SERIE RESEARCH MEMORANDA

NUMERICAL METHODS FOR QUEUEING MODELS

H.C. Tijms

Research Memorandum 1988-43 Sept. '88

VRIJE UNIVERSITEIT
Faculteit der Economische Wetenschappen en Econometrie
AMSTERDAM

NUMERICAL METHODS FOR QUEUEING MODELS

H.C.Tijms
Institute for Econometrics
Vrije Universiteit
Amsterdam, The Netherlands

This paper presents a survey of computational methods for basic queueing models with an isolated servicing node. We first discuss exact and approximate methods for the computation of the state and waiting-time probabilities in the single-server queue with general interarrival and service times. The approximate methods for complex queueing models are typically based on exact results for simpler related models and on asymptotic expansions. Finally, we give approximations for multi-server queues with Poisson arrivals.

1. Introduction

A remarkable change in queueing analysis took place about 1970. In the fifties and the sixties, the precomputer area, the emphasis was naturally on obtaining analytical results in a closed form. Recently queueing analysis has been much more directed to the goal of getting algorithms suited for (recursive) calculations on the computer. For example, the old and relatively simple 'birth-and-death' technique has seen a revival by the use of phase-type distributions and the ability of current computers to solve very large systems of linear equations. This basic technique goes at least back to Fry [13] and Molina [17] and was a powerful computational tool already in the early days of teletraffic analysis. Most of the new and important developments that took place in the seventies are nicely overviewed in the supplemented chapter of the second edition of Cohen's book 'The Single Server Queue' [5].

My own interest in algorithmic methods for queueing models was much stimulated by pleasant and useful discussions I had with Wim Cohen. It is therefore my great pleasure to give in this Liber Amicorum a survey on numerical methods for queueing models. In what follows I hope to make clear that algorithmic analysis of queueing models is more than getting numerical answers. The essence of computational probability is to find probabilistic ideas which make the computations transparent and natural.

In this survey I have made the choice to restrict my overview to basic queueing models with an isolated servicing node and not to discuss queueing networks. Even within this restriction I have to be selective. The paper is organised as follows. Section 2 discusses several basic algorithms for the computation of the state probabilities and the waiting-time probabilities for queueing models of the M/G/1 type. In section 3 generally applicable algorithms and approximations for the GI/G/1 queue are given. The final section 4 discusses approximations for multi-server queues.

2. Computational methods for M/G/1-type queueing models

In this section we discuss basic algorithms for the computation of the state probabilities and waiting-time probabilities.

2.1 The regenerative method for the state probabilities

The regenerative approach is a powerful computational tool that is generally applicable for the calculation of the state probabilities in a variety of queueing models of the M/G/1-type. This approach uses basic results from the theory of regenerative processes and a simple up- and downcrossings argument.

The regenerative approach is best explained via the standard M/G/1 queue having a Poisson arrival process with rate λ and a general probability distribution function $B(t) = P\{S \leq t\}$ for the service time S of a customer. It is assumed that the offered load $\rho=\lambda E(S)<1$. The equilibrium state probabilities p_j for $j=0,1,\ldots$ are defined by

$$p_j = \lim_{t\to\infty} P\{\text{there will be j customers present a time t from now}\}$$

This limiting distribution exists and is independent of the initial state. Alternatively, the probability p_j can be interpreted as the long-run fraction of time the system will be in state j. Here state j corresponds to the situation that j customers are present. By the theory of regenerative processes (see e.g. Cohen [4]), we have the basic result

$$p_j = \frac{E(T_j)}{E(T)} \qquad \text{for } j=0,1,\dots\ , \tag{2.1}$$

where

T = the length of a cycle which is defined as the time elapsed between two arrivals finding the system empty;

T_j = the amount of time that the process is in state j during one cycle.

Note that the epochs at which an arrival occurs and finds the system empty are regeneration epochs for the queue length process. The regenerative method is based on two fundamental relations between the expected values $E(T_j)$ and $E(N_k)$, where the random variable N_k is defined by

N_k = the number of service completion epochs in one cycle at which k customers are left behind, $k=0,1,\dots$.

The first relation is based on the up- and downcrossings argument that during each cycle the number of transitions out of the set of states $\{k+1,k+2,\dots\}$ must be equal to the number of transitions into the set of states $\{k+1,k+2,\dots\}$ for each k. The expected number of transitions out of the set $\{k+1,k+2,\dots\}$ per cycle equals $E(N_k)$ by definition, while the expected number of transitions into the set $\{k+1,k+2,\dots\}$ per cycle equals $\lambda E(T_k)$. For the latter result we use that the arrival process is a Poisson process and invoke a fundamental property of the Poisson process given in Wolff [30]. Thus we obtain the relation

$$\lambda E(T_k) = E(N_k) \qquad \text{for } k=0,1,\dots\ . \tag{2.2}$$

The second relation between $E(T_j)$ and $E(N_k)$ is obtained by splitting up the cycle in disjoint intervals via the service completion epochs and calculating $E(T_j)$ as the sum of the contributions of the disjoint intervals. Here we need again the memoryless property of the Poisson arrival process. Then, noting that during a service the queue size cannot decrease, we obtain as second relation

$$E(T_j) = A_{1j} + \sum_{k=1}^{j} E(N_k)A_{kj} \qquad \text{for } j=1,2,\ldots, \tag{2.3}$$

where the quantity A_{kj} is defined by

A_{kj} = the expected amount of time that j customers are present during one service time at whose beginning k customers are present.

Note that $E(T_0) = 1/\lambda$ and so, by (2.1), $E(T) = 1/(\lambda p_0)$. Inserting (2.2) into (2.3), dividing both sides of the resulting equation by E(T) and using (2.1), we obtain the recursion equation

$$p_j = \lambda A_{1j} p_0 + \sum_{k=1}^{j} \lambda A_{kj} p_k \qquad \text{for } j=1,2,\ldots . \tag{2.4}$$

Starting with p_0 $(=1-\rho)$, we can recursively compute $p_1, p_2, \ldots$. This recursive algorithm is numerically stable and is easy to apply. It remains to specify the constants A_{kj}. In general, we have the representation

$$A_{kj} = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^{j-k}}{(j-k)!} \{1-B(t)\}dt.$$

These integrals can efficiently be computed by numerical integration methods such as Gauss-Laguerre or Gauss-Legendre, cf. Van Hoorn [27]. For phase-type service time distributions the constants A_{kj} can also be computed from a simple recursion scheme, cf. also (2.8)-(2.9) below. For computational purposes it is important to note that A_{kj} depends on k and j only through j-k.

The regenerative approach having its origin in the paper Hordijk and Tijms [14] is a flexible and generally applicable method to single-server queueing models in which the arrival process is Markovian. For example, the analysis can be straightforwardly extended to models with state-dependent Poisson arrivals (including the machine repairman model) and to variants of the M/G/1 queue such as models with server vacations, retrials or exceptional first services, see Tijms [25].

As further illustration, consider the $M^X/G/1$ queue with batch arrivals. Let $\{\beta_j, j\geq 1\}$ denote the batch size distribution. It is assumed that $\lambda E(B)E(S) < 1$, where the generic variable B denotes the batch size. The above analysis leading to the recursion scheme (2.4) needs only a modification with respect to the up- and downcrossing relation (2.2). This relation becomes now

$$\lambda \sum_{i=0}^{k} E(T_i) \sum_{m>k-i} \beta_m = E(N_k), \qquad k=1,2,\ldots .$$

The recursion scheme for the state probabilities p_j is now given by

$$p_j = \lambda p_0 \sum_{m=1}^{j} \beta_m A_{mj} + \lambda \sum_{k=1}^{j} \Big(\sum_{i=0}^{k} p_i \sum_{m>k-i} \beta_m \Big) A_{kj}, \qquad j=1,2,\ldots , \tag{2.5}$$

where

$$p_0 = 1-\lambda E(B)E(S). \tag{2.6}$$

In general the computation of the constants A_{kj} is much more tricky for the $M^X/G/1$ queue than for the M/G/1 queue. However, an efficient recursion scheme for the constants A_{kj} can be given for the important case that the service time has a probability density of the form

$$b(t) = \sum_{i=1}^{r} q_i \mu^i \frac{t^{i-1}}{(i-1)!} e^{-\mu t}, \qquad t\geq 0, \tag{2.7}$$

where $q_i \geq 0$ and $\Sigma_i q_i = 1$. That is, the service time distribution is a mixture of Erlangian distributions with the *same* scale parameters. Any probability distribution function with mass on $(0,\infty)$ can be arbitrarily closely approximated by a mixture of Erlangian distribution functions with the same scale parameters. This result due to Schassberger [20] is an extremely useful result in queueing analysis both for analytical and computational purposes. As a consequence of this result it is sufficient in many practical applications to consider service times that can be represented as a random sum of independent *exponentials* with the same means. For a service time with density (2.7) we have the useful interpretation that, with probability q_i, the service time is distributed as the sum of i independent exponential phases each with the same mean $1/\mu$. This interpretation together with the memoryless property of the exponential distribution enables us to compute the constants A_{kj} from

$$A_{kj} = \sum_{i=1}^{m} q_i \ \alpha_{j-k}^{(i)} \tag{2.8}$$

where the numbers $\alpha_{j-k}^{(i)}$ are recursively computed from

$$\alpha_0^{(i)} = \frac{1}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}\alpha_0^{(i-1)}, \qquad 1 \le i \le r$$

$$\alpha_n^{(i)} = \frac{\lambda}{\lambda+\mu}\sum_{j=1}^{n}\beta_j\ \alpha_{n-j}^{(i)} + \frac{\mu}{\lambda+\mu}\alpha_n^{(i-1)}, \quad n \ge 1,\ 1 \le i \le r \tag{2.9}$$

with $\alpha_n^{(0)}=0$ by convention. Here the quantity $\alpha_n^{(i)}$ should be interpreted as the expected amount of time that during a service time consisting of i exponential phases there are n additional customers present on top of the customers present at the beginning of that service.

For computational purposes it is important to point out that the recursive calculations in (2.5) may be stopped earlier by using the result that for j large enough the state probability exhibits a geometric behaviour when the batch size distribution satisfies a weak regularity condition. This condition requires that the power series $\beta(x) = \Sigma_{j=1}^{\infty}\beta_j x^j$ has a convergence radius R>1 and that $\beta(x)\to\infty$ as $x\to R$. In words, the batch size distribution should not have an extremely long tail. Since the generating function $P(z) = \Sigma_{j=0}^{\infty} p_j z^j$ can be written as F(z)/G(z) for functions F(z) and G(z) that are analytic in the region $1<|z|<R$ of the complex plane, we can apply partial fraction expansion to obtain for t6he service time density (2.7) that

$$\frac{p_j}{p_{j-1}} \approx \tau \qquad \text{for j large enough} \tag{2.10}$$

where τ is the unique root on (1,R) of the equation

$$1-x-\lambda[1-\beta(x)]\left[\sum_{i=1}^{m} q_i \sum_{j=0}^{i-1} \frac{\mu^j}{\{\lambda(1-\beta(x))+\mu\}^{j+1}}\right] = 0. \qquad (2.11)$$

In practical applications it is not necessary to compute the constant τ from (2.11) on beforehand, but the constant τ can be estimated with any desired accuracy during the recursive computation of the p_j's. For nonlight traffic the asymptotic expansion (2.10) applies already for relatively small values of j. Using this empirical finding one may considerably reduce the computational effort of the recursion algorithm. The practical usefulness of asymptotic expansions in queueing analysis cannot be emphasized enough. Many other examples showing the practical importance of asymptotic expansions can be found in Tijms [25], see also the discussion below.

The above analysis of the algorithm (2.5) assumed a service time density of the form (2.7). In certain applications a hyperexponential density $b(t) = \Sigma_{i=1}^{r} q_i \mu_i e^{-\mu_i t}$ may be more convenient (usually r=2). Then constants A_{jk} are again given by (2.8), where the constants $\alpha_n^{(i)}$ are now recursively computed from $\alpha_0^{(i)} = 1/(\lambda+\mu_i)$ and $\alpha_n^{(i)} = [\lambda/(\lambda+\mu_i)]\Sigma_{j=1}^{n}\beta_j\alpha_{n-j}^{(i)}$, $n\geq 1$.

2.2 Computational methods for the waiting-time probabilities

In this subsection we continue with the $M^X/G/1$ queue. We discuss for this queueing system both exact methods and approximate methods for the computation of the waiting-time probabilities.

In the sequel it will be assumed that customers of different batches are served in order of arrival of the batches and that customers from a same batch are served in order of their (random) positions within that batch. We need the following preliminaries. Suppose that the customers of the batches are numbered as n=1,2,... in accordance with their order of service. By a standard result for backward recurrence times from renewal theory, we have

$\lim_{n\to\infty}$ P(the nth customer takes the kth position in its batch)

$$= \frac{1}{E(B)} \sum_{j=k}^{\infty} \beta_j, \qquad k=1,2,\ldots, \tag{2.12}$$

when we assume that the batch-size distribution $\{\beta_j\}$ is aperiodic (that is, there is no integer d≥2 such that $\Sigma_{n=1}^{\infty}\beta_{nd}=1$). This aperiodicity condition is not required for the following result. For each k≥1,

> the long-run fraction of customers taking the kth positon in their batch

$$= \frac{1}{E(B)} \sum_{j=k}^{\infty} \beta_j \tag{2.13}$$

with probability 1. An intuitive proof of this result is as follows. Fix k and assume that a reward of 1 is received for each customer taking the kth position in its batch. Then the left-hand side of (2.13) can be interpreted as the long-run average reward per customer. The latter average is given by the expected reward per batch divided by the expected batch size. The expected reward per batch equals $\Sigma_{j\geq k}\beta_j$. Note from (2.13) that on the average a customer finds $\frac{1}{2}E(B^2)/E(B)-\frac{1}{2}$ customers from its batch in front of him.

Define now the random variable D_n as the delay in queue of the nth customer, n≥1. Under the assumption that the batch-size distribution is aperiodic, the limit

$$W_q(x) = \lim_{n\to\infty} P(D_n \leq x) \tag{2.14}$$

exists for all x and represents a proper probability distribution, cf. Cohen [5]. Otherwise, define the waiting time distribution function $W_q(x)$ by

$$W_q(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P(D_k \leq x), \qquad x \geq 0. \tag{2.15}$$

The latter limit always exists by the theory of regenerative processes. We have the unifying interpretation that, with probability 1, $W_q(x)$ gives the long-run fraction of customers who have a waiting time of no more than x.

Let us now consider the case that the service time density is given by (2.7). For the computation of the waiting-time probabilities, the computational usefulness of a service time representation (2.7) becomes particularly evident. Then, by representing the service time of a customer by a random number of independent exponential phases, the powerful technique of continuous-time Markov chain analysis can be used. Since each service phase has an exponential distribution with the same mean, the process describing the number of uncompleted service phases in the system is a continuous-time Markov chain. Denote by $\{f_j, j\geq 0\}$ the limiting distribution of this continuous-time Markov chain. It is important to note that this process has the feature that the value of its state cannot decrease with more than 1 at each transition. Then the state probabilities can be *recursively* computed by equating the rate at which the process leaves the *macrostate* of having at least j uncompleted service phases present to the rate at which the system enters that macrostate. This type of argument has a long history in teletraffic analysis, cf. also Kosten [16]. Thus we obtain the recursive balance equations

$$\mu f_j = \sum_{k=0}^{j-1} (\lambda \sum_{i\geq j-k} \nu_i) f_k , \qquad j=1,2,\ldots, \tag{2.16}$$

where $f_0=p_0$ and ν_j denotes the probability that an arriving batch represents a total of j service phases. Obviously, $\nu_j = \Sigma_{n=1}^{j} \beta_n q_j^{n*}$ with q_j^{n*} denoting the probability that n customers represent a total of j service phases. In general the probabilities q_j^{n*} can be recursively computed from the convolution formula, but for the important case that only two q_j's are positive a binomial-type formula can be used.

The conditional waiting time of a customer who has just after arrival j uncompleted service phases in front of him is distributed as the sum of j independent exponentials with the same means and hence has an Erlang-j distribution. Denoting by z_j the long-run fraction of customers who just after arrival j have uncompleted service phases in front of them, we thus obtain

$$W_q(x) = 1 - \sum_{j=1}^{\infty} z_j \sum_{k=0}^{j-1} e^{-\mu x} \frac{(\mu x)^k}{k!}$$

$$= 1 - \sum_{k=0}^{\infty} e^{-\mu x} \frac{(\mu x)^k}{k!} \left(1 - \sum_{j=0}^{k} z_j\right) , \quad x \geq 0. \tag{2.17}$$

The latter representation for $W_q(x)$ is better suited for numerical calculations since it converges faster. The z_j's are easily expressed in the probabilities f_j. By the property Poisson arrivals see time averages, the equilibrium probability f_j has also the interpretation $f_j = \lim_{m\to\infty} P$ {there are j uncompleted service phases present just before the arrival of the mth batch}. Using this result together with (2.13), we obtain

$$z_j = \sum_{i=0}^{j} f_i \sum_{k=1}^{j-i+1} \eta_k \, q_{j-i}^{(k-1)*} \quad , \quad j=0,1,\ldots, \tag{2.18}$$

where $\eta_k = [1/E(B)]\Sigma_{j\geq k}\beta_j$. The desired probabilities z_j can be directly computed from a recursive scheme. It is a matter of simple algebra to derive from (2.16) and (2.18) that

$$z_j = f_0 \sum_{k=1}^{j+1} \eta_k \, q_j^{(k-1)*} + \frac{\lambda}{\mu} \sum_{i=0}^{j-1} z_i \left(1 - \sum_{n=0}^{j-1-i} \nu_n\right) , \quad j=1,2,\ldots, \tag{2.19}$$

where $z_0 = f_0/E(B)$ and $f_0 = 1-\lambda E(B)E(S)$. For nonlight traffic the computational effort in the recursion scheme (2.19) can be reduced by using that, for some constant τ, $z_j/z_{j-1} \approx \tau_0$ for j large enough provided the batch size distribution has no extremely long tail. The constant τ can be estimated during the recursive computations.

Unlike the algorithm for the state probabilities, the algorithm for the waiting time probabilities in the $M^X/G/1$ queue with hyperexponential services becomes more complicated than the corresponding algorithm for the case of Erlangian services. An exact analysis for the case of hyperexponential service requires a continuous-time Markov chain with two-dimensional states describing the number of customers present and the type of the exponential service in progress. The equilibrium state probabilities of this Markov chain are easily computed. However, the computations of the waiting time distribution is more elaborate and involves the computation of the conditional waiting time distributions as first-passage time distributions in a continuous-time Markov chain using the uniformization technique. We refer for details to Eikeboom and Tijms [11].

For practical applications a more simple approach to the $M^X/H_2/1$ queue can be based on the asymptotic expansion

$$1 - W_q(x) \approx \gamma e^{-\theta x} \qquad \text{for x large enough,} \qquad (2.20)$$

when it is assumed that the batch size distribution has no extremely long tail. The constants γ and θ are easily computed for the case of H_2 service, see Van Ommeren [29]. For nonlight traffic the asymptotically exponential expansion (2.20) applies already for moderate values of x. Also, using the asymptotic expansion (2.20) and explicit expressions for the first two moments of the waiting time, the delay probability and the derivative of $W_q(x)$ at x=0, a practically useful approximation to $1-W_q(x)$ for all $x \geq 0$ can be obtained by a sum of three exponential functions, see Van Ommeren [30].

To conclude this section we discuss two-moment approximations for the $M^X/G/1$ queue when c_S^2 is not too large. Such approximations are possible in view of the empirical finding that for nonlight traffic the waiting-time probabilities in the $M^X/G/1$ queue are quite insensitive to more than the first two moments of the service time S provided c_S^2 is not too large, say $0 \leq c_S^2 \leq 2$. Thus for the $M^X/G/1$ queue with $0<c^2{}_S \leq 1$ the waiting-time probabilities can be approximately obtained by fitting a mixture of E_{k-1} and E_k distributions with the same scale parameters to the service time S by matching the first two moments (cf. appendix C in Tijms [25]) and next applying the algorithm (2.17). In many practical applications only the probabilities in the tail of $1-W_q(x)$ are of interest. For the $M^X/G/1$ queue with $0 \leq c_S^2 \leq 2$ these tail probabilities (or, equivalently, the higher percentiles) can be accurately approximated by using the easily computed asymptotically exponential expansions of the waiting time distributions for the particular cases of deterministic services and exponential services. This simple and practically useful approach will be discussed in more detail in section 4.2 when considering the multi-server $M^X/G/c$ queue.

3. Computational methods for the GI/G/1 queue

This section presents algorithms and approximations for the waiting time probabilities in the single-server GI/G/1 queue with service in order of arrival. We first discuss exact methods for two particular cases of phase-type services. Next we show that the exact results for phase-type services can be used to obtain accurate approximations for the case of general service by applying the powerful approach of linear interpolation with respect to the squared coefficient of variation of the service time.

3.1 An embedded Markov chain algorithm for the $GI/E_{1,\ldots,r}/1$ queue

In the $GI/E_{1,\ldots,r}/1$ queue the customers arrive according to a renewal process and the service time of a customer is distributed as a mixture of Erlangians and has as density

$$b(t) = \sum_{i=1}^{r} q_i \mu^i \frac{t^{i-1}}{(i-1)!} e^{-\mu t}, \quad t \geq 0, \tag{3.1}$$

where $q_i \geq 0$ and $\Sigma_i q_i = 1$. Recall that any service-time density can be arbitrarily closely approximated by a density of the form (3.1). It is assumed that the offered load is smaller than 1.

Denoting by D_n the delay in queue of the nth customer, we wish to compute the stationary waiting-time distribution $W_q(x) = \lim_{n\to\infty} P(D_n \leq x)$ for $x \geq 0$ when service is in order of arrival. Following Bux [3], we use the basic technique of embedded Markov chains together with the physical interpretation of the density (3.1). Since the service time of a customer is distributed as the random sum of independent exponential phases with the same means and the exponential distribution is memoryless, the embedded process $\{X_n\}$ defined by

X_n = the number of uncompleted service phases present just prior to the arrival of the nth customer

is a discrete-time Markov chain. This Markov chain is aperiodic and positive recurrent and thus the limit $z_j = \lim_{n\to\infty} P(X_n = j)$ exists for all

$j \geq 0$. Since the conditional waiting time of a customer finding upon arrival j phases present has an Erlang-j distribution, it follows that

$$W_q(x) = 1 - \sum_{j=1}^{\infty} z_j \sum_{k=0}^{j-1} e^{-\mu x} \frac{(\mu x)^k}{k!}, \qquad x \geq 0 . \tag{3.2}$$

Hence we have a useful algorithm for the waiting time distribution provided the limiting probabilities z_j can be efficiently computed. As observed below (2.17), for numerical calculations it is recommended to rewrite (3.2) by interchanging the order of summation.

The probabilities z_j are the unique solution to the well-known balance equations

$$z_j = \sum_{i=0}^{\infty} p_{ij} z_i , \qquad j=0,1,\ldots \tag{3.3}$$

together with the normalization equation $\Sigma_{j=0}^{\infty} z_j = 1$. Here the p_{ij}'s are the one-step transition probabilities of the Markov chain $\{X_n\}$. Denoting by A(t) the probability distribution function of the interarrival time, it is easy to see that for any $i \geq 0$,

$$p_{ij} = \sum_{k=1}^{r} q_k \int_0^{\infty} e^{-\mu t} \frac{(\mu t)^{i+k-j}}{(i+k-j)!} dA(t) \quad \text{for } 1 \leq j \leq i+r , \tag{3.4}$$

while $p_{i0} = 1 - \Sigma_{j\geq 1} p_{ij}$. To solve numerically the linear equations for the equilibrium distribution of a Markov chain, iterative methods such as Gauss-Seidel or successive overrelaxation are highly recommended, see appendix D in Tijms [25]. However, the system of linear equations (3.3) is infinite and has therefore to be truncated to a finite system of linear equations. A straightforward truncation is based on choosing a sufficiently large integer M such that $\Sigma_{j>M} z_j < \varepsilon$ with ε some small accuracy number. For high traffic the truncation integer M will become very large which leads to excessive computing times. For the case of nonlight traffic there is a better alternative. We have the asymptotic result that

$$\frac{z_j}{z_{j-1}} \approx 1 - \frac{\theta}{\mu} \qquad \text{for j large enough,} \tag{3.5}$$

where θ is the unique solution on $(0,\mu)$ of the equation

$$\int_0^\infty e^{-\theta t}\, dA(t) \times \sum_{k=1}^{r} q_k \frac{\mu^k}{(\mu-\theta)^k} = 1. \tag{3.6}$$

A simple proof of (3.5) proceeds as follows. By general results in Neuts [18] for Markov chains with special structure (note that the Markov chain $\{X_n\}$ can also be associated with a batch-arrival $GI^Q/M/1$ queue), we have that $z_j \approx \alpha\xi^j$ for j large enough for some $\alpha>0$ and $0<\xi<1$. Substituting this asymptotic expansion into (3.2) and using the general result from Feller [12] that $1-W_q(x) \approx \gamma e^{-\theta x}$ for x large enough for some $\gamma>0$ and θ given by (3.6), the relation (3.5) follows.

In view of the empirical fact that for nonlight traffic (3.5) holds already for moderate values of j, the result (3.5) can be used as follows for computational purposes. For an appropriately chosen integer N we replace z_j by $z_N(1-\theta/\mu)^{j-N}$ for $j>N$ in the linear equations for the z_j's. In this way we get a finite system of linear equations. For the case of *nonlight* traffic the integer N for which $z_j \approx z_N(1-\theta/\mu)^{j-N}$ accurately holds for $j\geq N$ will be typically much smaller than the earlier discussed truncation integer M. It is noted that when using an iterative method for solving the equations the chosen value for N can be dynamically adjusted during the iterative calculations. Also, from a computational point of view, it is important to point out that the probabilities p_{ij} for $j\neq 0$ depend on i and j only through the difference i-j, see (3.4). This simple observation reduces considerably the memory requirements for the computations. For deterministic or phase-type arrivals, the integrals in (3.4) reduce to explicit expressions, otherwise they can be calculated by Gauss-Laguerre or Gauss-Legendre integration. Finally, the following remark is in order. When applying an iterative method such as Gauss-Seidel or successive overrelaxation to solve the balance equations, it is always important to have an extra accuracy check to verify whether the stopping criterion of the iterative method has served its purposes. For the present model, we have that the long-run fraction of time the server is idle equals $1-\rho$. On the other hand, this long-run fraction equals the average amount of time the server is idle between two arrivals divided by the average time between two arrivals. Thus we have the accuracy check

$$1-\rho = \frac{1}{E(A)} \sum_{j=0}^{\infty} z_j \sum_{i=1}^{r} q_i \left[\int_0^{\infty} t\left(1 - \sum_{k=0}^{i+j-1} e^{-\mu t} \frac{(\mu t)^k}{k!}\right) dA(t) \right.$$
$$\left. - \frac{(i+j)}{\mu} \int_0^{\infty} (1 - \sum_{k=0}^{i+j} e^{-\mu t} \frac{(\mu t)^k}{k!}) \, dA(t) \right] ,$$

where E(A) denotes the average interarrival time.

The above analysis can be straightforwardly extended to the $GI^X/E_{1,\dots,r}/1$ queue with batch arrivals. The analysis applies as well to the finite capacity $GI/E_k/1/N$ queueing system in which a customer finding upon arrival N other customers present is rejected. In the finite capacity model we restrict ourselves to pure Erlangian service since otherwise the number of uncompleted service phases present does not unambiguously determine the number of customers present. Nevertheless the exact results for the $GI/E_k/1/N$ queue can be used to calculate accurate approximations for several performance measures in the GI/G/1/N queue with a general service time S provided that c_S^2 is not too large. To do so, we consider the $GI/E_k/1/N$ queue for two (or three) appropriately chosen values of k. For each $GI/E_k/1/N$ considered we compute the exact value of the performance measure under consideration. Next the approximation for the GI/G/1/N queue is obtained by linear (or quadratic) interpolation of the exact results where the interpolation is based on the squared coefficient of variation of the service time, cf. formula (3.14) in section 3.2. This approach can be used for performance measures such as the average queue size and the average waiting time per entering customer, but not for the waiting-time probabilities and the rejection probability. However, both the waiting-time percentiles and the minimal buffer size guaranteeing a prespecified rejection level allow for interpolation with respect to the squared coefficient variation of the service time, cf. also Nobel [19] and Tijms [25].

To the end of this subsection we mention an alternative approach for the computation of certain long-run averages in the $GI/E_{1,\dots,r}/1$ queue. This alternative approach is the (modified) value-iteration algorithm from Markov decision theory, cf. Chapter 3 in Tijms [25]. Measures such as the delay probability and the average waiting time can be interpreted as an average cost by imposing an appropriate cost structure on the Markov chain $\{X_n\}$ describing the number of phases present just prior to an arrival. For the average waiting-time, suppose that the Markov chain incurs a cost of i/μ in state i. Similarly, for the delay probability the Markov chain incurs a cost of 1 in state $i\neq 0$ and a cost of 0 otherwise; more generally, for the waiting-time probability $1-W_q(t_0)$ a cost of $\Sigma_{k=0}^{i-1}\, e^{\mu t_0}(\mu t_0)^k/k!$ is associated with state i. The value-iteration algorithm does not require solving linear equations to compute the average cost but involves only recursive computations. Moreover, this approach has the important feature of providing at each iteration lower and upper bounds on the average cost. Thus the value-iteration approach may be used to calculate quick bounds for engineering purposes.

3.2 Approximations based on exact results for the $GI/C_2/1$ queue

The $GI/C_2/1$ queue denotes the particular GI/G/1 queue in which the service time S has a Coxian-2 (C_2) density. That is,

$$S = \begin{cases} U_1 & \text{with probability } 1-b \\ U_1+U_2 & \text{with probability } b \end{cases}$$

for some $0\leq b\leq 1$, where U_1 and U_2 are two independent exponentials with respective means $1/\mu_1$ and $1/\mu_2$. The class of C_2-densities coincides with the class of K_2-density, where the probability density of a posiive random variable is said to be a K_2-density when its Laplace transform is the ratio of a polynomial of at most degree 1 to a polynomial of degree 2. Any C_2-density has a squared coefficient of variation of at least 1/2. It is often convenient to fit a C_2-density to a positive random variable by matching its first two or its first three moments. Let S be a positive random variable with $c_S^2\geq \frac{1}{2}$ and $m_i=E(S^i)$ denoting the i^{th} moment of S. If a three-moment fit to S by a C_2-density exists, the three parameters of this fit are given by

$$\mu_{1,2} = \tfrac{1}{2}(a_1\pm\sqrt{a_1^2-4a_2}), \quad b = \frac{\mu_2}{\mu_1}(\mu_1 m_1-1), \qquad (3.7)$$

where $a_1 = 1/m_1 + \frac{1}{2} m_2 a_2/m_1$ and $a_2 = (6m_1 - 3m_2/m_1)/[(3m_2^2/2m_1) - m_3]$. An infinite number of C_2-densities can be fitted to S by matching only the first two moments. A very useful two-moment fit is the one with parameters

$$\mu_{1,2} = \frac{2}{m_1}\left[1 \pm \sqrt{\frac{c_S^2 - \frac{1}{2}}{c_S^2+1}}\right], \quad b = \frac{\mu_2}{\mu_1}(\mu_1 m_1 - 1). \tag{3.8}$$

This particular C_2-density has the same first three moments as a gamma density.

For the $GI/C_2/1$ queue a tractable closed-form expression can be given for the stationary waiting-time distribution function $W_q(x)$, $x \geq 0$. The closed-form solution requires the computation of the two real zeros η_1 and η_2 with $0<\eta_1<\min(\mu_1,\mu_2)\leq\eta_2$ of the equation

$$x^2 - (\mu_1+\mu_2)x + \mu_1\mu_2 - (\{\mu_1\mu_2 - (1-b)\mu_1 x\}\int_0^\infty e^{-xt}dA(t) = 0 . \tag{3.9}$$

Here A(t) is the probability distribution function of the interarrival time. It follows from the general relations (5.190) and (5.191) in Cohen [5] that

$$1 - W_q(x) = \sum_{i=1}^{2} a_i e^{-\eta_i x} \quad \text{for all} \quad x \geq 0 , \tag{3.10}$$

and

$$W_q(0) = \frac{\eta_1\eta_2}{\mu_1\mu_2}, \quad \int_0^\infty x dW_q(x) = -\left(\frac{\mu_1+\mu_2}{\mu_1\mu_2}\right) + \frac{1}{\eta_1} + \frac{1}{\eta_2} . \tag{3.11}$$

Here the constants a_i are given by

$$a_1 = \frac{-\eta_1^2\eta_2 + \eta_1\eta_2(\mu_1+\mu_2) - \eta_2\mu_1\mu_2}{\mu_1\mu_2(\eta_1-\eta_2)} \tag{3.12}$$

$$a_2 = \frac{\eta_1\eta_2^2 - \eta_1\eta_2(\mu_1+\mu_2) + \eta_1\mu_1\mu_2}{\mu_1 u_2(\eta_1-\eta_2)} . \tag{3.13}$$

The practical usefulness of the exact result (3.10) should not be underestimated and is not restricted to C_2-service. The easily computed results for C_2-service enable us as well to calculate simple and accurate approximations for the GI/G/1 queue with a general service time S. For the case of $c_S^2 \geq \frac{1}{2}$, we suggest to fit a C_2-density to the service time S by using the three-moment fit (3.7) or the two-moment fit (3.8) and to use the results (3.10)-(3.11) for this C_2-density. This will usually give excellent approximations provided the traffic load is not very low and c_S^2 is not very large. For the case of $c_S^2 < \frac{1}{2}$, we suggest to proceed as follows. To approximate a certain performance measure P in the GI/G/1 queue, the idea is to compute the exact values of the performance measure for two or three queueing systems of the $GI/C_2/1$ type and next to apply extrapolation with respect to the squared coefficient of variation of the service time. Each of the $GI/C_2/1$ models used should have the same interarrival-time distribution and the same mean service time as the original GI/G/1 model. The general form of the extrapolation formula is as follows. Suppose that the exact value P(i) of the performance measure is known for the $GI/C_2/1$ queue with squared coefficient of variation $cv^2(i)$ for $i=1,\ldots,n$. Then the exact value of the performance measure for the GI/G/1 queue with cv^2 as squared coefficient of variation of the service time can be approximated by

$$P_{app}(cv^2) = \sum_{i=1}^{n} P(i) \prod_{\substack{k=1 \\ k\neq i}}^{n} \frac{cv^2 - cv^2(i)}{cv^2(i) - cv^2(k)} . \qquad (3.14)$$

Usually n=2 (linear extrapolation) or n=3 (quadratic extrapolation). This approximation gives good results for performance measures such as the the average waiting time and the waiting-time percentiles, provided the traffic load is not very small. It is an empirical finding that the extrapolation works for all waiting-time percentiles but not for the (small) waiting-time probabilities. In particular, extrapolation should not be used for the delay probability when the traffic load is very small. For nonlight traffic the delay probability may be approximated by using extrapolation. A good source for verifying experimentally the quality of approximations is the tablebook Seelen et al. [21]. It should be pointed out that the approximation approach discussed above makes no approximation with respect to the interarrival time distribution. This observation is important because performance measures are typically much more sensitive to the form of the interarrival-time distribution than to the form of the service-time distribution, as is clearly demonstrated by the numerical results in table 3.1. In many practical applications it suffices to use only the first two moments of the service time, provided c_S^2 is not too large; the range of c_S^2 depends to some extent on the variability in the interarrival time and on the traffic load. However, one should be much more careful in using two-moment fits for the interarrival time distribution. A more extensive discussion of this subject can be found in Tijms [25].

In Table 3.1 we give some numerical results from which it can be verified that good approximations can be obtained by using exact results for the $GI/C_2/1$ queue in conjunction with (linear) extrapolation on the squared coefficient of variation of the service time. For several values of c_A^2, c_S^2 and ρ, we give the exact values of the mean waiting time $E(W_q)$, the delay probability $\Pi_W=1-W_q(0)$ and the conditional waiting time percentile $\eta(p)$ for $p=0.95$. Here c_A^2 and c_S^2 denote the squared coefficients of variation of the interarrival time A and the service time S. The conditional percentile $\eta(p)$ is defined by $P(W_q>\eta(p)|W_q>0)=1-p$, where W_q denotes the waiting time of a customer in the stationary situation. We vary c_A^2 as 0, ½ and 5 and c_S^2 as 0, 0.25, 0.50, 0.75, 1 and 2, where the value 0 corresponds to the deterministic distribution, the value 1/4 to the Erlang-4 distribution and the values 0.5, 0.75, 1 and 5 to C_2-distributions with the normalization (3.8). For $c_S^2=2$, we consider both the C_2 density with the gamma normalization (3.8) and the C_2 density with the normalization of balanced means (i.e. $p/\mu_1=(1-p)/\mu_2$, where $p=1-b\mu_1/(\mu_1-\mu_2)$). The offered load ρ is varied as ρ = 0.2, 0.5 and 0.8. In all examples we have taken $E(S)=1$. The results for the particular case of the $H_2/D/1$ queue ($c_A^2=5$) are obtained by the algorithm in Van Hoorn [28]. The exact results for the $E_2/D/1$ are obtained by using that the $E_k/D/1$ is equivalent to an $M/D/k$ queue, see also subsection 4.3. The other exact results are calculated by using (3.10)-(3.13). The notation 5E-4 in table 3.1 means 5×10^{-4}.

Table 3.1 Numerical results for some GI/G/1 queues

		$c_A^2=0$			$c_A^2=0.5$			$c_A^2=5$		
ρ	c_S^2	$E(W_q)$	Π_W	$\eta(.95)$	$E(W_q)$	Π_W	$\eta(.95)$	$E(W_q)$	Π_W	$\eta(.95)$
0.2	0	-	-	-	0.024	0.064	0.84	0.581	0.527	2.89
	0.25	0.000	3E-6	0.86	0.041	0.074	1.51	0.685	0.515	3.67
	0.50	0.000	5E-4	1.63	0.060	0.081	2.12	0.780	0.505	4.42
	0.75	0.002	0.003	2.42	0.081	0.088	2.72	0.869	0.496	5.15
	1	0.007	0.007	3.02	0.103	0.094	3.31	0.952	0.488	5.85
	2(g)	0.053	0.030	5.27	0.203	0.110	5.63	1.244	0.461	8.50
	2(b)	0.078	0.032	7.30	0.211	0.105	6.77	1.212	0.466	8.90
0.5	0	-	-	-	0.177	0.323	1.38	3.142	0.839	10.7
	0.25	0.017	0.047	1.07	0.281	0.344	2.26	3.320	0.826	11.7
	0.50	0.078	0.116	1.99	0.390	0.360	3.13	3.494	0.814	12.6
	0.75	0.160	0.165	2.90	0.503	0.372	3.99	3.664	0.803	13.6
	1	0.255	0.203	3.76	0.618	0.382	4.85	3.828	0.793	14.5
	2(g)	0.693	0.289	7.19	1.095	0.405	8.26	4.448	0.763	18.0
	2(b)	0.725	0.257	8.85	1.105	0.395	9.26	4.400	0.775	18.2
0.8	0	-	-	-	0.903	0.702	3.61	11.10	0.944	34.7
	0.25	0.319	0.446	2.08	1.381	0.716	5.59	11.61	0.940	36.6
	0.50	0.757	0.548	4.08	1.865	0.726	7.57	12.15	0.937	38.6
	0.75	1.217	0.598	6.08	2.353	0.734	9.54	12.67	0.933	40.6
	1	1.693	0.629	8.07	2.844	0.740	11.5	13.19	0.930	42.5
	2(g)	3.650	0.683	16.0	4.825	0.752	19.4	15.25	0.919	50.3
	2(b)	3.673	0.658	17.4	4.834	8.746	20.3	15.22	0.924	50.5

The above results can also be used to calculate an approximation to the average waiting time per customer in the general $GI^X/G/1$ queue with batch arrivals. This can be done by using the following result. Denoting by $E(W_q)$) the average waiting time per customer in the $GI^X/G/1$ queue and denoting by the generic variables B and S the number of customers in a batch and the service time of a customer, we have

$$E(W_q) = E(W_{q1}) + \tfrac{1}{2}E(S)\left\{ \frac{E(B^2)}{E(B)} - 1 \right\} . \qquad (3.15)$$

Here the first term in the right-hand side of (3.15) denotes the average waiting time per 'supercustomer' in the single-arrival $GI/G_X/1$ queue in which the interarrival time is distributed as in the $GI^X/G/1$ queue and the service time is distributed as the total time needed to service all customers from one batch in the $GI^X/G/1$ queue. The second term in the right side of (3.15) gives for the $GI^X/G/1$ queue the average amount of time per customer between the moment the batch of the customer arrives and the moment the customer enters service, cf. the relation (2.13) in section 2. Using that the mean and the squared coefficient of variation of the service time U of a customer in the $GI/G_X/1$ queue are given by

$$E(U) = E(B)E(S) \quad \text{and} \quad c_U^2 = \frac{1}{E(B)}\, c_S^2 + c_B^2 ,$$

the average waiting time per customer in the $GI^X/G/1$ can be approximated by using the approach discussed above.

4. Approximations for multi-server queues

In this section we discuss some simple approximations for multi-server queues. For the general multi-server queue it is difficult to give exact algorithms that lend themselves to practical computations. Therefore in many situations one has to resort to approximate methods for calculating measures of system performance. Useful approximations to complex queueing systems are often obtained by using exact results for simpler related queueing systems. It will be seen again that asymptotic expansions and linear interpolation with respect to the squared coefficient of variation of the service time are powerful tools.

4.1 The M/G/c queue

In this section we consider the multi-server queue with c identical servers, Poisson arrivals at rate λ and a generally distributed service time S. It is assumed that $\rho=\lambda E(S)/c$ is smaller than 1. The M/G/c queue does not permit a tractable exact analysis with the exception of the special cases of the M/M/c queue and the M/D/c queue. A tractable analytical solution is even not possible for a simple measure as the average waiting time per customer. Nevertheless useful approximations for the waiting time can be obtained by having a closer look at the structure of the Pollaczek-Khintchine formula for the case of c=1 server. Denote by $E(W_q)$ the average waiting time per customer in the general M/G/c queue and let $E_{exp}(W_q)$ and $E_{det}(W_q)$ represent $E(W_q)$ for the special cases of exponential services and deterministic services with the same means E(S). The Pollaczek-Khintchine formula for $E(W_q)$ for the case of c=1 server allows the following two representations,

$$E(W_q) = \tfrac{1}{2}(1+c_S^2)E_{exp}(W_q) \tag{4.1}$$

and

$$E(W_q) = (1-c_S^2)E_{det}(W_q) + c_S^2 E_{exp}(W_q) \ , \tag{4.2}$$

where c_S^2 denotes the squared coefficient of variation of the service time. For the case of c>1, the right-hand side of (4.1) can be used as a first-order approximation to $E(W_q)$. This is a practically useful approximation that is reasonably accurate provided c_S^2 is not too large. A more accurate approximation of a comparable simplicity is obtained by using the right-hand side of (4.2) in conjunction with the special-purpose approximation to $E_{det}(W_q)$,

$$E_{det}^{app}\,(W_q) = \tfrac{1}{2}\left[1 + (1-\rho)(c-1)\,\frac{\{\sqrt{4+5c}\,-\,2\}}{16\rho c}\right] E_{exp}(W_q) \ . \tag{4.3}$$

This approximation was advocated by Cosmetatos [7] and is a quite accurate approximation to $E(W_q)$ provided c_S^2 is not too large, say

$0 \le c_S^2 \le 2$. Unlike the M/G/1 queue, it is no longer true for the multi-server queue that the average waiting time is insensitive to more than the first two moments of the service time. The average waiting time becomes increasingly sensitive to the shape of the service time distribution as c_S^2 gets larger. The papers Boxma et al. [2] and Tijms et al. [24] give approximations of the form

$$E_{app}(W_q) = \tfrac{1}{2} f(1+c_S^2) E_{exp}(W_q), \qquad (4.4)$$

where f is some correction factor involving the service-time distribution. Unlike the approximations based on (4.1)-(4.3), the latter approximations can also be used for larger values of c_S^2, see Tijms [25] for details. This reference discusses also approximations for the state probabilities and the waiting-time probabilities. In particular, the delay probability $P(W_q>0)$ can be very well approximated by Erlang's delay probability $\Pi_W(\exp)$ for the M/M/c queue. We briefly discuss approximations to the waiting-time percentiles. It is convenient to do this for the conditional waiting-time percentiles $\eta(p)$ which are defined by $P(W_q>\eta(p)|W_q>0) = 1-p$ for *all* $0<p<1$. Here W_q has the limiting distribution of the waiting time of a customer. Denoting by $\eta_{exp}(p)$ and $\eta_{det}(p)$ the corresponding percentiles for the special cases of the M/M/c queue and the M/D/c queue with the same means E(S), an excellent approximation to $\eta(p)$ is given by

$$\eta_{app}(p) = (1-c_S^2)\eta_{det}(p) + c_S^2 \eta_{exp}(p) \quad \text{for all} \quad 0<p<1, \qquad (4.5)$$

provided that c_S^2 is not too large ($0 \le c_S^2 \le 2$) and the traffic load is not too small (say, $\Pi_W(\exp) \ge 0.2$). The percentile $\eta_{exp}(p)$ is trivial to compute, while a relatively simple algorithm can be given for the computation of $\eta_{det}(p)$, see Tijms [25]. For the higher values of p the asymptotic expansion of the waiting-time distribution can be used to compute $\eta_{det}(p)$, see also subsection 4.2 below. Also it is pointed out that for quick engineering calculations the first-order approximation $\tfrac{1}{2}(1+c_S^2)\eta_{exp}(p)$ to $\eta(p)$ can be used when p is close enough to 1 (say, $p \ge 1-\Pi_W(\exp)$), cf. Tijms [26].

4.2 The $M^X/G/c$ queue

In the $M^X/G/c$ queue the customers arrive in batches rather than singly. The arrival process of batches is a Poisson process with rate λ, the service time S of a customer has a general distribution and the batch size B has a probability distribution $\{\beta_j,\ j\geq 1\}$. It is assumed that $\lambda E(B)E(S)/c$ is less than 1. Also, we make the technical assumption that the power series $\beta(x) = \Sigma_{j=1}^{\infty}\beta_j x^j$ has a convergence radius R larger than 1 and that $\beta(x)\rightarrow\infty$ as $x\rightarrow R$. This assumption is satisfied in cases of practical interest, e.g. when $\{\beta_j\}$ has finite support.

Supposing that customers from different batches are served in order of arrival and that customers from the same batch are served according to their random positions in the batch, denote by D_n the delay in queue of the nth customer and let $W_q(x) = \lim_{n\rightarrow\infty}(1/n)\Sigma_{k=1}^{n}P(D_k\leq x)$ for $x\geq 0$. Also, denote by $\{p_j,\ j\geq 0\}$ the limiting distribution of the number of customers in the system. Using generating-function analysis and partial-fraction expansions, asymptotic expansions for $1-W_q(x)$ can be obtained for the particular cases of exponential and deterministic services. For the $M^X/M/c$ queue, it holds that (see Cromie et al. [8]),

$$1 - W_q(x) \approx \frac{(1/\rho)(\tau-1)\displaystyle\sum_{k=0}^{c-1}(c-k)p_k^{\exp}\tau^{k-c-1}}{(\lambda/\mu)[1-\tau(1-\tau)\beta'(\tau)-\beta(\tau)]}\, e^{-c\mu(1-1/\tau)x} \qquad (4.6)$$

for x large enough, where $\mu=1/E(S)$ and τ is the unique root on (1,R) of the equation

$$c(1-x) - (\lambda/\mu)x(1-\beta(x)) = 0. \qquad (4.7)$$

For the $M^X/D/c$ queue with the constant service time D, we have

$$1 - W_q(x) \approx \gamma e^{-\lambda(\beta(\tau)-1)x} \qquad \text{for x large enough} \qquad (4.8)$$

with

$$\gamma = \frac{(\beta(\tau)-1)}{(\tau-1)^2 E(B)}\left\{c\tau^{c-1} - \lambda D\beta'(\tau)e^{-\lambda D(1-\beta(\tau))}\right\}^{-1}\sum_{j=0}^{c-1}p_j^{\det}(\tau^j-\tau^c), \qquad (4.9)$$

where τ is the unique root on (1,R) of the equation

$$1 - x^c \ e^{\lambda D(1-\beta(x))} = 0. \tag{4.10}$$

The state probabilities p_j^{exp} for the $M^X/M/c$ are computed from a recursive scheme like (2.16), while the state probabilities p_j^{det} are computed by solving a generalization of the equilibrium equations derived by Crommelin [9] for the M/D/c queue, see Eikeboom and Tijms [11] for details. The first c state probabilities yield also the delay probability by

$$1-W_q(0) = 1 - \sum_{j=0}^{c-1} p_j \sum_{i=1}^{c-j} \eta_i \tag{4.11}$$

with $\eta_i = [1/E(B)] \ \Sigma_{k\geq i} \ \beta_k$, cf. (2.13) in section 2.

Denote by $\eta(p)$ the pth conditional waiting-time percentile for the $M^X/G/c$ queue. For the batch-arrival model the approximation based on interpolation of the percentiles for the particular cases of deterministic and exponential services works only for the higher percentiles. For p close enough to 1, a useful approximation to $\eta(p)$ is given by

$$\eta_{app}(p) = (1-c_S^2)\eta_{det}^{asy}(p) + c_S^2\eta_{exp}^{asy}(p) \tag{4.12}$$

provided that c_S^2 is not too large and the traffic load is not too small (that is, the probability that an arriving batch finds all servers busy should not be too small). Here $\eta_{det}^{asy}(p)$ and $\eta_{exp}^{asy}(p)$ are the asymptotic percentiles which are easily computed from (4.8) and (4.6) together with (4.11). As a rule of thumb, the approximation (4.12) can be used when $p\geq 1-\Pi_B(\exp)$ with $\Pi_B(\exp)=1-\Sigma_{j=0}^{c-1}p_j^{exp}$. The average waiting time per customer can be accurately approximated by using an interpolation as in (4.2), see Eikeboom and Tijms [11]. It should be pointed out that the simple first-order approximation (4.1) cannot be used for the batch-arrival model, as can be seen from formula (3.15). In table 4.1 we give for the $M^X/E_2/c$ queue the exact and approximate values for the conditional waiting-time percentiles $\eta(p)$ for several values of c and p. The batch size B is constant or geometrically distributed with mean E(B)=3. In all cases the normalization E(S)=1 is used.

Table 4.1 The conditional waiting-time percentiles $\eta(p)$ for the $M^X/E_2/c$ queue

			constant				geometric			
c	ρ	p	0.80	0.90	0.95	0.99	0.80	0.90	0.95	0.99
1	0.2	exa	2.927	3.945	4.995	7.458	5.756	8.122	10.49	15.98
		app	2.836	3.901	4.967	7.440	5.745	8.116	10.49	15.99
1	0.5	exa	5.107	7.170	9.231	14.02	9.044	12.84	16.64	25.45
		app	5.089	7.154	9.219	14.01	9.040	12.84	16.64	25.47
2	0.2	exa	1.369	1.897	2.431	3.661	2.989	4.172	5.355	8.101
		app	1.354	1.887	2.419	3.656	2.982	4.167	5.353	8.106
2	0.5	exa	2.531	3.561	4.592	6.985	4.600	6.498	8.395	12.80
		app	2.535	3.567	4.599	6.996	4.601	6.501	8.401	12.81
5	0.2	exa	0.621	0.845	1.063	1.560	1.298	1.773	2.246	3.345
		app	0.640	0.853	1.066	1.560	1.305	1.779	2.253	3.354
5	0.5	exa	1.063	1.476	1.889	2.846	1.898	2.657	3.417	5.179
		app	1.069	1.482	1.895	2.853	1.905	2.665	3.425	5.190
10	0.5	exa	0.553	0.764	0.971	1.451	0.980	1.360	1.740	2.622
		app	0.566	0.772	0.979	1.458	0.991	1.371	1.751	2.634
10	0.7	exa	0.923	1.295	1.667	2.530	1.547	2.181	2.815	4.287
		app	0.930	1.302	1.673	2.536	1.556	2.190	2.824	4.297

4.3 The GI/G/c queue

It seems obvious that the general GI/G/c queue offers enormous difficulties in getting practically useful results. Nevertheless, using powerful numerical techniques for solving large systems of linear equations for structured Markov chains, the continuous-time Markov chain method has proved to be extremely useful for analysing the GI/G/c queue when the interarrival time and service time both have special phase-type distributions. By a detailed state description involving sufficient information about the number of customers present and the status of both the services in progress and the arrival in progress, it is possible to set up the equilibrium equations for the microstate probabilities. This system of linear equations possesses a structure enabling the application of special aggregation/disaggregation algorithms to solve the equations numerically. We refer for details of this approach to Seelen [22] and Takahashi and Takami [23]. In this section we mention only some partial results obtained by simple approximation methods.

Numerical investigations reveal that in many practical GI/G/c queueing systems performance measures P such as the average queue length, the average waiting time and the (conditional) waiting time percentiles may reasonably well be approximated by using the interpolation formula

$$P^{app}_{GI/G/c} = (1-c_S^2)P_{GI/D/c} + c_S^2 P_{GI/M/c} \qquad (4.13)$$

provided c_S^2 is not too large. Here $c^2{}_S$ denotes the squared coefficient of variation of the service time and $P_{GI/D/c}$ and $P_{GI/M/c}$ denote the exact values of the performance measure under consideration for the special cases of the GI/D/c queue and the GI/M/c queue with the same mean service times. The approximation (4.13) takes into account the empirical finding that measures of system performance are typically much more sensitive to the shape of the interarrival-time distribution than to the shape of the service-time distribution. The particular models of the GI/D/c queue and the GI/M/c queue allow for a relatively simple algorithmic analysis, see Van Hoorn [28] and Cooper [6]. The special case of the E_k/D/c queueing model can be analysed by solving the M/D/ck queueing model with the same server utilization. It is not difficult to show that these two models have the same stationary waiting time distribution, see Iversen [15]. This result may be used as well to approximate measures of system performance in the GI/D/c by solving first the E_k/D/c queue for two (or three) appropriately chosen values of k and by interpolating next the exact results with respect to the squared coefficient of variation of the interarrival time.

Much research remains still to be done for the general GI/G/c queue. In particular, tractable exact methods for the GI/C_2/c queue would be very useful. Some progress for this queueing model has been made by Bertsimas [1] and De Smit [10]. Exact results for the GI/C_2/c queue with $c_s^2 \geq \frac{1}{2}$ together with interpolation with respect to the squared coefficient of variation of the service time would provide a powerful method to get practically useful approximations for the general GI/G/c queue.

References

1. Bertsimas, D. (1988), "An analytic approach to a general class of G/G/s queueing systems," (to appear in *Oper. Res.*)

2. Boxma, O.J., Cohen, J.W., and Huffels, N. (1979), "Approximations of the mean waiting time in an M/G/s queueing system", *Oper. Res.* 27, 1115-1127.

3. Bux, W. (1979), "Single server queues with general interarrival and phase type service time distributions, in *Proceedings 9th ITC*, Torremolinos, paper 413.

4. Cohen, J.W. (1976), *On Regenerative Processes in Queueing Theory*, Lecture Notes in Economics and Mathematical Systems, Vol. 121, Springer-Verlag, Berlin.

5. Cohen, J.W. (1982), *The Single Server Queue*, 2nd ed., North-Holland, Amsterdam.

6. Cooper, R.B. (1981), *Introduction to Queueing Theory*, Edward Arnold, London.

7. Cosmetatos, G.P. (1975), "Approximate explicit formulae for the average queueing time in the processes M/D/r and D/M/r", *INFOR* 13, 328-331.

8. Cromie, M.V., Chaudry, M.L., and Grassmann, W.K. (1979), "Further results for the queueing system $M^X/M/c$", *J.Oper.Res.Soc.* 30, 755-763.

9. Crommelin, C.D. (1932), "Delay probability formulae when the holding times are constant", *P.O.Elect.Engr.I.* 25, 41-50.

10. De Smit, J.H.A. (1983), "A numerical solution for the multi-server queue with hyperexponential service times", *Operat.Res. Letters* 2, 217-224.

11. Eikeboom, A.M., and Tijms, H.C. (1987), "Waiting time percentiles in the multi-server $M^x/G/c$ queue with batch arrivals", *Prob. Engr. Inform. Sc.* 1, 75-98.

12. Feller, W. (1971), *An Introduction to Probability Theory and its Applications*, Vol.II, 2nd ed., Wiley, New York.

. Fry, Th.C. (1928), *Probability and its Engineering Uses*, D.van Nostrand, New York.

14. Hordijk, A., and Tijms, H.C. (1976), "A simple proof of the equivalence of the limiting distributions of the continuous time and embedded process of the queue size in the M/G/1 queue", *Statistica Neerlandica* 30, 97-100.

15. Iversen, V.B. (1983), "Decomposition of an M/D/r.k queue into k E_k/D/r queues with FIFO", *Oper. Res. Letters* 2, 20-21.

16. Kosten, L. (1973), *Stochastic Theory of Service Systems,* Pergamon Press, London.

17. Molina, E.C. (1927), "Application of the theory of probability to telephone trunking problems", *Bell Syst. Techn. J.* 6, 461-494.

18. Neuts, M.F. (1981), *Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach*, The John Hopkins University Press, Baltimore.

19. Nobel, R.D. (1988), "Practical approximations for finite-buffer queueing models with batch arrivals" (to appear in *Europ. J. Oper. Res.*)

20. Schassberger, R. (1973), *Warteschlangen*, Springer-Verlag, Berlin.

21. Seelen, L.P., Tijms, H.C., and Van Hoorn, M.H. (1985), *Tables for. Multi-Server Queues*, North-Holland, Amsterdam.

22. Seelen, L.P. (1986), "An algorithm for Ph/Ph/c queues", *European J. Operat. Res.* 23, 118-127.

23. Takahashi, Y., and Takami, Y. (1976), "A numerical method for the steady-state probabilities of a GI/G/c queueing system in a general class", *J. Operat. Res. Soc. Japan* 19, 147-157.

24. Tijms, H.C., Van Hoorn, M.H., and Federgruen, A. (1981), "Approximations for the steady-state probabilities in the M/G/c queue", *Adv.Appl.Prob.* 13, 186-206.

25. Tijms, H.C. (1986), *Stochastic Modelling and Analysis: A Computational Approach*, Wiley, Chichester.

26. Tijms, H.C. (1987), "A simple and useful approximation to the waiting time distribution in the multi-server queue with priorities, pp. 161-169 in G. Iazeolla, P.J. Courtois and O.J. Boxma (eds.), *Applied Mathematics and Performance/Reliability Models of Computer/Communication Systems*, North-Holland,. Amsterdam.

27. Van Hoorn, M.H. (1984), *Algorithms and Approximations for Queueing Systems*, CWI Tract No.8, CWI, Amsterdam.

28. Van Hoorn, M.H. (1986), "Numerical analysis of multi-server queues

with deterministic services and special phase-type arrivals, *Zeitschrift für Oper. Res.* 30, 15-28.

29. Van Ommeren, J.C.W. (1988), "Exponential expansion for the tail of the waiting time probability in the single server queue with batch arrivals (to appear in *Adv. Appl. Prob.*).

30. Van Ommeren, J.C.W. (1988), "Simple approximations for the batch arrival $M^X/G/1$ queue", Research Report 1-88, Institute for Econometrics, Vrije Universiteit, Amsterdam.

31. Wolff, R.W. (1982), "Poisson arrivals see time averages", *Oper. Res.* 30, 223-231.