SIMPLE APPROXIMATIONS FOR THE BATCH-ARRIVAL $M^x/G/1$ QUEUE

J. C. W. VAN OMMEREN

University Twente, Enschede, The Netherlands

(Received February 1988; revisions received November 1988, April 1989; accepted May 1989)

In this paper we consider the $M^x/G/1$ queueing system with batch arrivals. We give simple approximations for the waiting-time probabilities of individual customers. These approximations are checked numerically and they are found to perform very well for a wide variety of batch-size and service-time distributions.

Batch-arrival queueing models can be used in many practical situations, such as the analysis of message packetization in data communication systems. In general it is difficult, if not impossible, to find tractable expressions for the waiting-time probabilities of individual customers. It is, therefore, useful to have easily computable approximations for these probabilities. This paper gives such approximations for the single server $M^x/G/1$ model.

Exact methods for the computation of the waiting-time distribution in the $M^x/G/1$ queue are discussed in Eikeboom and Tijms (1987), cf. also Chaudhry and Templeton (1983), Neuts (1981) and Tijms (1986). However, these methods apply only for special service-time distributions and are, in general, not suited for routine calculations in practice. A simple approximation for the tail probabilities of the waiting time was given in Eikeboom and Tijms by using interpolation of the asymptotic expansions for the particular cases of deterministic and exponential services. This approximate approach uses only the first two moments of the service time.

This paper presents an alternative approach that uses the actual service-time distribution rather than just its first two moments. This alternative approach starts with the asymptotic expansion of the waiting-time distribution. In Van Ommeren (1988) it is shown that the complementary waiting-time probability of an arbitrary customer in the $M^x/G/1$ queue has an exponentially fast decreasing tail under some mild assumptions. By calculating the decay parameter and the amplitude factor, we get the asymptotic expansion of the waiting-time distribution. Such asymptotic expansions already provide a very powerful tool in practical queueing analysis, cf. Cromie, Chaudhry and Grassman (1979) and Tijms. It turns out that for nonlight traffic this asymptotic expansion can be used as a good first-order approximation for the waiting-time probabilities. Next, by incorporating exact results for other quantities, such as the delay probability and the first two moments of the waiting time, we are able to give an improved second-order approximation. This approximate method performs very well for a wide range of values of the traffic intensity and the coefficients of variation of the service-time distribution and the batch-size distribution.

The organization of this paper is as follows. In Section 1, the model is defined and some preliminaries, including the asymptotic expansion of the waiting-time distribution, are given. The second-order approximation is given in Section 2. In Section 3 we give numerical results and discuss the performance of the approximations. The Appendix deals with the motivation of the second-order approximation.

1. THE MODEL AND PRELIMINARIES

In the $M^x/G/1$ queue, customers arrive in batches and are served individually by a single server. The batches arrive according to a Poisson process with rate λ. The number of customers in the batches are independent and identically distributed positive random variables. Denote the number of customers in a batch by X and the probability distribution of X by $\{g_i := \Pr\{X = i\}, i = 1, 2, \ldots\}$. The generating function of $\{g_i\}$ is denoted by $G(z) := \sum_{j=1}^{\infty} g_j z^j$. The service times of individual customers are independent identically distributed random variables. Denote the service time of a customer by S and the distribution of S by $B(x) := \Pr\{S \leq x\}$. We assume that $B(0) = 0$ and $B'(0) := \lim_{t\downarrow 0} B(t)/t$ exists. We denote the Laplace–Stieltjes transform of $B(\cdot)$ by $\hat{B}(s) := \int_0^{\infty} e^{-st}\, dB(t)$. Let $\hat{Q}(s)$

Subject classifications: Queues, approximations: approximations based on asymptotic analysis. Queues, batch/bulk: approximations.

0030-364X/90/3804-0678 $01.25

denote the Laplace–Stieltjes transform of the total amount of service time required by all customers belonging to one batch. It follows that

$$\hat{Q}(s) = G(\hat{B}(s)).$$

The offered load is denoted by $\rho := \lambda E(S)E(X)$ and it is assumed that $\rho < 1$.

Customers belonging to different batches are served in order of arrival, while customers belonging to the same batch are served according to their random position in the batch. Let D_n denote the delay in the queue of the nth served customer. The limit distribution $\lim_{n\to\infty}\Pr\{D_n \leq x\}$ exists only when the batch-size distribution $\{g_i\}$ is aperiodic (i.e., when the g.c.d. $\{j \mid g_j > 0\} = 1$), cf. Cohen (1976). As a counterexample, consider a constant batch size of 2 in which case $\Pr\{D_{2k} = 0\} = 0$ for all $k \geq 1$ and $\lim_{k\to\infty}\Pr\{D_{2k+1} = 0\} = (1 - \rho)$. In Van Ommeren it is proved that the limit

$$W_q(t) := \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \Pr\{D_j \leq t\}, \quad t \geq 0$$

always exists. Note that $W_q(t)$ represents with probability 1 the long-run fraction of customers having a waiting time of no more than t. Denote the Laplace–Stieltjes transform of $W_q(\cdot)$ by

$$\hat{W}_q(s) := \int_0^\infty e^{-st}\, dW_q(t).$$

From Cohen we can easily get the following theorem.

Theorem 1. *The Laplace–Stieltjes transform of the stationary waiting-time distribution $W_q(\cdot)$ is given by $\hat{W}_q(s) = \hat{W}_1(s)\hat{W}_2(s)$ with*

$$\hat{W}_1(s) = \frac{(1-\rho)s}{s - \lambda(1 - \hat{Q}(s))}$$

and

$$\hat{W}_2(s) = \frac{1 - \hat{Q}(s)}{E(X)(1 - \hat{B}(s))}.$$

Burke (1975) and Cooper (1981, pages 241–243), use another, more intuitive, probabilistic approach. Following their approach, we find that $\hat{W}_1(\cdot)$ is the Laplace–Stieltjes transform of the stationary waiting-time distribution of the first customer in a batch. $\hat{W}_2(\cdot)$ represents the Laplace–Stieltjes transform of the distribution of the waiting time of an arbitrary customer caused by customers who arrived in the same batch and are served before him. From $\hat{W}_q(\cdot)$, the Laplace–Stieltjes transform of the stationary waiting-time distribution, we can easily derive exact results for the delay probability, the derivative of $W_q(x)$ at $x = 0$ and the first two moments of the waiting time. The following results will be needed in order to obtain a second-order approximation to $W_q(\cdot)$

$$W_q(0) = \lim_{s\to\infty} \hat{W}_q(s) = \frac{1-\rho}{E(X)} \tag{1}$$

$$W_q'(0) = \lim_{s\to\infty} s(\hat{W}_q(s) - W_q(0)) = \frac{(1-\rho)[(1-g_1)B'(0) + \lambda]}{E(X)} \tag{2}$$

$$\int_0^\infty (1 - W_q(t))\, dt = -\hat{W}_q'(0) = -\hat{W}_1'(0) - \hat{W}_2'(0) \tag{3}$$

and

$$\int_0^\infty t(1 - W_q(t))\, dt = \frac{1}{2}\hat{W}_1''(0) + \hat{W}_1'(0)\hat{W}_2'(0) + \frac{1}{2}\hat{W}_2''(0). \tag{4}$$

Here the derivatives should be interpreted as the right derivatives in $t = 0$. In order to evaluate the right-hand sides of (3) and (4) we have to use l'Hôpital's rule repeatedly to obtain

$$\hat{W}_1'(0) = \frac{\lambda\hat{Q}''(0)}{2(1-\rho)}$$

$$\hat{W}_1''(0) = \frac{-\lambda\hat{Q}'''(0)}{3(1-\rho)} + 2[\hat{W}_1'(0)]^2$$

$$\hat{W}_2'(0) = \frac{\hat{Q}''(0) - E(X)\hat{B}''(0)}{2E(X)\hat{B}'(0)}$$

and

$$\hat{W}_2''(0) = \frac{2[\hat{Q}'''(0)\hat{B}'(0) - \hat{Q}'(0)\hat{B}'''(0)]\hat{B}'(0) - 3[\hat{Q}''(0)\hat{B}'(0) - \hat{Q}'(0)\hat{B}''(0)]\hat{B}''(0)}{6E(X)(\hat{B}'(0))^3}$$

where

$$\hat{B}'(0) = -E(S), \quad \hat{B}''(0) = E(S^2)$$

$$\hat{B}'''(0) = -E(S^3)$$

$$\hat{Q}'(0) = -E(S)E(X)$$

$$\hat{Q}''(0) = E(X(X-1))E(S)^2 + E(X)E(S^2)$$

and

$$\hat{Q}'''(0) = -E(X(X-1)(X-2))E(S)^3 - 3E(X(X-1))E(S)E(S^2) - E(X)E(S^3).$$

Here, it is assumed that the service time and the batch size have finite third moments.

To give the asymptotically exponential expansion of the complementary waiting-time distribution $1 - W_q(t)$ for $t \to \infty$, we need a mild assumption on the service-time and batch-size distributions. Essentially, we have to require that both the service-time and the batch-size distributions do not have an extremely long tail. This can be stated precisely as follows.

Assumption

a. The radius of convergence R of the power series $G(z) = \sum_{j=1}^{\infty} g_j z^j$ is larger than one and the integral $\hat{B}(s) = \int_0^\infty e^{-st}\, dB(t)$ is finite for some $s < 0$.

b. $\mathrm{Lim}_{s \downarrow T} \hat{B}(s) = R$

where

$$T = \inf\left\{ s \,\middle|\, \int_0^\infty e^{-st}\, dB(t) < R \right\}.$$

c. $\mathrm{Lim}_{z \uparrow R} [G(z)]^{-1} = 0.$

This assumption is satisfied in most cases of practical interest. It is always satisfied if the service-time distribution is of phase type and if the batch-size distribution has finite support. Assumption c excludes batch-size distributions of the form $\tau n^{-2} p^n$ though.

Theorem 2. *The stationary waiting-time distribution $W_q(\cdot)$ satisfies*

$$\lim_{t \to \infty} e^{\beta t}(1 - W_q(t)) = \alpha, \tag{5}$$

where β is the smallest positive solution to

$$\lambda(\hat{Q}(-\beta) - 1) = \beta$$

and α is defined by

$$\alpha := \frac{(1-\rho)\beta}{\lambda E(X)(\lambda \hat{Q}'(-\beta) + 1)(1 - \hat{B}(-\beta))}.$$

This theorem follows as an immediate corollary of the general results for the $GI^X/G/1$ queue in Van Ommeren. It can also be derived from Theorem 12 in Gaver (1959), where an asymptotic expansion is given for the virtual waiting-time distribution in the $M^X/G/1$ queue.

2. APPROXIMATIONS

The asymptotic expansion stated in Theorem 2 suggests the following first-order approximation to $W_q(t)$

$$1 - W_q(t) \approx \alpha e^{-\beta t} \quad \text{for } t \text{ large enough.}$$

This approximation gives practically useful results for moderate values of t as long as the traffic load is not too low. The performance of this first-order approximation improves as ρ increases. As a rule of thumb, in terms of the pth percentile of the waiting-time distribution function $W_q(t)$, the first-order approximation can also be used for practical purposes when $p \geqslant 1 - \rho(1 - W_q(0))$, cf. also the numerical results in Section 3.

A refinement of the first-order approximation for the complementary waiting-time distribution $1 - W_q(t)$ that can be used for all values of t is given by

$$\bar{W}_{\mathrm{app}}(t) = \alpha e^{-\beta t} + \gamma e^{-\delta t} + \eta e^{-\varphi t}, \quad t \geqslant 0. \tag{6}$$

Here α and β are the coefficients of the asymptotic expansion (5). The motivation of an approximation with three (rather than two) exponential terms is as follows. A close look at the derivation of the asymptotic expansion reveals that the behavior of the waiting-time distribution is determined by the poles of the Laplace–Stieltjes transform of this waiting-time distribution and the residues at these points. The pole with the largest negative real part is simple and real and gives a first-order approximation. The poles that have the second largest negative real parts lead to a second-order approximation. However, it is difficult to find these poles because they no longer have to be real. We, therefore, try to determine γ, δ, η and φ by matching the exact explicit results for the delay probability, the derivative of $W_q(x)$ at $x = 0$ and the first two moments of the stationary waiting time. This yields the relations

$$\bar{W}_{\mathrm{app}}(0) = 1 - W_q(0) \tag{7}$$

$$\bar{W}'_{\mathrm{app}}(0) = -W'_q(0) \tag{8}$$

$$\int_0^\infty \bar{W}_{\mathrm{app}}(t)\, dt = \int_0^\infty (1 - W_q(t))\, dt \tag{9}$$

and

$$\int_0^\infty t\bar{W}_{\mathrm{app}}(t)\, dt = \int_0^\infty t(1 - W_q(t))\, dt. \tag{10}$$

The numbers γ, δ, η and φ may or may not be real. If they are not real, they are complex conjugates, that is, $\delta = \bar{\varphi}$ and $\gamma = \bar{\eta}$, and we find after some algebra that (6) involves a cosine term, cf. Remark 1 in the

Appendix. Furthermore, we require that both $\mathrm{Re}(\delta) > \beta$ and $\mathrm{Re}(\varphi) > \beta$ hold. Otherwise, the asymptotically exponential expansion would be violated for t large. If the poles are complex, it is preferable to approximate $1 - W_q(t)$ by three exponential terms rather than by two exponential terms. Moreover, numerical investigations indicate that for small values of t, the three-term approximation to $1 - W_q(t)$ usually performs much better than the two-term approximation.

As further support for the approximation (6), note that the approximation is exact when the service time as a K_2-distribution and the batch size is geometrically distributed (cf. Van Ommeren). If the service time has a K_3-distribution and the batch size is geometrically distributed the approximation is also exact provided that the function $\lambda(\hat{Q}(-s) - 1) - s$ has no double zero, which is usually the case.

In order to give closed form expressions for γ, δ, η and φ, we define the constants

$$C_1 := (1 - W_q(0)) - \alpha$$

$$C_2 := W_q'(0) - \alpha\beta$$

$$C_3 := \int_0^\infty (1 - W_q(t))\,dt - \alpha/\beta \tag{11}$$

$$C_4 := \int_0^\infty t(1 - W_q(t))\,dt - \alpha/\beta^2$$

and

$$\Delta := (C_1C_3 - C_2C_4)^2 - 4(C_3^2 - C_4C_1)(C_1^2 - C_2C_3).$$

Note that the C_i''s represent the deviations of the *moments* of $1 - W_q(x)$ and the *moments* of the first-order approximation $\alpha e^{-\beta t}$. We can give a simple scheme for computation of the numbers γ, δ, η and φ. For clarity of presentation, we give only the results and refer to the Appendix for their derivation. Let β_0 denote some constant with $\beta_0 > \beta$, e.g., $\beta_0 = 2\beta$. We have to distinguish between the following cases.

Case i. $C_1 = C_2 = C_3 = C_4 = 0$. In this case, the first-order approximation already matches the four proposed conditions (7)–(10), and we set $\gamma = \eta = 0$ and $\delta = \varphi = \beta_0$.

Case ii. $C_3 \neq 0$, $C_1/C_3 > \beta$, $C_1^2 = C_2C_3$ and $C_3^2 = C_1C_4$. In this case, only two exponential terms are needed in the approximation (6) and we set $\gamma = C_1$, $\delta = C_1/C_3$, $\eta = 0$ and $\varphi = \beta_0$.

Case iii. $C_3^2 \neq C_1C_4$, $(C_1C_3 - C_2C_4)/2(C_3^2 - C_4C_1) > \beta$ and $0 < \Delta < [(C_1C_3 - C_2C_4) - 2(C_3^2 - C_4C_1)\beta]^2$. We then get three exponential terms in the approximation (6) where the numbers γ, δ, η and φ are all real and are given by

$$\delta = \frac{C_1C_3 - C_2C_4 + \sqrt{\Delta}}{2(C_3^2 - C_4C_1)}$$

$$\varphi = \frac{C_1C_3 - C_2C_4 - \sqrt{\Delta}}{2(C_3^2 - C_4C_1)}$$

$$\gamma = \frac{C_2 - \varphi C_1}{\delta - \varphi} \quad \text{and} \quad \eta = \frac{C_2 - \delta C_1}{\varphi - \delta}.$$

Case iv. $C_3^2 \neq C_1C_4$, $(C_1C_3 - C_2C_4)/2(C_3^2 - C_4C_1) > \beta$ and $\Delta < 0$. In this case, the numbers γ, δ, η and φ are complex and $\gamma = \bar{\eta}$ and $\delta = \bar{\varphi}$. This gives the following representation for the approximation (6)

$$\bar{W}_{\mathrm{app}}(t) = \alpha e^{-\beta t} + \gamma^* \cos(\varphi^* t + \psi^*) e^{-\delta^* t} \tag{12}$$

where $\delta^* = \mathrm{Re}(\delta)$, $\varphi^* = \mathrm{Im}(\delta)$ and γ^* are given by

$$\delta^* = \frac{C_1C_3 - C_2C_4}{2(C_3^2 - C_4C_1)}$$

$$\varphi^* = \frac{\sqrt{-\Delta}}{2(C_3^2 - C_4C_1)} \tag{13}$$

$$\gamma^* = \sqrt{C_1^2 + ((C_2 - \delta^* C_1)/\varphi^*)^2}$$

and ψ^* is defined by $\cos(\psi^*) = C_1/\gamma^*$ and $\sin(\psi^*) = (\delta^* C_1 - C_2)/(\gamma^* \varphi^*)$.

Remark 1. In most applications, the constants C_1, C_2, C_3 and C_4 match one of these four cases. In case they do not, we propose to try approximation (6) with two exponential terms (i.e., $\eta = 0$), where the approximation matches the exact results for the delay probability and the first moment of the waiting time. This approximation is given by

$$\bar{W}_{\mathrm{app}}(t) = \alpha e^{-\beta t} + C_1 e^{-(C_1/C_3)t}$$

and should, of course, satisfy the requirement that $C_3 \neq 0$ and $\beta < C_1/C_3$. If this requirement is not satisfied, we propose to use the asymptotically exponential expansion. However, in our numerical investigations, this contingency never arose.

Remark 2. For the $M^x/D/1$ queue with deterministic services it may be hazardous to use the above approximations, particularly when the traffic load is low. Due to batches that arrive in an empty system, the waiting-time distribution has a positive mass at nD when

$$\sum_{k=n+1}^{\infty} g_k \neq 0.$$

The effect of this phenomenon is considerable when the traffic load is low, but becomes less important for high traffic when most batches arrive in a nonempty system. It was shown in Eikeboom and Tijms that the total mass of the waiting-time distribution at the discrete points nD with $n \geqslant 1$ equals $(1-\rho)(1-1/E(X))$. Since the approximations given in this paper are all continuous, it cannot be expected that they perform well in the case of deterministic services and low traffic. In this particular case we, therefore, suggest to use the approximate method given in Eikeboom and Tijms. Next, assume that the service-time distribution has a similar shape as the discount distribution, i.e., there is a large probability that the service time is within a relatively narrow interval. For these systems we have the same effect as in the case where the service times are constant: When ρ is small the waiting-time distribution will have *most* of its mass in a number of narrow intervals. This explains the fact that our approximation performs slightly less for the E_{10}-distribution and $\rho = 0.2$ (see Section 3).

To conclude this section, we give a short outline of the approximate method.

Step 1. Determine α and β (see Theorem 2).

Step 2. Compute C_1, C_2, C_3, C_4 and Δ (see (11)) by using (1), (2), (3) and (4).

Step 3. If one of the cases i, ii or iii (see below (11)) applies, then use the formulas given there to compute γ, δ, η and φ. In these three cases, the approximation is given by (6). In Case iv, use (13) to find γ^*, δ^*, φ^* and ψ^*. The approximation is then given by (12). If none of these four cases apply, then follow the procedure described in Remark 1.

Table I
Conditional Waiting-Time Percentiles When $E(X) = 2$

			E_{10}, $C_S^2 = 0.1$				E_2, $C_S^2 = 0.5$				H_2, $C_S^2 = 2$			
		C_X^2	0.00	0.17	0.50	2.00	0.00	0.17	0.50	2.00	0.00	0.17	0.50	2.00
$\rho = 0.2$	$p = 0.2$	asy	0.56	0.80	0.85	0.52	0.41	0.68	0.75	0.38	0.00	0.00	0.00	0.00
		app	0.70	0.82	0.88	1.18	0.48	0.61	0.72	1.06	0.24	0.33	0.46	0.81
		exa	0.75	0.83	0.90	1.11	0.48	0.60	0.72	1.06	0.24	0.34	0.46	0.82
	$p = 0.5$	asy	0.94	1.28	1.73	3.34	0.95	1.31	1.76	3.32	0.00	0.05	0.97	2.98
		app	1.07	1.35	1.73	3.35	1.04	1.34	1.76	3.41	0.84	1.12	1.57	3.38
		exa	1.07	1.35	1.74	3.38	1.04	1.34	1.76	3.41	0.84	1.12	1.57	3.37
	$p = 0.8$	asy	1.71	2.22	3.46	8.85	2.01	2.54	3.73	9.03	2.15	2.85	4.25	9.64
		app	1.64	2.24	3.46	8.85	2.04	2.58	3.73	9.03	2.53	3.13	4.35	9.70
		exa	1.63	2.27	3.46	8.85	2.04	2.58	3.73	9.03	2.52	3.11	4.35	9.70
	$p = 0.9$	asy	2.28	2.94	4.76	13.01	2.81	3.48	5.21	13.36	4.18	4.97	6.73	14.67
		app	2.23	2.92	4.76	13.01	2.82	3.50	5.21	13.36	4.31	5.04	6.75	14.68
		exa	2.27	2.88	4.76	13.01	2.81	3.49	5.21	13.36	4.30	5.05	6.75	14.69
$\rho = 0.5$	$p = 0.2$	asy	0.72	0.92	1.04	0.86	0.63	0.84	0.99	0.83	0.00	0.00	0.21	0.53
		app	0.83	0.97	1.05	1.41	0.67	0.83	0.98	1.39	0.47	0.61	0.79	1.32
		exa	0.85	0.95	1.05	1.33	0.67	0.83	0.98	1.39	0.47	0.62	0.79	1.32
	$p = 0.5$	asy	1.52	1.88	2.47	4.68	1.65	2.02	2.62	4.82	1.50	1.95	2.74	5.21
		app	1.50	1.89	2.47	4.68	1.66	2.03	2.62	4.86	1.83	2.21	2.88	5.40
		exa	1.48	1.92	2.47	4.70	1.66	2.04	2.62	4.86	1.83	2.21	2.88	5.40
	$p = 0.8$	asy	3.08	3.76	5.26	12.14	3.65	4.32	5.79	12.59	5.50	6.20	7.68	14.34
		app	3.09	3.76	5.26	12.14	3.65	4.32	5.79	12.59	5.52	6.21	7.68	14.36
		exa	3.09	3.76	5.26	12.14	3.65	4.32	5.79	12.59	5.52	6.21	7.68	14.36
	$p = 0.9$	asy	4.27	5.18	7.37	17.78	5.16	6.06	8.19	18.48	8.52	9.42	11.41	21.25
		app	4.27	5.18	7.37	17.78	5.16	6.06	8.19	18.48	8.53	9.42	11.41	21.25
		exa	4.27	5.18	7.37	17.78	5.16	6.06	8.19	18.48	8.53	9.42	11.41	21.25
$\rho = 0.8$	$p = 0.2$	asy	1.40	1.68	2.02	2.95	1.47	1.74	2.12	3.12	1.32	1.65	2.18	3.58
		app	1.39	1.69	2.02	3.06	1.47	1.75	2.12	3.27	1.55	1.84	2.29	3.77
		exa	1.38	1.70	2.03	3.05	1.47	1.75	2.12	3.27	1.53	1.83	2.29	3.77
	$p = 0.5$	asy	3.71	4.38	5.64	10.98	4.27	4.95	6.22	11.59	6.16	6.86	8.21	13.77
		app	3.71	4.38	5.64	10.98	4.27	4.95	6.22	11.59	6.16	6.86	8.21	13.78
		exa	3.71	4.38	5.64	10.98	4.27	4.95	6.22	11.59	6.16	6.86	8.21	13.78
	$p = 0.8$	asy	8.21	9.66	12.69	26.65	9.75	11.20	14.20	28.09	15.58	17.02	19.96	33.63
		app	8.21	9.66	12.69	26.65	9.75	11.20	14.20	28.09	15.58	17.02	19.96	33.63
		exa	8.21	9.66	12.69	26.65	9.75	11.20	14.20	28.09	15.58	17.02	19.96	33.63
	$p = 0.9$	asy	11.61	13.65	18.03	38.50	13.90	15.92	20.24	40.57	22.72	24.70	28.86	48.65
		app	11.61	13.65	18.03	38.50	13.90	15.92	20.24	40.57	22.72	24.70	28.86	48.65
		exa	11.61	13.65	18.03	38.50	13.90	15.92	20.24	40.57	22.72	24.70	28.86	48.65

3. NUMERICAL RESULTS

In this section, we present numerical results for various models. We consider four different batch-size distributions: i) the constant batch size ($C_X^2 = 0$), ii) the uniformly distributed batch size ($C_X^2 = E(X-1)/3E(X)$), iii) the geometrically distributed batch size ($C_X^2 = E(X-1)/E(X)$), and iv) a batch size with a mixed-geometric distribution with balanced means, where C_X^2 is taken to be equal to 2. A batch-size distribution $\{b_n, n \geq 1\}$ is said to be a mixed-geometric distribution with balanced means when $b_n = qp_1(1-p_1)^{n-1} + (1-q)p_2(1-p_2)^{n-1}$, $n \geq 1$ with $q/p_1 = (1-q)/p_2$. Here C_X^2 denotes the squared coefficient of variation of the batch size X (i.e., the ratio of the variance to the squared mean). For the service time S of a customer, we consider the Erlang-10 distribution ($C_S^2 = 1/10$), the Erlang-2 distribution ($C_s^2 = 1/2$) and the hyperexponential distribution of order 2 with balanced means where $C_s^2 = 2$ is taken for the latter distribution. In all cases, we have taken $E(S) = 1$.

In Tables I and II we present numerical results which are obtained by: a) the first-order approximation (asy), b) the second-order approximation (app), and c) the exact solution (exa). The results are displayed by using the waiting-time percentiles. Since the percentiles $\nu(p)$ of the conditional waiting-time distribution of the delayed customer are defined for all $0 < p < 1$, it is convenient to use this conditional percentile rather than the percentiles $\xi(p)$ of the unconditional waiting-time distribution $W_q(\cdot)$. Note that $\nu(p)$ is determined by $(1 - (W_q(\nu(p))/(1 - W_q(0)) = 1 - p$ and thus $\xi(p_0) = \gamma(p_1)$ when $p_0 = 1 - (1-p_1)(1 - W_q(0))$. The numerical investigations reveal that for

Table II
Conditional Waiting-Time Percentiles When $E(X) = 5$

			$E_{10}, C_S^2 = 0.1$				$E_2, C_S^2 = 0.5$				$H_2, C_S^2 = 2$			
		C_X^2	0.00	0.27	0.80	2.00	0.00	0.27	0.80	2.00	0.00	0.27	0.80	2.00
$\rho = 0.2$	$p = 0.2$	asy	1.64	2.29	1.71	0.00	1.53	2.18	1.58	0.00	0.00	0.93	0.73	0.00
		app	1.53	1.70	1.71	1.94	1.27	1.61	1.58	1.78	0.78	0.98	1.16	1.38
		exa	1.41	1.56	1.72	1.89	1.23	1.40	1.58	1.78	0.80	0.97	1.16	1.38
	$p = 0.5$	asy	2.56	3.56	4.23	4.10	2.60	3.59	4.36	4.02	1.77	3.02	4.04	3.61
		app	2.69	3.45	4.23	5.78	2.62	3.37	4.36	5.82	2.29	3.05	4.11	5.79
		exa	2.84	3.51	4.23	5.78	2.70	3.44	4.36	5.82	2.28	3.06	4.11	5.79
	$p = 0.8$	asy	4.36	6.03	9.57	16.97	4.68	6.34	9.79	17.10	5.28	7.10	10.49	17.54
		app	4.44	6.15	9.57	17.05	4.77	6.29	9.79	17.21	5.39	7.11	10.49	17.84
		exa	4.33	6.26	9.57	17.05	4.74	6.49	9.79	17.21	5.37	7.10	10.49	17.84
	$p = 0.9$	asy	5.71	7.89	13.50	26.71	6.25	8.42	13.90	27.00	7.94	10.18	15.38	28.08
		app	5.72	8.02	13.50	26.71	6.31	8.40	13.90	27.01	7.96	10.19	15.38	28.13
		exa	5.54	7.86	13.50	26.71	6.23	8.47	13.90	27.01	7.97	10.19	15.38	28.13
$\rho = 0.2$	$p = 0.2$	asy	1.88	2.44	2.44	0.00	1.76	2.32	2.36	0.00	0.78	1.48	1.79	0.00
		app	1.94	2.25	2.44	2.82	1.73	2.12	2.36	2.74	1.29	1.62	1.98	2.44
		exa	1.93	2.17	2.44	2.78	1.74	2.04	2.36	2.74	1.30	1.63	1.98	2.44
	$p = 0.5$	asy	3.81	4.98	6.71	8.98	3.91	5.08	6.81	9.08	3.87	5.14	7.03	9.37
		app	3.85	5.00	6.71	9.56	3.94	5.05	6.81	9.72	4.00	5.18	7.04	10.20
		exa	3.87	5.10	6.71	9.57	3.96	5.14	6.81	9.72	3.98	5.17	7.04	10.20
	$p = 0.8$	asy	7.56	9.93	15.03	26.58	8.09	10.45	15.51	27.00	9.89	12.26	17.25	28.56
		app	7.56	9.95	15.03	26.59	8.09	10.45	15.51	27.01	9.89	12.27	17.25	28.61
		exa	7.59	9.90	15.03	26.59	8.09	10.43	15.51	27.01	9.89	12.27	17.25	28.61
	$p = 0.9$	asy	10.40	13.67	21.33	39.90	11.25	14.51	22.08	40.55	14.44	17.65	24.98	43.08
		app	10.40	13.68	21.33	39.90	11.25	14.51	22.08	40.56	14.44	17.65	24.98	43.09
		exa	10.39	13.68	21.33	39.90	11.25	14.51	22.08	40.56	14.44	17.65	24.98	43.09
$\rho = 0.2$	$p = 0.2$	asy	3.49	4.39	5.47	6.22	3.50	4.41	5.52	6.34	3.30	4.28	5.57	6.67
		app	3.52	4.39	5.47	7.00	3.52	4.39	5.52	7.13	3.41	4.32	5.59	7.46
		exa	3.57	4.45	5.47	7.00	3.54	4.43	5.52	7.13	3.39	4.32	5.59	7.46
	$p = 0.5$	asy	9.07	11.54	16.15	25.63	9.58	12.05	16.68	26.18	11.35	13.86	18.57	28.20
		app	9.07	11.54	16.15	25.64	9.58	12.05	16.68	26.20	11.35	13.86	18.57	28.23
		exa	9.06	11.54	16.15	25.64	9.58	12.05	16.68	26.20	11.35	13.86	18.57	28.23
	$p = 0.8$	asy	19.95	25.48	36.98	63.47	21.44	26.96	38.43	64.87	27.03	32.53	43.90	70.17
		app	19.95	25.48	36.98	63.47	21.44	26.96	38.43	64.87	27.03	32.53	43.90	70.17
		exa	19.95	25.48	36.98	63.47	21.44	26.96	38.43	64.87	27.03	32.53	43.90	70.17
	$p = 0.9$	asy	28.18	36.03	52.74	92.09	30.40	38.24	54.88	94.14	38.90	46.66	63.06	101.9
		app	28.18	36.03	52.74	92.09	30.40	38.24	54.88	94.14	38.90	46.66	63.06	101.9
		exa	28.18	36.03	52.74	92.09	30.40	38.24	54.88	94.14	38.90	46.66	63.06	101.9

nonlight traffic, the first-order approximation can be used for relatively small values of t. In terms of the conditional waiting-time percentile $\nu(p)$, the first-order approximation $(1/\beta)\ln(\alpha/(1-p)\rho)$ to $\nu(p)$ can be used for practical purposes when $p \geqslant 1-\rho$. This rule of thumb reflects the fact that the performance of the first-order approximation improves as ρ gets larger. The numerical results show the excellent performance of the second-order approximation to $W_q(t)$ for all values of t. Therefore, this approximation is well suited for practical purposes because it combines accuracy with ease of computation.

APPENDIX

The Derivation of the Second-Order Approximation

In this Appendix, we derive the approximation given in Section 2. The proposed conditions (7) to (11) for the determination of the numbers γ, δ, η and φ in approximation (6) lead to the following equations for these numbers

$$C_1 = \gamma + \eta, \quad C_2 = \gamma\delta + \eta\varphi$$

$$C_3 = \gamma/\delta + \eta/\varphi \tag{A1}$$

and

$$C_4 = \gamma/\delta^2 + \eta/\varphi^2.$$

In these four nonlinear equations, we restrict the feasible (complex) numbers as follows: for $\gamma \neq 0$ we require that $\mathrm{Re}(\delta) > \beta$ and for $\eta \neq 0$ we require that $\mathrm{Re}(\varphi) > \beta$. From (6)–(11), when $\gamma = 0$, the number δ is not determined by (A1) and, hence, can be taken as any real (or complex) number with $\mathrm{Re}(\delta) > \beta$. The same applies for φ when $\eta = 0$. Also, it will be used below that the roles of γ and δ in (A1) are interchangeable with the roles of η and φ, respectively. In the following, let $\Delta := (C_1C_3 - C_2C_4)^2 - 4(C_3^2 - C_4C_1)(C_1^2 - C_2C_3)$ and let β_0 denote some real constant with $\beta_0 > \beta$, e.g., $\beta_0 = 2\beta$.

Theorem A1. *The four nonlinear equations in* (A1) *have a solution if and only if one of the following four* (*exclusive*) *cases applies*:

i. $C_1 = C_2 = C_3 = C_4 = 0$

ii. $C_3 \neq 0,\ \dfrac{C_1}{C_3} > \beta,\ C_1^2 = C_2C_3$ *and* $C_3^2 = C_1C_4$

iii. $C_3^2 \neq C_1C_4,\ \dfrac{C_1C_3 - C_2C_4}{2(C_3^2 - C_1C_4)} > \beta$

and $0 < \Delta < [(C_1C_3 - C_2C_4) - 2(C_3^2 - C_1C_4)\beta]^2$

iv. $C_3^2 \neq C_1C_4,\ \dfrac{C_1C_3 - C_2C_4}{2(C_3^2 - C_1C_4)} > \beta,\ \Delta < 0.$

For the respective cases, we have as solutions

$$\text{i. } \gamma = \eta = 0 \ \text{ and } \ \delta = \varphi = \beta_0 \tag{A2}$$

$$\text{ii. } \gamma = C_1,\ \delta = \frac{C_1}{C_3},\ \eta = 0 \ \text{ and } \ \varphi = \beta_0 \tag{A3}$$

iii. *and* iv.

$$\delta = \frac{C_1C_3 - C_2C_4 + \sqrt{\Delta}}{2(C_3^2 - C_4C_1)}$$

$$\varphi = \frac{C_1C_3 - C_2C_4 - \sqrt{\Delta}}{2(C_3^2 - C_4C_1)} \tag{A4}$$

$$\gamma = \frac{C_2 - \varphi C_1}{\delta - \varphi} \ \text{ and } \ \eta = \frac{C_2 - \delta C_1}{\varphi - \delta}.$$

Proof. a. Suppose that $(\gamma, \delta, \eta, \varphi)$ is a solution to (A1) satisfying the restriction stated below (A1). First consider the case of $\gamma\eta = 0$. Because of (6), the pairs (γ, δ) and (η, φ) are interchangeable. Without loss of generality, we can thus assume that $\eta = 0$. This means that the set of equations (A1) reduces to

$$C_1 = \gamma, \quad C_2 = \gamma\delta,$$

$$C_3 = \frac{\gamma}{\delta} \ \text{ and } \ C_4 = \frac{\gamma}{\delta^2}. \tag{A5}$$

If $C_1 = 0$ it follows from (A5) that all C_i's are zero and so $\gamma = \eta = 0$ and $\delta = \varphi = \beta_0$ is a solution (Case i). If $C_1 \neq 0$, then $\gamma \neq 0$ and so by the convention below (A1) $\mathrm{Re}(\delta) > \beta$ which implies $\delta \neq 0$. Hence $C_1 \neq 0$ implies that all C_i's are nonzero and, therefore, $\delta = C_2/C_1 = C_1/C_3 = C_3/C_4$. This leads to the results for Case ii of Theorem A1.

Secondly consider the case that $\gamma\eta \neq 0$ and $\delta = \varphi$. By our convention $\delta \neq 0$. The set of equations (A1) reduces to

$$C_1 = \gamma + \eta, \quad C_2 = (\gamma + \eta)\delta,$$

$$C_3 = \frac{\gamma + \eta}{\delta} \ \text{ and } \ C_4 = \frac{\gamma + \eta}{\delta^2}. \tag{A6}$$

This set of equations is identical to (A5) with $\gamma' = \gamma + \eta$ and $\delta' = \delta$ and, thus, either i or ii applies. Next, we can replace the solution $(\gamma, \delta, \eta, \varphi)$ with a solution as in (A2) or (A3).

Finally, consider the remaining case of $\gamma\eta \neq 0$ and $\delta \neq \varphi$. Since $\gamma\eta(\delta - \varphi) \neq 0$ it is easily derived from (A1) that $(\delta C_1 - C_2)$, $(\delta C_3 - C_1)$ and $(\delta C_4 - C_3)$ are all nonzero and that φ is equal to both $(\delta C_1 - C_2)/(\delta C_3 - C_1)$ and $(\delta C_3 - C_1)/(\delta C_4 - C_3)$. Thus, δ must satisfy the relation

$$(C_3^2 - C_1C_4)\delta^2 - (C_1C_3 - C_2C_4)\delta + (C_1^2 - C_2C_3) = 0. \quad \text{(A7)}$$

Since the pairs (γ, δ) and (η, φ) can be interchanged, the same relation applies to φ, that is

$$(C_3^2 - C_1C_4)\varphi^2 - (C_1C_3 - C_2C_4)\varphi + (C_1^2 - C_2C_3) = 0. \quad \text{(A8)}$$

Since $\delta \neq \varphi$ and $C_3^2 - C_1C_4 = \gamma\eta(1/\delta - 1/\varphi)^2 \neq 0$, it follows from (A7) and (A8) that δ and φ can be taken as in (A4). From the condition that $\text{Re}(\delta) > \beta$ and $\text{Re}(\varphi) > \beta$, it easily follows that $(C_1C_3 - C_2C_4)/2(C_3^2 - C_1C_4) > \beta$ and $\Delta < [(C_1C_3 - C_2C_4) - 2(C_3^2 - C_1C_4)\beta]^2$. We must also have $\Delta \neq 0$, since otherwise $\delta = \varphi$. Next, using (A1), we find the equations $\gamma = (\varphi C_1 - C_2)/(\varphi - \delta)$ and $\eta = (\delta C_1 - C_2)/(\delta - \varphi)$. Note that η and γ are nonzero since $(\delta C_1 - C_2)$ and $(\varphi C_1 - C_2)$ are nonzero. Hence $\gamma\eta \neq 0$ and $\delta \neq \varphi$ imply the conditions of Case iii or Case iv.

b. By the construction of the solutions given in a, it follows that under the conditions stated in Cases i–iv the corresponding solutions satisfy the nonlinear equations (A1) with restrictions.

Remark A1. Note that in Case iv of the previous theorem, $\Delta < 0$, which implies that the numbers γ, δ, η and φ are not real. In this case, we have $\gamma = \bar{\eta}$ and $\delta = \bar{\varphi}$ and, therefore, we also have that $\eta e^{-\varphi t}$ is the complex conjugate of $\gamma e^{-\delta t}$. In the remaining analysis, we use the relations $e^{ix} = \cos(x) + i\sin(x)$ and $\theta\cos(x) + \omega\sin(x) = (\theta^2 + \omega^2)^{1/2}\cos(x + y)$ with $\theta^2 + \omega^2 > 0$ and y such that $\cos(y) = \theta/(\theta^2 + \omega^2)^{1/2}$ and $\sin(y) = -\omega/(\theta^2 + \omega^2)^{1/2}$. After some algebra, we find that $\gamma e^{-\delta t} + \eta e^{-\varphi t} = \gamma^*\cos(\varphi^* t + \psi^*)e^{-\delta^* t}$ where $\delta^* = \text{Re}(\delta)$, $\varphi^* = \text{Im}(\delta)$ and γ^* are given by

$$\delta^* = \frac{C_1C_3 - C_2C_4}{2(C_3^2 - C_4C_1)}$$

$$\varphi^* = \frac{\sqrt{-\Delta}}{2(C_3^2 - C_4C_1)}$$

$$\gamma^* = \sqrt{C_1^2 + ((C_2 - \delta^* C_1)/\varphi^*)^2}$$

and ψ^* is defined by $\cos(\psi^*) = C_1/\gamma^*$ and $\sin(\psi^*) = (\delta^* C_1 - C_2)/\gamma^*\varphi^*$.

ACKNOWLEDGMENT

I thank H. C. Tijms and the referees for their many suggestions to improve the paper.

REFERENCES

Chaudhry, M. L., and J. G. C. Templeton. 1983. *A First Course in Bulk Queues.* John Wiley, New York.

Burke, P. J. 1975. Delays in Single-Server Queues With Batch Input. *Opns. Res.* **23,** 830–833.

Cohen, J. W. 1976. On a Single Server Queue With Group Arrivals. *J. Appl. Prob.* **13,** 619–622.

Cooper, R. B. 1981. *Introduction to Queueing Theory,* 2nd ed. Elsevier North-Holland, New York.

Cromie, M. V., M. L. Chaudhry and W. K. Grassmann. 1979. Further Results for the Queueing System $M^x/M/c$. *J. Opns. Res. Soc.* **30,** 755–763.

Gaver, D. P. 1959. Imbedded Markov Chain Analysis of a Waiting-Line Process in Continuous Time. *Ann. Math. Stat.* **30,** 698–720.

Eikeboom, A. M., and H. C. Tijms. 1987. Waiting Time Percentiles in the Multiserver $M^x/G/1$ Queue With Batch Arrivals. *Prob. Eng. Inform. Sci.* **1,** 75–96.

Neuts, M. F. 1981. *Matrix Geometric Solutions in Stochastic Models—An Algorithmic Approach.* Johns Hopkins University Press, Baltimore.

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

Van Ommeren, J. C. W. 1988. Exponential Expansion for the Tail of the Waiting Time Probability in the Single Server Queue With Batch Arrivals. *Adv. Appl. Prob.* **20,** 880–895.