20. Sennott, L. I. (1986). A new condition for the existence of optimal stationary policies in average cost Markov decision processes. *Operations Research Letters* 5:17–23.
21. Sennott, L. I. (1986). A new condition for the existence of optimum stationary policies in average cost Markov decision processes—unbounded cost case. *Proceedings of the 25th IEEE Conference on Decision and Control.* Athens, Greece, pp. 1719–1721.
22. Sennott, L. I. Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs. *Operations Research* (to appear).
23. Stidham, S., Jr. (1978). Socially and individually optimal control of arrivals to a GI/M/1 queue. *Management Science* 24:1598–1610.
24. Weber, R.R. & Stidham, Jr. S. (1987). Optimal control of service rates in networks of queues. *Advances in Applied Probability* 19:202–218.
25. Widder, D.V. (1946). *The Laplace transform.* New Jersey: Princeton University Press.

Probability in the Engineering and Informational Sciences, 3, 1989, 273–287. Printed in the U.S.A.

A MOMENT-ITERATION METHOD FOR APPROXIMATING THE WAITING-TIME CHARACTERISTICS OF THE GI/G/1 QUEUE

A. G. DE KOK
Nederlandse Philips Bedrijven B.V.
Center for Quantitative Methods
The Netherlands

In this paper, a moment-iteration method is introduced. The method is used to solve Lindley's integral equation for the GI/G/1 queue. From several forms of this integral equation, we derive the first two moments of the waiting-time distribution, the waiting probability, and the percentiles of the conditional waiting time. Numerical evidence is given that the method yields excellent results. The flexibility of the method provides the opportunity to solve the GI/G/1 queue for all interarrival time distributions of practical interest. To show that the moment-iteration method is generally applicable, we give some results for an (s,S)-model with order-size-dependent lead times and finite production capacity of the supplier.

1. INTRODUCTION

In this paper, we present a new practical method to compute approximations for the delay probability and the average waiting time for the GI/G/1 queue. In fact, an approximation for the waiting-time distribution is derived. The method is based on an iterative scheme to approximate the first two or three moments of the waiting time of an arbitrary arriving customer from Lindley's integral equation (cf. Kleinrock [7]). It turns out that the iterative method gives excellent approximations for the notorious GI/D/1 queue and the D/G/1

practical solutions to problems that are unsolved at present (cf. De Kok [4,5]). To illustrate the general applicability of the moment-iteration method, we describe in Section 5 the (s,S)-inventory model with lead times depending on the order size as studied in De Kok [5].

In the literature on the GI/G/1 queue, a lot of solution procedures are suggested to find exact or approximate expressions for the delay probability and the average waiting time.

Exact solutions to the GI/G/1 queue can be found in Neuts [9] and Seelen [10] for the Ph/Ph/1 queue with phase-type interarrival times and phase-type service times and in Bux [3] for the GI/$E_{r,s}$/1 queue in which the service-time distribution is a mixture of Erlangian distributions with the same scale parameters (cf. also Tijms [13]). The latter reference gives an excellent survey on exact and approximate methods for queueing models, including the GI/G/1 queue.

Approximation methods aim at accurate approximations at low computational costs. However, most of these methods appear to be either dependent on the type of interarrival and service-time distributions, such as the excellent approximation for the D/G/1 queue of Fredericks [6], or dependent on the traffic intensity ρ, such as the diffusion approximation by Whitt [17], which applies for moderate and heavy traffic. The same holds for the approximations given by Seelen and Tijms [11], which apply when the coefficients of variation of both the interarrival time and the service time is less then 1. A robust approximate solution method for the GI/G/c queue is given by Van Hoorn and Seelen [16], but implementation of this method is not straightforward as is the case with the other approximation methods. The idea of iterating Lindley's integral equation was employed by Ackroyd [1]. He uses signal-processing methods applied to the discretized waiting-time distribution.

The iterative method presented in this paper is based on a simple idea, is easy to implement on a personal computer, and shows an overall excellent performance. A weak point of the iterative method is that the time to compute the approximations is highly dependent on the traffic load. As the traffic load becomes very heavy (i.e., the delay probability gets close to 1), the iterative method converges more slowly. For the case of heavy traffic (say $\rho > 0.95$), we suggest the use of the diffusion approximation in Whitt [17] or the well-known approximation in Krämer and Langenbach-Belz [8].

The paper is organized as follows. In the next section, we describe the basic moment-iteration method. In Section 3, this method is elaborated for the GI/D/1 and GI/G/1 queue with nondeterministic service times, respectively. In Section 4, we present some numerical results and discuss the performance of the approximations. In Section 5, we present another application of the moment-iteration scheme.

2. THE MOMENT-ITERATION METHOD

In this section, we present the basic ideas behind the moment-iteration method. We restrict ourselves to a description of the moment-iteration method based on

the first two moments of the distribution that is approximated. First, we describe the GI/G/1 queue in detail and introduce some notation.

We consider the standard GI/G/1 queue. Customers arrive according to a renewal process, where the interarrival time has probability distribution function $A(t)$. The service time of an arbitrary customer has probability distribution function $B(t)$. The service times are independent of each other and independent of the arrival process.

We assume that at epoch 0 the zeroth customer arrives having a service time B_0. Define for $n \geq 1$,

A_n := the time between the arrival of the $(n-1)$th customer and the nth customer.

B_n := the service time of the nth customer.

W_n := the waiting time of the nth customer.

It is easily seen that

$$W_n = \max(0, W_{n-1} + B_{n-1} - A_n), \qquad n \geq 1. \tag{2.1}$$

Under the assumption that the queue is in a stationary state, this recursive equation can be translated into Lindley's integral equation. Also, this equation is the starting point for our iterative method which approximates the following performance characteristics of the GI/G/1 queue.

$$\pi_w := \lim_{n\to\infty} P\{W_n > 0\}, \qquad E[W] := \lim_{n\to\infty} E[W_n].$$

To assure that π_w and $E[W]$ are properly defined, we assume that

$$\rho := E[B]/E[A] < 1,$$

where the generic random variables A and B are distributed according to $A(t)$ and $B(t)$, respectively.

Through Eq. (2.1), the waiting-time distribution of the nth arriving customer is related to the waiting-time distribution of the $(n-1)$th customer. From this equation, we derive the following expression for $P\{W_n > 0\}$ and the first two moments of W_n:

$$P\{W_n > 0\} = \int_0^\infty [1 - F_{W_{n-1}+B}(t)]\, dA(t), \tag{2.2}$$

$$E[W_n] = \int_0^\infty \int_t^\infty (y-t)\, dF_{W_{n-1}+B}(y)\, dA(t), \tag{2.3}$$

$$E[W_n^2] = \int_0^\infty \int_t^\infty (y-t)^2\, dF_{W_{n-1}+B}(y)\, dA(t) \tag{2.4}$$

One might call the Eqs. (2.2–2.4) low-order versions of Lindley's integral equation. The basic iteration scheme can be given as follows:

2.1. Moment-Iteration Algorithm

Step 0 (Initialization): Choose initial values for $E[W_0]$ and $E[W_0^2]$ (e.g., equal to 0).

Step 1 (Iteration): Compute

$$E[W_{n-1} + B] = E[W_{n-1}] + E[B].$$

$$E[(W_{n-1} + B)^2] = E[W_{n-1}^2] + 2E[W_{n-1}]E[B] + E[B^2].$$

Fit tractable distributions $\tilde{F}_{W_{n-1}+B}$ and $\tilde{A}$ to the probability distributions of the random variables $W_{n-1} + B$ and A by matching the respective first two moments. Compute

$$E[W_n] = \int_0^\infty \int_t^\infty (y - t)\, d\tilde{F}_{W_{n-1}+B}(y)\, d\tilde{A}(t), \tag{2.5}$$

$$E[W_n^2] = \int_0^\infty \int_t^\infty (y - t)^2\, d\tilde{F}_{W_{n-1}+B}(y)\, d\tilde{A}(t). \tag{2.6}$$

Step 2 (Convergence): If the maximum of

$$|E[W_n] - E[W_{n-1}]| < E \quad \text{and} \quad |E[W_n^2] - E[W_{n-1}^2]| < \epsilon,$$

then stop; otherwise repeat Step 1.

Stop: Approximate $E[W]$ by $E[W_n]$, $E[W^2]$ by $E[W_n^2]$ and $P\{W > 0\}$ by

$$P\{W > 0\} \approx \int_0^\infty [1 - \tilde{F}_{W_{n-1}+B}(t)]\, d\tilde{A}(t).$$

In Section 3, we discuss how to fit tractable distributions to the first two moments of $W_{n-1} + B$ and A. For these distributions, the integrals involved are easy to compute. Note that for all $n \geq 1$, $E[W_n]$ and $E[W_n^2]$ are two-moment approximations for the first two moments of the waiting time of the nth arriving customer. Since the stability condition $\rho < 1$ assures that for the true W_n and W we have that $W_n \to W$ in distribution, it may be expected that good approximations for the true $E[W_n]$ and $E[W_n^2]$ will also converge. Numerical experiments yield the following conjecture:

2.2. Convergence Conjecture

The two-moment-iteration algorithm always terminates in a finite number of steps.

Also, numerical experiments reveal that the convergence of $E[W_n]$ and $E[W_n^2]$ is geometrical, which enables us to speed up convergence considerably by extrapolation as follows.

In each step, we compute

$$\tau_n^{(1)} := \frac{E[W_n] - E[W_{n-1}]}{E[W_{n-1}] - E[W_{n-2}]},$$

$$\tau_n^{(2)} := \frac{E[W_n^2] - E[W_{n-1}^2]}{E[W^2] - E[W_{n-2}^2]}.$$

It appears that $\tau_n^{(1)}$ and $\tau_n^{(2)}$ both converge to the same constant τ. Hence, in a limiting sense the convergence of $E[W_n]$ and $E[W_n^2]$ is geometrical. Using this empirical finding, we replace $E[W_n]$ and $E[W_n^2]$ in Step 2 by $E[W_n^g]$ and $E[(W_n^g)^2]$, respectively, which are defined by

$$E[W_n^g] := E[W_{n-1}] + \frac{E[W_n] - E[W_{n-1}]}{1 - \tau_n^{(1)}},$$

$$E[(W_n^g)^2] := E[W_{n-1}^2] + \frac{E[W_n^2] - E[W_{n-1}^2]}{1 - \tau_n^{(2)}}.$$

Modifying the moment-iteration algorithm in this way ensures faster termination of the iteration.

We have not been able to prove this conjecture. The conjecture is supported by the following result. Instead of fitting a distribution using the first two moments of the true distribution, we fit an exponential distribution to the first moment of the true distribution in Step 1 of the iteration algorithm. Then it can be shown that the moment-iteration algorithm terminates in a finite number of steps. In fact, the moment-iteration algorithm solves the following functional equation:

$$x = (x + E[B])\mathrm{LS}_A((x + E[B])^{-1}),$$

where LS_A denotes the Laplace–Stieltjes transform of A. Note that we have assumed that in Step 1 the true interarrival distribution $A(t)$ is used. Next, substituting

$$\mu := E[B] \text{ and } \sigma := x(x + E[B])^{-1},$$

we obtain the following functional equation in the variable σ.

$$\sigma = \mathrm{LS}_A(\mu(1 - \sigma)).$$

This is the well-known functional equation associated with the GI/M/1 queue, which can be solved by iteration and where this iterative scheme converges geometrically. It follows from these arguments that the moment-iteration algorithm yields exact results for the GI/M/1 queue.

The proof of the conjecture should proceed along the same lines as above. The difficulty lies in the degree of freedom one has in choosing the two-moment

fit of the true distributions $F_{W_{n-1}+B}(t)$ and $A(t)$. The underlying functional equation depends on this two-moment fit.

We note that in principle the algorithm can be modified to yield better results through iteration of higher moments and fitting distributions to all these moments. It is to be expected that computations become more intricate, apart from the problem of finding *tractable* distributions that can be fitted. This extra effort should be weighed against the desired accuracy. For an important set of GI/G/1 queues, we have fitted tractable distributions to the first three moments of $W_{n-1} + B$. We shall discuss this three-moment iteration scheme in more detail in the next section.

Finally, we remark that the method can be used to study the time-dependent behavior of the GI/G/1 queue, where time should be interpreted as successive arrival epochs. The algorithm starts off with an initial distribution of the waiting time of the zeroth customer, e.g., the degenerate distribution with probability mass 1 at zero corresponding to the zeroth customer arriving at an empty system. Next the algorithm approximates the first two moments of the waiting time of all arriving customers until stationarity is reached. Hence, the algorithm can be used to gain insight in so-called relaxation times (cf. Blanc and Van Doorn [2]), i.e., the time that elapses until stationarity is reached. It is clear that there is a close relationship between these relaxation times and the parameter τ, which governs the geometric convergence.

In the next section, we give suggestions for implementation of the algorithm for the cases of either deterministic service times or nondeterministic service times.

3. IMPLEMENTATION OF THE ALGORITHM

To implement the algorithm, we should distinguish between two cases. In order to make this distinction, we define c_A^2 and c_B^2 as the squared coefficients of variation of A and B, respectively. We distinguish between the case of $c_B^2 = 0$ and the case of $c_B^2 > 0$. Below we propose which distributions should be fitted to $F_{W_{n-1}+B}(t)$ and $A(t)$ in Step 2.

Case 1 ($c_B^2 > 0$): For the case of nondeterministic service times, we proceed as follows. If the interarrival time is deterministic, we fit the exact distribution to $A(t)$, since this yields tractable results. If $c_A^2 > 0$, then we fit mixtures of Erlang distributions to the first two moments of $W_{n-1} + B$ as explained in Tijms [13]. We also fit mixtures of Erlang distributions to $E[A]$ and $E[A^2]$. Using these mixtures, it is easy to obtain explicit expressions for the right-hand-sides of Eqs. (2.5) and (2.6).

Case 2 ($c_B^2 = 0$): The approach outlined above failed to yield good approximations for the case of $c_B^2 = 0$, i.e., the service times are constant. Therefore, proceed with more care. Let us reconsider Eq. (2.5). For deterministic service times B, this equation can be rewritten as follows:

$$E[W_n] = E[W_{n-1}]A(B) + B - E[A] + \int_B^\infty (t - B)\, dA(t)$$

$$+ P\{W_{n-1} > 0\} \int_B^\infty \int_{t-B}^\infty (y - t + B)\, dF_{W_{n-1}|W_{n-1}>0}(y)\, dA(t),$$

where

$$F_{W_{n-1}|W_{n-1}>0}(y) := P\{W_{n-1} \le y | W_{n-1} > 0\}.$$

The probability distribution function $F_{W_{n-1}|W_{n-1}>0}$ has no probability mass at zero. Now it is reasonable to fit mixtures of Erlang distributions to $E[W_{n-1}|W_{n-1} > 0]$ and $E[W_{n-1}^2|W_{n-1} > 0]$. A similar analysis can be done for Eq. (2.6). Hence, during each iteration step we compute successively approximations for $E[W_n]$, $E[W_n^2]$, $P\{W_n > 0\}$, $E[W_n|W_n > 0]$, and $E[W_n^2|W_n > 0]$.

For all possible cases, we have specified the two-moment iteration algorithm. This leaves us with the question whether the approximations are of practical use. This question is answered in detail in the next section.

It should be emphasized that the performance characteristics of the GI/G/1 queue are sensitive to more than the first two or three moments of the interarrival-time distribution (cf. Tijms [13]). Fortunately, the moment-iteration scheme can easily be adapted to other tractable interarrival time distributions than used above (e.g., uniform distribution, shifted exponential distribution, discrete distributions, etc.).

In the next section, we discuss the quality of the approximations for the delay probability and the expected conditional waiting time $E[W|W > 0]$ obtained with the moment-iteration method. We also give some results concerning approximations for the waiting-time distribution.

4. NUMERICAL RESULTS AND CONCLUSIONS

In this section, we report on our extensive numerical experiments to check the quality and robustness of the approximations obtained. Towards this end, we compared our approximations with the exact results as can be found in Seelen et al. [12] and Tijms [13]. In the latter reference, the embedded Markov chain method is given to compute the exact results for the GI/G/1 queue with phase-type service times.

We consider the following cases. The traffic intensity ρ is varied as 0.2, 0.5, 0.8, and 0.95. The squared coefficient of variation c_A^2 is varied as 0, $\frac{1}{3}$, $\frac{1}{2}$, and 2. The expected service time is normalized at 1, the squared coefficient of variation of the service time c_B^2 runs through the values 0, $\frac{1}{4}$, $\frac{1}{2}$, and $2\frac{1}{2}$. A coefficient of variation equalling 0 corresponds to the deterministic distribution. When the coefficient of variation of the random variable considered equals $\frac{1}{4}$, $\frac{1}{3}$, or $\frac{1}{2}$, we assume that the random variable is distributed according to an Erlang distribution. When the coefficient of variation of the random variable considered equals 2 or $2\frac{1}{2}$, we assumed that the random variable is distributed

TABLE 4.1. The Delay Probability π_W

		$c_B^2 = 0$			$c_B^2 = 1/3$			$c_B^2 = 1/2$			$c_B^2 = 2.5$		
c_A^2	ρ	$\pi_W(2)$	$\pi_W(3)$	$\pi_W(\text{ex})$	$\pi_W(2)$	$\pi_W(3)$	$\pi_W(\text{ex})$	$\pi_W(2)$	$\pi_W(3)$	$\pi_W(\text{ex})$	$\pi_W(2)$	$\pi_W(3)$	$\pi_W(\text{ex})$
0	0.2				0.0000		0.0000	0.0005	0.0005	0.0005	0.0422	0.0413	0.0413
0	0.5				0.0724		0.0726	0.1166	0.1164	0.1164	0.2785	0.2715	0.2716
0	0.8				0.5072		0.4904	0.5645	0.5475	0.5476	0.6411	0.6657	0.6657
0	0.95				0.8786	0.8547	0.8558	0.8938	0.8758	0.8757	0.9023	0.9117	0.9118
1/3	0.2	0.0236		0.0236	0.0373		0.0373	0.0434		0.0434	0.0884	0.0813	0.0813
1/3	0.5	0.2252		0.2253	0.2783		0.2757	0.2960	0.2933	0.2933	0.3603	0.3590	0.3590
1/3	0.8	0.6329	0.6324	0.6325	0.6826		0.6735	0.6964	0.6861	0.6861	0.7162	0.7251	0.7251
1/3	0.95	0.9021	0.9017	0.9019	0.9226	0.9140	0.9146	0.9243	0.9185	0.9184	0.9259	0.9294	0.9294
1/2	0.2	0.0645		0.0645	0.0764		0.0763	0.0815		0.0813	0.1144	0.1089	0.1089
1/2	0.5	0.3233		0.3233	0.3522		0.3496	0.3624	0.3596	0.3596	0.3997	0.3993	0.3994
1/2	0.8	0.7022	0.7018	0.7019	0.7266	0.7189	0.7198	0.7328	0.7262	0.7262	0.7430	0.7483	0.7483
1/2	0.95	0.9229	0.9226	0.9227	0.9324	0.9277	0.9280	0.9331	0.9299	0.9298	0.9339	0.9359	0.9354
2	0.2	0.2610		0.2610	0.2594		0.2594	0.2586	0.2587	0.2587	0.2565	0.2519	0.2519
2	0.5	0.6137	0.6139	0.6140	0.6037	0.6050	0.6050	0.5998	0.6012	0.6012	0.5788	0.5779	0.5779
2	0.8	0.8754	0.8768	0.8768	0.8621	0.8682	0.8682	0.8600	0.8646	0.8647	0.8565	0.8471	0.8871
2	0.95	0.9714	0.9720	0.9720	0.9665	0.9694	0.9695	0.9663	0.9684	0.9685	0.9661	0.9636	0.9635

according to a hyperexponential distribution with balanced means (cf. Tijms [13]).

For these sets of parameters, we computed the delay probability π_w for which the results are tabulated in Table 4.1. For all cases, we computed the two-moment-approximation $\pi_w(2)$. If possible, we also computed a three-moment approximation based on a three-moment iteration method. This three-moment iteration method employs the property that if the squared coefficient of variation of the approximated random variable $W_n + B$ (or $W_n | W_n > 0$ in the case of the G/D/1 queue) exceeds $\frac{1}{2}$, we may be able to fit a so-called K_2-distribution to the first three moments of $W_n + B$ $(W_n | W_n > 0)$. For more details, we refer to Tijms [13]. The three-moment approximation is denoted by $\pi_w(3)$. The exact value of the delay probability is denoted by $\pi_w(\text{ex})$.

In Table 4.2, we display the computed approximations and the exact values for the expected conditional waiting time $E[W | W > 0]$. Similarly, as for the delay probability π_w, we denote the two-moment approximation by $W(2)$, the three-moment approximation by $W(3)$, and the exact value of the expected conditional waiting time by $W(\text{ex})$.

The results in Tables 4.1 and 4.2 show the satisfactory performance of the two-moment approximation. Only when c_B^2 gets large, say larger than 2, the results should be used with care. The reason is simply that the performance measures become increasingly sensitive to more than the first two moments of the service times when c_B^2 gets large (cf. Tijms [13]). We note that our two-moment approximations for $P\{W > 0\}$ and $E[W | W > 0]$ turned out to be exact for the case of the GI/M/1 queue and the M/G/1 queue.

The three-moment approximations yield excellent results. However, in contrast with the two-moment iteration scheme, the three-moment iteration scheme does not always converge. This is mostly caused by the fact that a three-moment fit is not always possible. Typically, if a three-moment iteration is not possible then the algorithm starts cycling in some sense. The implementation of the algorithm detects whether cycling occurs; and if so, then a two-moment iteration scheme is used.

The computations are, as said before, simple but the number of iterations involved is highly dependent on the traffic load: the number of iterations increases as ρ increases. To a lesser extent, the number of iterations depends on the coefficients of variation c_A^2 and c_B^2: the number of iterations increases with increasing c_A^2 or c_B^2.

Since, from Eq. (2.2),

$$P[W_n > y] = \int_0^\infty [1 - F_{W_{n-1}+B}(t + y)]\, dA(t),$$

the same approach, as used to compute an approximation for the delay probability $P\{W > 0\}$, can be applied to compute two-moment and three-moment approximations to the waiting-time distribution.

TABLE 4.2. The Expected Conditional Waiting Time

		$c_B^2 = 0$			$c_B^2 = 1/3$			$c_B^2 = 1/2$			$c_B^2 = 2.5$		
c_A^2	ρ	$W(2)$	$W(3)$	$W(ex)$	$W(2)$	$W(3)$	$W(ex)$	$W(2)$	$W(3)$	$W(ex)$	$W(2)$	$W(3)$	$W(ex)$
0	0.2				0.377		0.377	0.546	0.546	0.546	2.182	3.004	3.008
0	0.5				0.463		0.471	0.660	0.673	0.673	3.454	3.561	3.560
0	0.8				0.878		0.939	1.308	1.382	1.382	7.309	7.013	7.013
0	0.95				3.283	3.446	3.427	4.964	5.117	5.118	25.96	25.67	25.67
1/3	0.2	0.279		0.279	0.544		0.547	0.675		0.678	2.278	2.648	2.648
1/3	0.5	0.387		0.387	0.745		0.762	0.928	0.947	0.947	3.383	3.398	3.397
1/3	0.8	0.875	0.876	0.876	1.702	1.702	1.742	2.121	2.173	2.173	7.614	7.505	7.505
1/3	0.95	3.369	3.371	3.371	6.641	6.743	6.735	8.334	8.414	8.415	28.83	28.71	28.71
1/2	0.2	0.376		0.376	0.618		0.621	0.739		0.742	2.261	2.454	2.454
1/2	0.5	0.547		0.547	0.893		0.907	1.070	1.086	1.086	3.378	3.382	3.382
1/2	0.8	1.286	1.287	1.287	2.108	2.147	2.142	2.531	2.568	2.568	7.858	7.792	7.793
1/2	0.95	5.031	5.033	5.033	8.322	8.391	8.387	10.01	10.06	10.06	30.34	30.26	30.26
2	0.2	0.681		0.681	0.905		0.904	1.016	1.015	1.015	2.307	2.288	2.288
2	0.5	1.332	1.330	1.330	1.730	1.713	1.713	1.919	1.901	1.901	3.934	3.950	3.950
2	0.8	4.333	4.311	4.311	5.319	5.214	5.215	5.742	5.659	5.659	10.57	10.69	10.69
2	0.95	19.38	19.35	19.35	22.91	22.74	22.74	24.56	24.44	24.43	44.27	44.44	44.45

From the algorithm, one can compute two-moment and three-moment approximations for the conditional waiting-time distribution. From this, we compute the α-percentiles of the conditional waiting-time distribution, where α runs through the values 0.5, 0.8, 0.9, 0.95, and 0.99. We consider deterministic, Erlang-2, and hyperexponential distributed interarrival times, where in the latter case we choose c_A^2 equal to 2 and the first three moments of the interarrival-time distribution according to the gamma distribution (cf. Tijms [13]). For the service-time distribution, we choose an Erlang-4 distribution instead of the deterministic distribution. The traffic intensity is chosen equal to 0.5. Table 4.3 presents the exact and approximate α-percentiles. The two-moment and

TABLE 4.3. Exact and Approximate Values of the Conditional Waiting-Time Percentiles

		α-Percentiles				
	α	0.5	0.8	0.9	0.95	0.99
$D/E_4/1$	exact	0.26	0.59	0.83	1.07	1.61
	app(2)	0.26	0.58	0.82	1.05	1.58
	app(3)					
$D/E_2/1$	exact	0.48	1.09	1.54	1.99	3.01
	app(2)	0.47	1.07	1.51	1.94	2.92
	app(3)	0.48	1.09	1.54	1.99	3.01
$D/H_2/1$	exact	1.66	3.86	5.52	7.19	11.05
	app(2)	1.79	4.27	6.15	8.03	12.39
	app(3)	1.66	3.86	5.53	7.19	11.05
$E_2/E_4/1$	exact	0.62	1.29	1.77	2.26	3.37
	app(2)	0.64	1.29	1.73	2.14	3.05
	app(3)					
$E_2/E_2/1$	exact	0.79	1.73	2.43	3.13	4.74
	app(2)	0.80	1.72	2.36	2.99	4.39
	app(3)	0.79	1.73	2.43	3.13	4.74
$E_2/H_2/1$	exact	1.83	4.39	6.33	8.26	12.76
	app(2)	1.72	4.57	6.77	8.97	14.09
	app(3)	1.83	4.39	6.33	8.27	12.77
$H_2/E_4/1$	exact	1.44	3.09	4.34	5.59	8.48
	app(2)	1.59	3.18	4.28	5.33	7.66
	app(3)	1.44	3.09	4.33	5.57	8.45
$H_2/E_2/1$	exact	1.60	3.54	5.00	6.46	9.86
	app(2)	1.73	3.63	4.95	6.23	9.07
	app(3)	1.60	3.53	4.99	6.46	9.85
$H_2/H_2/1$	exact	2.46	6.07	8.80	11.54	17.88
	app(2)	2.14	5.94	8.99	12.05	19.16
	app(3)	2.46	6.07	8.80	11.54	17.89

three-moment approximations for the α-percentiles are denoted by app(2) and app(3), respectively. The exact results have been taken from Tijms [13]. Again, the two-moment approximations are satisfactory and the three-moment approximations are excellent.

It is well-known that the waiting-time characteristics of the GI/G/1 queue are highly dependent on the form of the arrival-time distribution. Especially for small values of ρ, it appears that the first two moments of the interarrival time do not characterize the interarrival-time distribution sufficiently to obtain accurate results if the fitted distribution is not the actual distribution. The moment-iteration scheme can be easily adapted to other tractable distributions of the interarrival times, such as finite discrete distributions, uniform and shifted exponential distributions and all finite mixtures of Erlangian distributions. This provides a means to do an extensive sensitivity analysis of the GI/G/1 queue. This also gives the opportunity to analyze GI/G/1 queues different from Ph/Ph/1 queues. Table 4.4 compares approximations for the $U/E_k/1$ queue with uniform interarrival times and Erlangian service times with exact results obtained by Tijms [15]. It is assumed that ρ equal 0.5.

The results of Table 4.4 show again the excellent performance of the moment-iteration method. It should be noted that the results for the $U/E_2/1$ queue are obtained from a three-moment iteration, and the results for the $U/E_4/1$ and $U/E_3/1$ queues are obtained from a two-moment-iteration.

In conclusion, the approximations in this paper are of good quality, robust, and easy to implement. Computation times can become prohibitively long on personal computers when ρ exceeds 0.95. Fortunately, in that case the heavy traffic approximations for π_w and $E[W|W>0]$ of Whitt [17] can be used. The algorithm can give insight in the time-dependent behavior of the GI/G/1 queue. The algorithm also provides great flexibility with respect to the choice of the interarrival-time distribution. The basic idea can be applied to other classes of

Table 4.4. Performance of the Moment-Iteration Method for the $U/E_k/1$ Queue

				α-Percentiles Conditional Waiting Time				
		$E[W_q]$	$P\{W_q=0\}$	0.50	0.80	0.90	0.95	0.99
$U/E_4/1$	exact	0.278	0.6822	0.681	1.367	1.862	2.355	3.501
	approx.	0.274	0.6822	0.692	1.371	1.822	2.245	3.164
$U/E_3/1$	exact	0.306	0.6764	0.719	1.489	2.047	2.603	3.892
	approx.	0.312	0.6761	0.731	1.488	1.999	2.483	3.541
$U/E_2/1$	exact	0.365	0.6655	0.803	1.737	2.425	3.109	4.693
	approx.	0.365	0.6655	0.803	1.737	2.425	3.108	4.692

problems. Further research has been conducted on GI/G/1 queues with finite waiting time in De Kok [4]. In the next section, we describe an (s,S)-production-inventory model where lead times depend on the production capacity.

5. APPLICATION OF THE MOMENT-ITERATION METHOD TO AN (s,S)-PRODUCTION-INVENTORY MODEL

In this section, we show that the moment-iteration method is applicable to other problems, in addition to the GI/G/1 queue. We consider a production facility with a warehouse. Customers arrive at the warehouse according to a Poisson process and the demand per customer has some arbitrary distribution. The demand of a customer is satisfied from stock on hand, otherwise it is backlogged. The inventory is replenished according to an (s,S)-rule. As soon as an order from the warehouse arrives at the production facility, production on behalf of this order starts, unless the production facility is still busy producing another order. In that case, the production of the new order starts as soon as production of the preceding order has finished. The production time depends on the production rate π and the order size Q. We assume that the production time equals Q/π. Hence, the total production lead time equals Q/π plus some waiting time until the preceding production order has finished. The goal of the manager of the production facility is to determine an (s,S)-rule, such that the fraction of demand delivered directly from stock on hand equals some target value β.

It is important to note that the above-described model explicitly incorporates the finite production capacity of the production facility. In the literature on (s,S)-models, one implicitly assumes that the production capacity of the supplier is infinite. The only way to account for the finite production capacity in these models is to overestimate the production lead time. The above model shows the interaction between subsequent orders.

The analysis is based on the derivation of an equation that describes the relation between subsequent production lead times (cf. Lindley's integral equation for the GI/G/1 queue). Using the moment-iteration scheme, this equation is approximately solved for the first two moments of the "effective" production lead time. Next, we compute the (s,S)-rule that yields the target service level for the "effective" production lead time (e.g., using the results from Tijms and Groenevelt [14]). It is shown in De Kok [5] that the analysis yields excellent results. It is also shown that (s,S)-rules determined under the assumption of infinite production capacity yield poor results in terms of service levels. In Table 5.1, we show some results obtained in De Kok [5] to show the accuracy of the moment-iteration method. We compare the performance in terms of service level of the standard (s,S)-rule, which implicitly assumes infinite production capacity, with the finite capacity (s,S)-rule, obtained from the application of the moment-iteration method. For the infinite capacity (s,S)-model, we assumed that the lead time equals $(S-s+U)/\pi$, where U denotes the undershoot of the order level s.

Table 5.1. The Infinite Capacity (s,S)-Rule Versus the Finite Capacity (s,S)-Rule

		$c_D^2 = 1/3$		$c_D^2 = 2/3$	
π	β	β_{inf}	β_{fin}	β_{inf}	β_{fin}
1.25	0.95	0.81	0.94	0.80	0.95
1.5	0.95	0.90	0.95	0.90	0.96
2	0.95	0.94	0.96	0.94	0.96
1.25	0.99	0.90	0.98	0.89	0.99
1.25	0.95	0.97	0.99	0.96	0.99
2	0.99	0.99	0.99	0.98	0.99

In Table 5.1, we varied π as 1.25 and 2, β as 0.95 and 0.99, and the variation coefficient of demand, c_D^2, as $\frac{1}{3}$ and $\frac{2}{3}$. The customers' arrival rate equals 1 and the expected demand per customer equals 1. The difference $S - s$ has been computed from the EOQ formula, assuming a fixed-order cost equal to 25 and a holding cost equal to 1. For both the infinite capacity (s,S)-rule and the finite capacity (s,S)-rule, we have given the actual service levels obtained from computer simulation.

Remark: The GI/G/1 queue arises in the context of another important finite-capacity production-inventory model. Consider a periodic review (R,S) system. At the beginning of each review period, the economic stock is replenished up to S. However, the maximum production capacity in a period of length R equals Q. Hence, the maximum replenishment equals Q. The production lead time equals L. The demand per review period is distributed according to some arbitrary probability distribution. Then it can be seen that the economic-stock process can be translated into a D/G/1 queue.

Acknowledgments

I wish to thank Professor Henk Tijms of the Free University, Amsterdam, for his careful reading of the manuscript and for providing additional exact results. I also thank Jan van Doremalen for his useful comments.

References

1. Ackroyd, M.H. (1980). Computing the waiting-time distribution for the GI/G/1 queue by signal-processing methods. *IEEE Transactions on Communication* 28: 52–58.
2. Blanc, J.P.C. & Van Doorn, E.Q. (1986). Relaxation times for queueing systems. In *Proceedings of the CWI Symposium on Mathematics and Computer Science.* CWI Monograph 1, North-Holland, pp. 139–162.
3. Bux, W. (1979). Single-server queues with general interarrival and phase-type service times distributions. In *Proceedings of the 9th International Teletraffic Congress*, Torremolinos, Paper 413.
4. De Kok, A.G. (1987). Approximations for the waiting-time characteristics of the GI/G/1 queue with impatience. CQM-note Centre for Quantitative Methods, Philips, Eindhoven, The Netherlands.
5. De Kok, A.G. (1987). Computing optimal (m,M) control rules for production-inventory models with compound Poisson demand. CQM-note 55, paper presented at EURO VIII, 1986, Lisbon, Portugal.
6. Fredericks, A.A. (1982). A class of approximations for the waiting-time distribution in a GI/G/1 queueing system. *Bell System Technical Journal* 61: 295–325.
7. Kleinrock, L. (1975). *Queueing systems, Vol. 1.* New York: Wiley.
8. Krämer, W. & Langenbach-Belz, M. (1976). Approximate formulae for the delay in the queueing system GI/G/1. In *Proceedings of the 8th International Teletraffic Congress*, Melbourne, 235-1/8.
9. Neuts, M.F. (1981). *Matrix-geometric solutions in stochastic models—an algorithmic approach.* Baltimore, Maryland: The John Hopkins University Press.
10. Seelen, L.P. (1986). An algorithm for Ph/Ph/c queues. *European Journal of Operations Research* 23: 118–127.
11. Seelen, L.P. & Tijms, H.C. (1984). Approximations for the conditional waiting times in the GI/G/c queue. *Operations Research Letters* 3: 183–190.
12. Seelen, L.P., Tijms, H.C., & Van Hoorn, M.H. (1985). *Tables for multiserver queues.* Amsterdam: North-Holland.
13. Tijms, H.C. (1986). *Stochastic modelling and analysis: A computational approach.* New York: Wiley.
14. Tijms, H.C. & Groenevelt, H. (1984). Simple approximations for the reorder point in periodic and continuous review (s,S) inventory systems with service-level constraints. *European Journal of Operations Research* 17: 175–190.
15. Tijms, H.C. (1987). Private communication.
16. Van Hoorn, M.H. & Seelen, L.P. (1986). Approximations for the GI/G/c queue. *Journal of Applied Probabilities* 23: 484–494.
17. Whitt, W. (1982). Refining diffusion approximations for queues. *Operations Research Letters* 5: 165–169.