Figure 1 we give graphs of the exact probabilities (p_n^{EXA}), the WKB approximation (p_n^{WKB}) and the diffusion approximation (p_n^{DIFF}) when $\rho = 3.0$ and $M = 20$ for $b(z) = \mu e^{-\mu z}$. In this case it is easy to obtain the exact answer and demonstrate the accuracy of the WKB approximation. Let us review the procedure we have used. We started from a complicated system ((2.13)-(2.19)) of differential-difference equations, with global transitions in the y variable, and linear (non-constant) coefficients in n. Then, using perturbation methods, we reduced the problem to solving a set of simpler equations. To obtain the WKB approximation we need to solve only the transcendental equation (2.26) and then a sequence of linear first order ODEs to get the functions ϕ_j for $j \geq 1$. To get the "boundary layer" approximation, valid for $n = O(1)$, it was necessary to solve the "constant coefficient version" of (2.13)-(2.17), which is easy to do using transforms. Using this approach we could also treat more complicated queues, e.g. ones with a general state dependent arrival rate $\tilde{\lambda}(n)$. For such problems transform methods are not applicable.

The details of the analysis presented here appear in [28]. These results were extended to the finite source $M/G/2$ queue in [33] and to problems with more general state-dependent parameters in [25], and a similar analysis was done for the $GI/M/m$ queue in [20].

2.2 State-dependent queues described by the unfinished work

Consider an $M/G/1$ model with an arrival rate that depends on the workload, i.e. $\lambda = \tilde{\lambda}(U(t))$. Also, we consider a state-dependent service density that is allowed to depend on the value of $U(t)$ when that particular customer entered the system, hence Pr[service time $\in (x, x+dx)|U(t^*) = w] = \tilde{b}(x,w)dx$ where t^* is the arrival time of the customer that requests x units of service. The forward equation for this state-dependent model is a generalization of the equation for the distribution of the unfinished work formulated by Takács (see equation (1.2) in [23]). We assume that the arrivals are fast and service times (jumps in $U(t)$) are small, so that we write $\tilde{\lambda}$, $\tilde{b}$ in the scaled forms

$$\tilde{\lambda}(w) = \frac{1}{\varepsilon}\lambda(w), \quad \tilde{b}(x,w) = \frac{1}{\varepsilon}b(\frac{x}{\varepsilon}, w)$$

where ε may be defined, say, as $(\tilde{\lambda}(0))^{-1}$. Denoting by $p(w,t)$ the unfinished work density for $w > 0$ and by $A(t)$ the probability that the system is empty, the forward equation is

$$p_t = p_w - \frac{1}{\varepsilon}\lambda(w)p + \frac{1}{\varepsilon}\int_0^{w/\varepsilon} p(w-\varepsilon z)\lambda(w-\varepsilon z)b(z, w-\varepsilon z)dz + \frac{A(t)\lambda(0)}{\varepsilon^2}b\Big(\frac{w}{\varepsilon}, 0\Big) \quad (2.38)$$

with the boundary condition

$$A'(t) = -\frac{\lambda(0)}{\varepsilon}A(t) + p(0,t).$$

Also, if $\tau(w)$ is the mean residual busy period (i.e. time to empty the system given $U(0) = w$), it satisfies the backward equation

$$-\tau_w(w) - \frac{1}{\varepsilon}\lambda(w)\tau(w) + \frac{1}{\varepsilon}\lambda(w)\int_0^\infty \tau(w+\varepsilon z)b(z,w)dz = -1 \quad (2.39)$$

with $\tau(0) = 0$. It is also of interest to compute the time needed for the workload $U(t)$ to exceed a certain value; call it K. Letting $n(w)$ be the mean time for $U(t)$ to exceed K given $U(0) = w$, it satisfies

$$-n_w(w) - \tfrac{1}{\varepsilon}\lambda(w)n(w) + \tfrac{1}{\varepsilon}\lambda(w)\textstyle\int_0^{(K-w)/\varepsilon} n(w+\varepsilon z)b(z,w)dz = -1, \quad 0 < w < K, \quad (2.40)$$
$$-\tfrac{1}{\varepsilon}\lambda(0)n(0) + \tfrac{1}{\varepsilon}\lambda(0)\textstyle\int_0^{K/\varepsilon} n(\varepsilon z)b(z,0)dz = -1$$

with $n(w) = 0$ for $w \geq K$.

We have analyzed the problems (2.38)-(2.40) using perturbation methods in [23]. Again the WKB method and asymptotic matching proved useful in analyzing this integro-differential equation, which cannot be solved exactly. Finite capacity queues, in which $U(t)$ is not allowed to exceed a given level, were analyzed in [24,26]. In [31,32] we analyzed Markov-modulated state-dependent queues, in [30] we obtained the distribution of the maximum value of $U(t)$ during a busy period, and in [29] we obtained the (time-dependent) busy period distribution for the model described here. State-dependent $GI/G/1$ queues were considered in [35].

We close this section by mentioning the work of Keller on time-dependent queues ([14]) that is similar in approach to that discussed here. Time-dependent queues also lead to non-constant coefficient equations, where L and L^* in (1.1)-(1.2) now explicitly involve time. State-dependent and time-dependent queues are qualitatively very different, but the analysis of each usually involves several different (space/time) scales, and asymptotic matching allows us to relate the various scales to one another.

3 Processor-shared Queues

Processor-shared (PS) queues are used to model time-sharing computer systems. In some respects their analysis is more difficult than that for FIFO queues. For PS queues, one wishes to compute the **sojourn time.** This is the time period from when a "tagged" customer enters the system, to when that customer leaves (after obtaining the required service).

Consider a Markovian, state-dependent PS model which has (i) a single server, (ii) a state-dependent arrival rate $= \lambda_n$ and (iii) a state-dependent service rate $= \mu_n$, where the rates are conditioned on $N(t) = n$. We denote the sojourn time by R and its conditional density by

$$p_n(t,x)dt = \Pr[R \in (t, t+dt)|N(0^-) = n, X = x]. \tag{3.1}$$

Here X is the tagged customer's required service, and n is the number of customers already in the system as the tagged customer arrives, which is assumed to occur at $t = 0$.

The density satisfies the evolution equation

$$\begin{aligned}\frac{\partial}{\partial t}p_n(t,x) &= -\frac{1}{n+1}\frac{\mu_{n+1}}{\mu_0}\frac{\partial}{\partial x}p_n(t,x) + \lambda_{n+1}[p_{n+1}(t,x) - p_n(t,x)] \\ &\quad + \frac{n}{n+1}\mu_{n+1}[p_{n-1}(t,x) - p_n(t,x)]; \quad n = 0, 1, \ldots\end{aligned} \tag{3.2}$$

for $x > 0$, and the initial condition is $p_n(t, 0) = \delta(t)$ $(n \geq 0)$, since a customer with zero service does not spend any time in the system according to the PS discipline. There are several important special cases of this model which we summarize below as models $A - F$:

model A : $\lambda_n = \lambda$; $\mu_n = \mu$

model B : $\lambda_n = \lambda(M - n)$; $\mu_n = \mu$

model C : $\lambda_n = \begin{cases} \lambda, n < K \\ 0, n \geq K \end{cases}$; $\mu_n = \mu$

model D : λ_n general; $\mu_n = \mu$

$$\text{model } E: \lambda_n = \lambda(M-n);\ \mu_n = \frac{\mu_0}{1+\omega n}$$

$$\text{model } F: \lambda_n = \begin{cases} \lambda, n < K \\ 0, n \geq K \end{cases};\ \mu_n = \frac{\mu_0}{1+\omega n}\ .$$

The exact solution to model A was obtained in [5] and asymptotic properties of the sojourn time were studied in [54]. Model B corresponds to a closed network consisting of a PS-server in series with a set of M terminals (IS node). Model C is a finite capacity model where at most K customers are allowed into the system (with additional arrivals lost). Model E again corresponds to the closed network of model B, with a state-dependent server. The parameter ω has been used to model the "switching time," which is the time needed for the server to switch between individual customers. Model F has a finite capacity and also considers the switching times.

Sometimes it may not be possible to compute the full conditional density $p_n(t,x)$. Hence we define

(a) $E[R^j|n,x] = \int_0^\infty t^j p_n(t,x)dt;\ j = 1,2,\ldots$

(b) $p(t,x) = \sum_n p_n(t,x)\pi_n$

(c) $E[R^j|x] = \int_0^\infty t^j p(t,x)dt;\ j = 1,2,\ldots$

(d) $p(t) = \int_0^\infty \mu e^{-\mu x} p(t,x)dx$

(e) $E[R^j] = \int_0^\infty t^j p(t)dt;\ j = 1,2,\ldots$

which denote respectively the conditional sojourn time moments; the distribution conditioned only on the service request; the moments conditioned only on the service request; the unconditional sojourn time distribution; and the unconditional moments. Here π_n is the (steady-state) probability that the tagged customer finds n others at the arrival instant. By multiplying (3.2) by t^j and integrating with respect to t over $[0,\infty)$, we can obtain recursive equations for the conditional moments, the first two have the form $\mathcal{L}_{n,x}E[R|n,x] = -1$, $\mathcal{L}_{n,x}E[R^2|n,x] = -2E[R|n,x]$ where the operator $\mathcal{L}$ is that in the right side of (3.2).

The various models $B-F$ were in recent years analyzed using singular perturbation methods. For model B, (d) was computed in [55], (c) was computed in [56] and (b) was computed in [57]. The asymptotic limit assumes that $M \to \infty$ with $\lambda M/\mu$ fixed, and the structure of the problem depends on whether $\lambda M/\mu < 1$ (normal usage), $\lambda M/\mu = 1 + O(M^{-1/2})$ (heavy usage), or $\lambda M/\mu > 1$ (very heavy usage). Some of these results have been extended to systems with multiple customer classes in [53,58,59]. For the finite capacity model C, (a) was computed in [16], (d)-(e) were computed in [22] and (b) was considered in [64]. The asymptotics assume that $K \to \infty$ and depend on whether $\lambda/\mu < 1$, $\lambda/\mu = 1 + O(K^{-1})$, $\lambda/\mu = 1 + O(K^{-1/2})$, or $\lambda/\mu > 1$. For model D, the case of a general (but smooth) λ_n was considered in [34] and (a), (c) were computed for a certain scaling of λ_n and μ. Note that model D contains model B as a special case. The closed model E with switching times was analyzed in [2], where the authors computed asymptotic approximations to (d). In [41,42] (c) was computed under various assumptions: in the asymptotic limit $M \to \infty$, $\mu_0 = O(M)$, $\omega = O(M^{-1})$. The size of the "switching time" parameter ω significantly affects the asymptotics, and this shows that the introduction of ω is an important consideration in modeling real systems. For model F, the conditional moments (c) were computed asymptotically in [63] for $K \to \infty$, $\omega = O(K^{-1})$, and several cases of the parameters (ρ, β) where $\rho = \lambda/\mu_0$, $\beta = K\omega$.

We also mention some work on the $GI/M/1 - PS$ model, which generalizes model A by allowing for a general (renewal) input. Approximations for (a) and (d) were obtained in [21] and (b) was analyzed in [65]. For the $GI/M/1 - PS$ model with finite capacity K (which generalizes model C), (c) was computed in [68] and (d) appears in [69]. For this model π_n is not known explicitly, but asymptotic approximations for $K \to \infty$ have been obtained in [70].

There are basically two asymptotic approaches that have been developed to analyze problems of the form (3.2). The first is to use a generating function to transform over n and then analyze the resulting equation, which may be an ODE, PDE or a functional equation, depending on the specific model and on the quantity that is to be computed. A second approach is to analyze (3.2) directly after some appropriate scalings. The first approach was used in [2,55-58] and the second was used in [16, 34, 41-42, 63-64]. Of course, where both are applicable they yield identical results, as is shown in [22].

In the next section we also discuss related work on sojourn times in networks, which may contain PS nodes.

4 Networks

Next we consider networks of queues. We will discuss (i) asymptotic expansions for the partition function for large product form networks, (ii) the time needed for large queue lengths to build up, (iii) bottleneck analysis, (iv) sojourn times in networks with overtaking, and (v) approximations for non-product form networks.

4.1 Partition functions

To illustrate the basic ideas and results, we consider a closed BCMP network which has a single class (chain) consisting of M customers, a single IS (infinite server) node and K single-server nodes with constant service rates. For this model the steady-state queue length distribution is well known to have the product form

$$p(n_1, \dots n_K) = \frac{1}{G(M,K)} \frac{M!}{(M - n_1 - n_2 - \cdots - n_K)!} \rho_1^{n_1} \dots \rho_K^{n_K} \tag{4.1}$$

where n_j is the number of customers in node j, ρ_j is the relative utilization and G is the partition function (normalizing constant)

$$G(M,K) = \sum_{n_1 + \cdots + n_K \leq M} \frac{M!}{(M - n_1 - n_2 - \cdots - n_K)!} \rho_1^{n_1} \dots \rho_K^{n_K}. \tag{4.2}$$

The performance measures, such as mean queue lengths and throughputs, can be easily calculated from G. The numerical evaluation of G is, however, difficult for M and/or K large. Various ideas have been introduced to simplify the calculation of G, such as computational algorithms (see [3,7,48]) and asymptotic expansions using integral representation (see [46,49,50,52]).

In [39] we developed an asymptotic approach to treat networks where **both** the population M and the number of nodes K are large. It has been shown in [3] that G can be computed from the recursion

$$\begin{gathered} h(m,k) = h(m,k-1) + m\rho_k h(m-1,k); \ 1 \leq m \leq M, 1 \leq k \leq K \\ h(0,k) = 1; \ 1 \leq k \leq K \\ h(m,0) = 1; \ 1 \leq m \leq M \end{gathered} \tag{4.3}$$

with $G(M,K) = h(M,K)$. We have obtained asymptotic expansions for G using the ray method and asymptotic matching, applied to the difference equation (4.3). The asymptotics depend on the relative sizes of the ρ_k. For example, if each of the ρ_k is $O(M^{-1})$ and the utilizations are not much different from one another, our final result for G is

$$G(M,K) \approx \frac{e^{M[U_0 - 1 - \log U_0]}}{\left\{\prod_{k=1}^{K}(1-U_k)\right\}[U_0 + U_0 \sum_{k=1}^{K} \frac{\rho_k}{(1-U_k)^2}]^{1/2}} \tag{4.4}$$

where $U_0 < 1$ is determined by

$$M = MU_0 + \sum_{k=1}^{K} \frac{U_k}{1-U_k}. \tag{4.5}$$

with $U_k = M\rho_k U_0 < 1$. Equation (4.5) is the fixed-population mean (FPM) approximation in which the closed network is replaced by an equivalent open network, but with the network population constrained to be M. The unknown constant U_0 in (4.5) represents the utilization in the IS node of this open network. We show in [39] that (4.4) - (4.5), as well as the results derived under other assumptions on ρ_k, are very accurate numerically. We have extended this analysis to multi-class networks with all single-server nodes [40], and to multi-class networks with IS node(s) [51].

4.2 Buildup of large queue lengths

The steady-state queue length distribution is easy to obtain for tandem Jackson networks consisting of d $M/M/1$ queues in series. However, it proves much harder to analyze the busy period or to compute the time until the network population becomes large. Recently, we have computed asymptotic approximations to the mean time until the total population reaches M for $M \to \infty$. In a stable network with $\rho_j = \lambda/\mu_j < 1 \ \ (1 \leq j \leq d)$, this is a rare event and the mean time grows exponentially with M. When $d = 2$ our final results take the form ([43])

$$\begin{aligned} T &\sim \frac{1}{\lambda}\frac{\rho_> - \rho_<}{(1-\rho_<)(1-\rho_>)^2}\rho_>^{-M};\ \rho_> > \rho_< \\ T &\sim \frac{1}{\lambda}\frac{1}{(1-\rho)^3}\frac{a}{\sinh(a/\rho)}\frac{1}{M}\rho^{-M};\ \rho_1 \approx \rho_2. \end{aligned} \tag{4.6}$$

Here $T = E[\tilde{T}|N_1(0) = N_2(0) = 0]$, $\tilde{T} = \min\{t : N_1(t) + N_2(t) = M\}$, $N_j(t) =$ queue length in node j, $\rho_> = \max\{\rho_1, \rho_2\}$, $\rho_< = \min\{\rho_1, \rho_2\}$, $\rho = (\rho_1 + \rho_2)/2$, and $a = (\rho_1 - \rho)M = (\rho - \rho_2)M$. The second formula in (4.6) applies to the limit $M \to \infty$ with $|\rho_2 - \rho_1| = O(M^{-1})$, and $a = 0$ corresponds to the case of identical nodes. The result (4.6) and the analogous formulae for $d > 2$ were obtained by applying singular perturbation methods to a d-dimensional recursion equation. Using such techniques, we have also computed the time until the network population is no longer large, given a large initial value in [44].

4.3 Bottleneck analysis

Mean value analysis (MVA) computes **mean** performance measures for product-form networks, without computing the partition function (see [60]). To illustrate the procedure, consider a single class, closed network with population size $= M$ and K single-server nodes. Let $N_i(m)$ be the mean queue length at node i for a network with population size $= m$. Then MVA corresponds to the nonlinear iteration

$$N_i(m) = \frac{mD_i[1 + N_i(m-1)]}{\sum_{k=1}^{K} D_k[1 + N_k(m-1)]};\ 1 \leq i \leq K,\ \ 1 \leq m \leq M \tag{4.7}$$

where D_i is the load at node i and the initial condition is $N_i(0) = 0$. A "bottleneck" node is one where D_i is largest. Since (4.7) may be difficult to compute if M is very large, approximations have been developed that assume a unique bottleneck ([61,62]). In particular, [62] develops a perturbation method which scales (4.7) and expands $N_i(m)$ in powers of $\varepsilon \equiv M^{-1}$. This approximation method also applies to multi-class networks. However, it works poorly when there are 2 or more bottleneck nodes. This is because the perturbation expansion becomes invalid if several of the D_i are nearly equal. Recently, we have shown how to treat such non-uniformities. For example, assume that nodes 1 and 2 are the bottlenecks with $D_1 \approx D_2 > D_j (j \geq 3)$. Then scaling $z = m/M = \varepsilon m$, $h_i(z) = N_i(m)/M$ $(i = 1, 2)$ leads, to leading order in ε, to the ODE

$$zDH_1'(z) + (D - 2az)H_1(z) + 2aH_1^2(z) = zD;\; H_1(0) = 0 \tag{4.8}$$

where $H_1(z)$ is the limit of $h_1(z)$ as $\varepsilon \to 0$, $D = (D_1 + D_2)/2$, $a = (D_1 - D)M = (D - D_2)M$, and $h_2(z) \sim H_2(z) = z - H_1(z)$ as $\varepsilon \to 0$. (4.8) is a Ricatti equation whose solution is

$$H_1(z) = -\frac{D}{2a} + \frac{z}{1 - e^{-2az/D}} \tag{4.9}$$

The approximation based on (4.9) leads to accurate numerical results, regardless of the relative sizes of D_1, D_2, and contains the single-node bottleneck approximations as limiting cases.

We have extended this approach to networks with many bottleneck nodes. There it is necessary to solve a system of non-linear ODEs of the type (4.8), but this can nevertheless be done explicitly. We can also treat multi-class networks near "switch-points," which correspond to regions where the bottleneck configuration changes ([45]).

4.4 Networks with overtaking

It is of interest to compute the sojourn time through a network of queues, as a customer goes through the network along a specified path. This can be easily computed for (paths in) product-form networks which are "overtake free." This means that along the specified path, customers cannot overtake others that are initially in front of them. Most networks are not overtake-free as service disciplines such as processor-sharing, and also network topologies, allow customers to overtake one another. The simplest examples of networks with overtaking were studied in [4,10], but no explicit expressions could be obtained for the sojourn time distribution. In [38] we considered a two-node network with an $M/M/1$ PS node in series with a FIFO node. Asymptotic approximations to the sojourn time moments were obtained, where it was assumed that as the tagged customer entered the PS node, the number of customers in at least one of the nodes was large. These results were shown to be in excellent agreement with simulations.

In [36,37] we considered networks with overtaking in the heavy traffic limit. In [37] we analyzed the sojourn time distribution for tandem PS-FIFO and PS-PS queues, assuming that $\rho_i = \lambda/\mu_i \uparrow 1 \quad (i = 1, 2)$. A similar analysis was done in [36] for a 3-node network where overtaking was caused by the network topology. In each case we derived three-term asymptotic approximations to the sojourn time density which gave simple quantitative measures of the effects of overtaking. Our approach involved using generating functions and then analyzing the resulting functional equations using singular perturbation ideas (e.g. scaling, matching).

4.5 Non-product form networks

The analysis of these problems is difficult even for two coupled (Markovian) queues. Classic examples of such problems are the shortest queue problem in [12], the fork-join model in [11], and two parallel queues where one server helps the other during periods when one of the queues is empty in [47]. Exact solutions are available ([6,9]) for some of these problems, but their analytic complexity makes them difficult to interpret.

Multi-dimensional Markovian networks correspond to solving problems of the type (1.1)-(1.2) with L, L^* being multi-dimensional difference operators with complicated boundary conditions. Diffusion approximations have been formulated for queueing networks ([13]), but their solution again involves analyzing (1.1)-(1.2). Now L is a partial-differential operator of second order, but the "oblique-derivative" boundary conditions make these problems difficult.

We have developed a singular perturbation approach for obtaining the tails of the distributions for multi-dimensional models. Such an approach can be applied either to difference equations ([27,38,44]) or to PDEs ([18,19]). For many of these problems there is no natural large or small parameter, so that it seems that the only way to obtain simple formulas is to assume that space and/or time is large. So far this approach has only been applied to two-dimensional models, but we believe that these ideas can also be used on higher-dimensional networks.

5 Transient Behavior of Queues

It is generally much harder to solve for the transient distribution of a stochastic model than it is to obtain the steady-state distribution. We show how to use perturbation methods to simplify the former task. As an example we consider the Erlang loss model (the $M/M/m/m$ queue), which is important in the analysis of blocking in teletraffic.

Let $N(t)$ be the number of occupied servers, each with rate μ, and let λ_0 be the arrival rate. Then $p_n(t) =$Pr$[N(t) = n]$ satisfies

$$p_n'(t) = \lambda_0 p_{n-1}(t) + \mu(n+1)p_n(t) - (\lambda_0 + \mu n)p_n(t);\ 1 \le n \le m-1 \tag{5.1}$$

$$p_0'(t) = \mu p_1(t) - \lambda_0 p_0(t) \tag{5.2}$$

$$p_m'(t) = \lambda_0 p_{m-1}(t) - \mu m p_m(t) \tag{5.3}$$

$$p_n(0) = \delta_{n,n_0};\ 0 \le n_0 \le m \tag{5.4}$$

where we assume that n_0 servers are occupied at $t = 0$. We assume that the number of servers m is large and that the arrival rate λ_0 is also large. Thus we scale $\lambda_0 = \lambda m$ with $\lambda = O(1)$ as $m \to \infty$. Letting $\xi = n/m$, $p_n(t) = P(\xi, t)$ and $\varepsilon = m^{-1}$, the scaled version of (5.1) is

$$\varepsilon P_t(\xi, t) = \lambda P(\xi - \varepsilon, t) + \mu(\xi + \varepsilon)P(\xi + \varepsilon, t) - (\lambda + \mu\xi)P(\xi, t). \tag{5.5}$$

As shown in [17,66], the asymptotic analysis depends on the size of $\rho = \lambda/\mu = \lambda_0/(\mu m)$, and we must consider

(i) $\rho < 1$ (ii) $\rho > 1$ (iii) $\rho - 1 = O(m^{-1/2})$.

Also, the problem is very sensitive to the initial condition n_0 and we must consider the three cases

(a) $n_0 = O(1)$ (b) $0 < \xi_0 = n_0/m < 1$ (c) $m - n_0 = O(1)$

which correspond respectively to starting the process with only a few occupied servers, a fixed fraction of the servers occupied, and with almost all the servers occupied. Thus there are nine cases that must be analyzed. In addition, within a particular case of (ρ, n_0), it is necessary to analyze several regions in the (n, t) (or (ξ, t)) plane. All these scales are treated in [17] if $\rho < 1$ and in [66] if $\rho > 1$ or $\rho \approx 1$. Here we give a few of the main results.

Assume first that $\rho < 1$ and $0 < \xi_0 < 1$. We first analyze (5.5) for short times $t = \varepsilon\tau = O(\varepsilon)$ and localize space near the initial condition by setting $k = (\xi - \xi_0)/\varepsilon = n - n_0$. To leading order we obtain

$$p_n(t) \sim e^{-(\lambda+\mu\xi_0)\tau}\left(\frac{\lambda}{\mu\xi_0}\right)^{k/2} I_k(2\sqrt{\lambda\mu\xi_0\tau});\ \tau = \frac{t}{\varepsilon},\quad k = \frac{\xi - \xi_0}{\varepsilon} \tag{5.6}$$

where $I_k(\cdot)$ is the modified Bessel function. Relation (5.6) corresponds to a free space birth-death process with birth rate λ and death rate $\mu\xi_0$.

For $t > 0$ and $0 < \xi < 1$, we obtain the approximation to $p_n(t)$ using the **ray method**. We set

$$p_n(t) = P(\xi, t) = C(\varepsilon) e^{-\phi(\xi,t)/\varepsilon}[A(\xi, t) + \varepsilon A_1(\xi, t) + \cdots] \tag{5.7}$$

in (5.5) and find that ϕ, A satisfy the PDEs

$$-\phi_t = \lambda(e^{\phi_\xi} - 1) + \mu\xi(e^{-\phi_\xi} - 1) \tag{5.8}$$

$$A_t + (\lambda e^{\phi_\xi} - \mu\xi e^{-\phi_\xi})A_\xi = [\mu e^{-\phi_\xi} - \frac{1}{2}\phi_{\xi\xi}(\lambda e^{\phi_\xi} + \mu\xi e^{-\phi_\xi})]A. \tag{5.9}$$

These are analogous to the "eiconal" and "transport" equations of geometrical optics. They can be solved using standard methods for PDEs. Both are first order equations, though (5.8) is nonlinear. To specify uniquely the functions ϕ and A, we must match the ray approximation to the short time approximation (5.6). This leads to the final results

$$\begin{aligned}
\phi(\xi, t) &= \xi \log z_0 - \rho(z_0 - 1)e^{-\mu t} - \xi_0 \log(1 - e^{-\mu t} + z_0 e^{-\mu t});\\
z_0(\xi, t) &= \frac{1}{2}\left\{\frac{\xi - \xi_0}{\rho(1 - e^{-\mu t})} + 1 - e^{\mu t} + \sqrt{\left(\frac{\xi_0 - \xi}{\rho(1 - e^{-\mu t})} + e^{\mu t} - 1\right)^2 + \frac{4\xi e^{\mu t}}{\rho}}\right\};\\
A(\xi, t) &= \frac{1}{\sqrt{2\pi}}\left[\xi - \xi_0\Big(\frac{z_0}{e^{\mu t} - 1 + z_0}\Big)^2\right]^{-1/2}
\end{aligned} \tag{5.10}$$

and $C(\varepsilon) = \sqrt{\varepsilon}$. This approximation becomes invalid near the boundaries $\xi = 0, 1$. There other expansions must be constructed (see [17,66]). The latter boundary region is especially important if one wants an accurate approximation to the blocking probability $p_m(t)$. The leading order approximation to this is

$$p_m(t) \sim \sqrt{\varepsilon} A(1,t) e^{-\phi(1,t)/\varepsilon} \left[1 + \frac{1 - \rho z_0(1,t)}{1 - z_0(1,t)} \frac{1}{\rho z_0(1,t)} \right]. \tag{5.11}$$

Note that this is different than setting $\xi = 1 (n = m)$ in (5.7), and again indicates the importance of treating the various scales inherent to the problem.

Now we examine the steady-state limit of the ray approximation by letting $t \to \infty$. From (5.10) we have $A(\xi, \infty) = (2\pi\xi)^{-1/2}$ and $\phi(\xi, \infty) = \xi \log \xi + \rho - \xi - \xi \log \rho$. The exact stationary distribution is

$$p_n(\infty) = \frac{(\rho m)^n}{n!} p_0(\infty), \; p_0(\infty) = 1 \Big/ \sum_{\ell=0}^{m} \frac{(\rho m)^\ell}{\ell!}. \tag{5.12}$$

If $\rho < 1$, we obtain $p_0(\infty) \sim e^{-\rho m}$ as $m \to \infty$. Then expanding $n!$ by Stirling's formula, we see that $p_n(\infty) \sim (2\pi n)^{-1/2} (\rho m / n)^n e^n e^{-\rho m}$, which is the same as the limit of the ray expansion as $t \to \infty$ (recall that $m = \varepsilon^{-1}$, $\xi = n/m = \varepsilon n$).

When $\rho < 1$, the blocking probability is exponentially small for all times t (cf. (5.11)). Now we consider more heavily loaded systems which have $\rho > 1$ (again taking $0 < \xi_0 < 1$). For this case the main result(s) for $p_n(t)$ in [66] are (away from boundary and initial layers)

$$\begin{aligned} p_n(t) \;&\sim\; \sqrt{\varepsilon} A(\xi, t) e^{-\phi(\xi,t)/\varepsilon}, \; t < T(\xi) \\ &\sim\; \left(1 - \rho^{-1}\right) \xi^{-1/2} e^{-\psi(\xi)/\varepsilon}, \; t > T(\xi) \end{aligned} \tag{5.13}$$

where $\psi(\xi) = 1 + \log \rho + \xi \log \xi - \xi - \xi \log \rho$ and ϕ, A are as in (5.10). The second formula in (5.13) is precisely the expansion of $p_n(\infty)$, as when $\rho > 1$ we have $p_0(\infty) \sim (1 - \rho^{-1}) m! (\rho m)^{-m}$. In (5.13), $T(\xi)$ is defined implicitly be the relation

$$\psi(\xi) = \phi(\xi, T(\xi)) \tag{5.14}$$

which defines a curve in the (ξ, t) plane which passes through the point $\xi = 1$, $\mu t = \log[(\rho - \xi_0)/(\rho - 1)]$. The curve $t = T(\xi)$ may be viewed as a "front" above which ($t > T(\xi)$) the process forgets the initial condition and settles to its steady-state behavior. Below the front ($t < T(\xi)$) transient effects are still important.

Thus, our analysis of $\rho < 1$ and $\rho > 1$ reveals two mechanisms by which a process approaches its steady state behavior. If $\rho < 1$, the ray approximation depended on time for all $t < \infty$, and approached the steady state smoothly as $t \to \infty$. When $\rho > 1$ there was a sharp transition to the equilibrium distribution at $t = T(\xi)$. The case $\rho \approx 1$ is more complicated and is treated in detail in [66].

The blocking probability $p_m(t)$ now ($\rho > 1, 0 < \xi_0 < 1$) has the asymptotic expansions

(a) $\mu t < \log\left(\frac{\rho - \xi_0}{\rho - 1}\right) = \mu T(1)$

$$p_m(t) \sim \sqrt{\varepsilon} A(1,t) e^{-\phi(1,t)/\varepsilon} \left[1 + \frac{1 - \rho z_0(1,t)}{1 - z_0(1,t)} \frac{1}{\rho z_0(1,t)}\right]; \tag{5.15}$$

(b) $\mu t - \log\left(\frac{\rho - \xi_0}{\rho - 1}\right) = \frac{\sqrt{\varepsilon}}{\rho - 1}\Delta = O(\sqrt{\varepsilon}\,)$

$$p_m(t) \sim \left(1 - \frac{1}{\rho}\right) \frac{1}{\sqrt{2\pi}} \int_{-\Delta/\sqrt{\beta}}^{\infty} e^{-u^2/2} du, \; \beta = 1 - \frac{\xi_0(\rho - 1)^2}{(\rho - \xi_0)^2};$$

(c) $\mu t > \log\left(\frac{\rho - \xi_0}{\rho - 1}\right)$

$$p_m(t) \sim p_m(\infty) \sim \left(1 - \frac{1}{\rho}\right).$$

This shows that for times $t < T(1)$ the blocking probability is exponentially small, and settles to its equilibrium value as t passes through the critical time $T(1)$. Note also that if $\rho > 1$, $z_0(1, T(1)) = 1$, so that (5.15) becomes infinite (and is thus invalid) as $t \uparrow T(1)$.

We believe that the type of structure discussed here (e.g. the front at $t = T(\xi)$) is canonical to many applied probability models. It has also been observed in the finite capacity $M/M/1$ queue and in the infinite capacity $M/M/1$ queue, if we take the initial condition $N(0) = n_0$ to be large ([67]). These ideas should also apply to models in more than one space dimension; there the PDEs (5.8)-(5.9) will involve time t and, say, 2 space variable ξ_1, ξ_2. They will still be of the general form (5.8)-(5.9) and can be readily solved.

Finally, we note that in [17,66] detailed numerical comparisons are presented, which show that the various asymptotic approximations are in very good agreement with the exact (numerical) values. This is true even for modest size systems which have, say, $m = 10$. Since this type of asymptotics reveals much about the qualitative structure of the models, our results achieve both of the goals outlined in section 1.

6 Open Problems and Research Directions

We have indicated how to use singular perturbation methods to analyze a variety of queueing models. We have shown that such methods may be used to obtain asymptotic information for models with general interarrival/service time distributions, state-dependent and time-dependent queues, large product-form networks, non-product form networks, and analyzing time-dependent behavior. Except for the work on large product-form networks, most of our work has involved computing steady-state behavior for models in one or two dimensions, and transient behavior for one-dimensional models. We expect that these techniques may also be used on models in three or more dimensions. For non-product form networks, very little is known for models in more than two dimensions, even under Markovian assumptions on the arrival and service processes. We believe that the "geometric optics" approach outlined in [18,19,66,67] should be useful for computing asymptotically, steady-state probabilities in more than two dimensions, and also transient probabilities in more than one dimension. It should be possible to treat multi-dimensional non-Markovian models using these methods. We are presently investigating such possibilities.

We have shown that the methods discussed are general enough to be used on many specific problems in queueing theory. At the same time, our results show that they capture the subtle structure of the individual model and thus lead to better understanding of the differences between models. A good applied mathematics technique should be applicable to a large class of problems, yet it should also be able to lead to an in-depth understanding of specific problems. In this article we hope that we have demonstrated that perturbation methods satisfy both of these essential criteria.

REFERENCES

[1] BENDER, C.M. AND ORSZAG, S.A., *Advanced Mathematical Methods for Scientists and Engineers*, McGraw-Hill, New York, 1978.

[2] BERSANI, A. AND SCIARRETTA, C., Asymptotic analysis for a closed processor-sharing system with switching times, *SIAM J. Appl. Math.* **51**, 525-541, 1991.

[3] BUZEN, J.P., Computational algorithms for closed queueing networks with exponential servers, *Commun. ACM* **16**, 527-531, 1973.

[4] COFFMAN, JR., E.G., FAYOLLE, G. AND MITRANI, I., Sojourn times in a tandem queue with overtaking: reduction to a boundary value problem, *Comm. Statist.-Stochastic Models* **2**, 43-65, 1986.

[5] COFFMAN, JR., E.G., MUNTZ, R.R., AND TROTTER, H., Waiting time distributions for processor-sharing systems, *J. ACM* **17**, 123-130, 1970.

[6] COHEN, W. AND BOXMA, O.J., *Boundary Value Problems in Queueing Systems Analysis*, North-Holland, Amsterdam, 1983.

[7] CONWAY, A.E. AND GEORGANAS, N.D., RECAL - A new efficient algorithm for the exact analysis of multiple-chain closed queueing networks, *J. ACM* **33**, 768-791, 1986.

[8] COX, D.R., The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables, *Proc. Camb. Phil. Soc.* (Math. and Phys. Sci.) **51**, 433-441, 1955.

[9] FAYOLLE, G., On functional equations for one or two complex variables arising in the analysis of stochastic models, in *Mathematical Computer Performance and Reliability*, G. Iazeolla, P.J. Curtois, and A. Hordijk, eds., North-Holland, Amsterdam, 55-76, 1984.

[10] FAYOLLE, G., IASNOGORODSKI, R. AND MITRANI, I., The distribution of sojourn times in a queueing network with overtaking: reduction to a boundary value problem, in *Performance 1983*, A.K. Agrawala and S.K. Tripathi, eds., North-Holland, Amsterdam, 477-486, 1983.

[11] FLATTO, L. AND HAHN, S., Two parallel queues created by arrivals with two demands I, *SIAM J. Appl. Math.* **44**, 1041-1054, 1984.

[12] FLATTO, L. AND MCKEAN, H., Two queues in parallel, *Comm. Pure and Appl. Math.* **30**, 255-263, 1977.

[13] HARRISON, M.J. AND REIMAN, M.I., On the distribution of multi-dimensional reflected Brownian motion, *SIAM J. Appl. Math.* **41**, 345-361, 1981.

[14] KELLER, J. B., Time-dependent queues, *SIAM Rev.* **24**, 401-412, 1982.

[15] KEVORKIAN, J. AND COLE, J.D., *Perturbation Methods in Applied Mathematics*, Springer-Verlag, Berlin, New York, 1981.

[16] KNESSL, C., On finite capacity processor shared queues, *SIAM J. Appl. Math.* **50**, 264-287, 1990.

[17] KNESSL, C., On the transient behavior of the $M/M/m/m$ loss model, *Comm. Statist.-Stochastic Models* **6**, 749-776, 1990.

[18] KNESSL, C., Diffusion approximation to a fork and join queueing model, *SIAM J. Appl. Math.* **51**, 160-171, 1991.

[19] KNESSL, C., Diffusion approximation to two parallel queues with processor sharing, *IEEE Trans. on Automatic Control* **36**, 1356-1367, 1991.

[20] KNESSL, C. The WKB approximation to the $G/M/m$ queue, *SIAM J. Appl. Math.* **51**, 1119-1133, 1991.

[21] KNESSL, C., Asymptotic approximations for the $GI/M/1$ queue with processor sharing service, *Comm. Statist.-Stochastic Models* **8**, 1-34, 1992.

[22] KNESSL, C., On the sojourn time distribution in a finite capacity processor shared queue, *J. ACM* **40**, 1238-1301, 1993.

[23] KNESSL, C., MATKOWSKY, B.J., SCHUSS, Z. AND TIER, C., Asymptotic analysis of a state-dependent $M/G/1$ queueing system, *SIAM J. Appl. Math.* **46**, 483-505, 1986.

[24] KNESSL, C., MATKOWSKY, B.J., SCHUSS, Z. AND TIER, C., A finite capacity single-server queue with customer loss, *Comm. Statist.-Stochastic Models* **2**, 97-121, 1986.

[25] KNESSL, C., MATKOWSKY, B.J., SCHUSS, Z. AND TIER, C., On the performance of state-dependent single-server queues, *SIAM J. Appl. Math.* **46**, 657-697, 1986.

[26] KNESSL, C., MATKOWSKY, B.J., SCHUSS, Z. AND TIER, C., System crash in a finite capacity $M/G/1$ queue, *Comm. Statist.-Stochastic Models* **2**, 171-201, 1986.

[27] KNESSL, C., MATKOWSKY, B.J., SCHUSS, Z. AND TIER, C., Two parallel queues with dynamic routing, *IEEE Trans. Comm.* **34**, 1170-1175, 1986.

[28] KNESSL, C., MATKOWSKY, B.J., SCHUSS, Z. AND TIER, C., Asymptotic expansions for a closed multiple access system, *SIAM J. Comp.* **16**, 378-398, 1987.

[29] KNESSL, C., MATKOWSKY, B.J., SCHUSS, Z. AND TIER, C., Busy period distribution in state-dependent queues, *Queueing Systems: Theory and Applications* **2**, 285-305, 1987.

[30] KNESSL, C., MATKOWSKY, B.J., SCHUSS, Z. AND TIER, C., Distribution of the maximum buffer content during a busy period for state-dependent $M/G/1$ queues, *Comm. Statist.-Stochastic Models* **3**, 191-226, 1987.

[31] KNESSL, C., MATKOWSKY, B.J., SCHUSS, Z. AND TIER, C., A Markov-modulated $M/G/1$ queue I: Stationary distribution, *Queueing Systems: Theory and Applications* **1**, 355-374, 1987.

[32] KNESSL, C., MATKOWSKY, B.J., SCHUSS, Z. AND TIER, C., A Markov-modulated $M/G/1$ queue II: Busy period and time for buffer overflow, *Queueing Systems: Theory and Applications* **1**, 375-397, 1987.

[33] KNESSL, C., MATKOWSKY, B.J., SCHUSS, Z. AND TIER, C., The two repairmen problem: a finite source $M/G/2$ queue, *SIAM J. Appl. Math.* **47**, 367-397, 1987.

[34] KNESSL, C., MATKOWSKY, B.J., SCHUSS, Z. AND TIER, C., Response times in processor-shared queues with state-dependent arrival rates, *Comm. Statist.-Stochastic Models* **5**, 83-113, 1989.

[35] KNESSL, C., MATKOWSKY, B.J., SCHUSS, Z. AND TIER, C., A state-dependent $GI/G/1$ queue, *European Journal of Applied Math.* **5**, 217-241, 1994.

[36] KNESSL, C. AND MORRISON, J.A., Heavy traffic analysis of the sojourn time in a three node Jackson network with overtaking, *Queueing Systems: Theory and Applications* **8**, 165-182, 1991.

[37] KNESSL, C. AND MORRISON, J.A., Heavy traffic analysis of the sojourn time in tandem queues with overtaking, *SIAM J. Appl. Math.* **51**, 1740-1763, 1991.

[38] KNESSL, C. AND TIER, C., Approximations to the moments of the sojourn time in a tandem queue with overtaking, *Comm. Statist.-Stochastic Models* **6**, 499-524, 1990.

[39] KNESSL, C. AND TIER, C., Asymptotic expansions for large closed queueing networks, *J. ACM* **37**, 144-174, 1990.

[40] KNESSL, C. AND TIER, C., Asymptotic expansions for large closed queueing networks with multiple job classes, *IEEE Trans. on Computers* **41**, 480-488, 1992.

[41] KNESSL, C. AND TIER, C., A processor shared queue which models switching times: normal usage, *SIAM J. Appl. Math.* **52**, 883-899, 1992.

[42] KNESSL, C. AND TIER, C., A processor shared queue which models switching times: heavy usage asymptotics, *SIAM J. Appl. Math.* **54**, 854-875, 1994.

[43] KNESSL, C. AND TIER, C., Asymptotic properties of first passage times for tandem Jackson networks I: buildup of large queue lengths, *Comm. Statist.-Stochastic Models*, *to appear*.

[44] KNESSL, C. AND TIER, C., Asymptotic properties of first passage times for tandem Jackson networks II: time to empty the system, *preprint*.

[45] KNESSL, C. AND TIER, C., Asymptotic bottleneck analysis in single and multi-class networks, *in preparation*.

[46] KOGAN, Y., Another approach to asymptotic expansions for large closed queueing networks, *Oper. Res. Lett.* **11**, 317-321, 1992.

[47] KONHEIM, A.G., MEILIJSON, I. AND MELKMAN, A., Processor-sharing of two parallel lines, *J. Appl. Probab.* **18**, 952-956, 1981.

[48] McKenna, J., Extensions and applications of RECAL in the solution of closed product-form queueing networks, *Comm. Statist.-Stochastic Models* **4**, 235-276, 1988.

[49] McKenna, J. and Mitra, D., Integral representations and asymptotic expansions for closed Markovian queueing networks: normal usage, *Bell Syst. Tech. J.* **61**, 661-683, 1982.

[50] McKenna, J. and Mitra, D., Asymptotic expansions and integral representations of moments of queue lengths in closed Markovian networks, *J. ACM* **31**, 346-360, 1984.

[51] Mei, J.D. and Tier, C., Asymptotic approximations for a queueing network with multiple classes, *SIAM J. Appl. Math.* **54**, 1147-1180, 1994.

[52] Mitra, D. and McKenna, J., Asymptotic expansions for closed Markovian networks with state-dependent service rates, *J. ACM* **33**, 568-592, 1986.

[53] Mitra, D. and Morrison, J.A., Asymptotic expansions of moments of the waiting time in closed and open processor-sharing systems with multiple job classes, *Adv. Appl. Probab.* **15**, 813-839, 1983.

[54] Morrison, J.A., Response-time distribution for a processor-sharing system, *SIAM J. Appl. Math.* **45**, 152-167, 1985.

[55] Morrison, J. A., Asymptotic analysis of the waiting-time distribution for a large closed processor-sharing system, *SIAM J. Appl. Math.* **46**, 140-170, 1986.

[56] Morrison, J. A., Moments of the conditioned waiting time in a large closed processor-sharing system, *Comm. Statist.-Stochastic Models* **2**, 293-321, 1986.

[57] Morrison, J. A., Conditioned response-time distribution for a large closed processor-sharing system in very heavy usage, *SIAM J. Appl. Math.* **47**, 1117-1129, 1987.

[58] Morrison, J. A., Conditioned response-time distribution for a large closed processor-sharing system with multiple classes in very heavy usage, *SIAM J. Appl. Math.* **48**, 1493-1509, 1988.

[59] Morrison, J. A. and Mitra, D., Heavy-usage asymptotic expansions for the waiting time in closed processor-sharing systems with multiple classes, *Adv. Appl. Probab.* **17**, 163-185, 1985.

[60] Reiser, M. and Lavenberg, S.S., Mean-value analysis of closed multi-chain queueing networks, *J. ACM* **27**, 313-322, 1980.

[61] Schweitzer, P., A fixed-point approximation to product-form networks with large population, presented at Second ORSA Telecommunications Conference, Boca Raton, Florida, March 1992.

[62] Schweitzer, P., Serazzi, G. and Broglia, M., A survey of bottleneck analysis in closed networks of queues, in *Performance Evaluation of Computer and Communications Systems*, L. Donatiello and R. Nelson, eds., Lecture Notes in Computer Science, No. 729, Springer-Verlag, Berlin, 491-508, 1993.

[63] Tan, X. and Knessl, C., A finite capacity PS queue which models switching times, *SIAM J. Appl. Math.* **53**, 491-554, 1993.

[64] Tan, X. and Knessl, C., Sojourn time distribution in some processor shared queues, *European Journal of Applied Math.* **4**, 437-448, 1993.

[65] TAN, X., YANG, Y. AND KNESSL, C., The conditional sojourn time distribution in the $GI/M/1$ processor sharing queue in heavy traffic, *Queueing Systems: Theory and Applications* **14**, 99-109, 1993.

[66] XIE, S. AND KNESSL, C., On the transient behavior of the Erlang loss model: heavy usage asymptotics, *SIAM J. Applied Math.* **53**, 555-559, 1993.

[67] XIE, S. AND KNESSL, C., On the transient behavior of the $M/M/1$ and $M/M/1-K$ queues, *Studies in Applied Math.* **88**, 191-240, 1993.

[68] YANG, Y. AND KNESSL, C., Conditional sojourn time moments in the finite capacity $GI/M/1$ queue with processor sharing service, *SIAM J. Applied Math.* **53**, 1132-1193, 1993.

[69] YANG, Y. AND KNESSL, C., The unconditional sojourn time distribution in the $GI/M/1-K$ queue with processor sharing service, *Studies in Applied Math.* **93**, 1994, 29-91.

[70] YANG, Y. AND KNESSL, C., Heavy traffic asymptotics of the queue length in the $GI/M/1-K$ queue, *preprint*.