Waiting-Time Approximations in Multi-Queue Systems with Cyclic Service

O.J. Boxma * and B.W. Meister
IBM Zürich Research Laboratory, Säumerstrasse 4, 8803 Rüschlikon, Switzerland

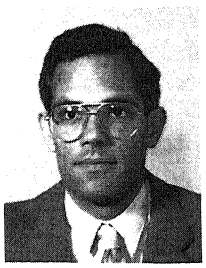
Received 28 February 1985
Revised 9 October 1985

This study is devoted to mean waiting-time approximations in a single-server multi-queue model with cyclic service and zero switching times of the server between consecutive queues. Two different service disciplines are considered: exhaustive service and (ordinary cyclic) nonexhaustive service. For both disciplines it is shown how estimates of the mean waiting times at the various queues can be obtained when no explicit information on arrival intensities and service-time distributions is available, while only the utilizations at the queues and the lengths of the busy periods of the system can be measured. In the exhaustive case, a known mean waiting-time approximation is shown to be suitable for our purposes; in the nonexhaustive case, a new approximation has been derived which is simple and yet more accurate than existing approximations. Extensive simulation validates the approximation methods.

Keywords: Waiting-Time Approximations, Single-Server Multi-Queue System, Exhaustive and Nonexhaustive Cyclic Services.

1. Introduction

In computer-communication systems which employ some variant of time-division multiplexing to share communication channels, the loop network is an important network structure. The queueing model of a



Onno J. Boxma received his Master's degree from Delft Technological University, The Netherlands, in 1974, and his Ph.D. from the University of Utrecht, The Netherlands, in 1977, both in Mathematics. During 1978–1979 he was an IBM Postdoctoral Fellow in Yorktown Heights, New York; in 1984, he spent three months at the IBM Zürich Research Laboratory. Since August 1985 he has been with the Centre for Mathematics and Computer Science (CWI), where he leads a small research group in queueing theory and performance evaluation. His research interests include queueing theory, computer performance, and stochastic scheduling.

He is a member of IFIP W.G. 7.3 and serves on the editorial board of the *Queueing Systems: Theory and Applications* journal.

Bernd Werner Meister received the M.S degree from Humboldt University, Berlin, Dem. Rep. Germany, in 1958, and the Ph.D. degree from the University of Freiburg im Breisgau, Fed. Rep. Germany, in 1962, both in Mathematics. From 1962 to 1964, he was a Scientific Assistant at the Institute for Applied Mathematics and Mechanics, University of Freiburg im Breisgau. In 1964 he joined the IBM Zürich Research Laboratory, Rüschlikon, Switzerland. Since then he has worked at the Heidelberg Scientific Center from 1972 to 1974, in the Department of Mathematics and Computer Science, University of Stuttgart, Fed. Rep. Germany, as a Visiting Professor for six months during 1977, and in the same Department at the University of Stuttgart as a Lecturer from 1977 on. His research interests include performance evaluation of local area networks. He has published numerous papers on hydrodynamic stability theory, numerical analysis, queueing theory and applications, and performance evaluation.

* On leave from the Mathematical Institute, University of Utrecht, Utrecht, The Netherlands. Present affiliation: Centre for Mathematics and Computer Science, P.O. Box 4079, 1009 AB Amsterdam, The Netherlands.

North-Holland
Performance Evaluation 7 (1987) 59–70

0166-5316/87/$3.50 © 1987, Elsevier Science Publishers B.V. (North-Holland)

loop network is a single-server multi-queue system with a cyclic service discipline: A server (the single communication channel of the loop network) is shared by customers (e.g., terminals) and the cyclic service discipline specifies how this sharing is accomplished. The analysis of this queueing model is finding a new application in token-bus and token-ring systems in local-area computer networks. For example, in a token-ring system, a permission token to access a shared transmission medium is passed around cyclically among the stations attached. The queueing of model of such loop and ring systems, the subject of our analysis, is described below.

1.1. Model description

A single service facility serves N queues $Q_1, \ldots, Q_N$ (with infinite buffer capacities) in a cyclic manner. We consider two different service disciplines:

(a) *Exhaustive service* (also called polling, or alternating priority): when the server visits a queue, he serves its customers until that queue is empty.

(b) (*Ordinary cyclic*) *nonexhaustive service* (also called chaining, or alternating service): When the server visits a queue, he only serves one customer (if any is present).

(In the literature, the term nonexhaustive service is used for the general case that the server serves at most a fixed number of customers, K, at each queue he visits; we only consider $K = 1$, sometimes adding 'ordinary cyclic' to the term nonexhaustive service to make this distinction.)

In both cases, switch-over times from each queue to the next are considered to be negligible. Customers (messages) arrive at all queues according to independent Poisson processes with rates $\lambda_1, \ldots, \lambda_N$; the total arrival rate is Λ. Customers arriving at Q_i will be called type-i customers. The service times (transmission times) of type-i customers are independent, identically-distributed, stochastic variables, with distribution $B_i(\cdot)$ with first and second moments β_i, $\beta_i^{(2)}$ and Laplace–Stieltjes transform (LST) $\beta_i(\cdot)$. The service-time processes at the various queues are independent of each other and of the arrival processes. The utilization at Q_i is denoted by

$$\rho_i \overset{\text{def}}{=} \lambda_i \beta_i, \quad i = 1, 2, \ldots, N. \tag{1}$$

We assume that

$$\rho \overset{\text{def}}{=} \rho_1 + \cdots + \rho_N < 1 \tag{2}$$

to ensure that the stationary distributions of all relevant queueing quantities exist.

1.2. Problem description

Important performance measures in local-area networks with ring or bus topology and access through a permission token, and in time-division multiplexing loop networks, are the mean waiting times Ew_i at Q_i, $i = 1, 2, \ldots, N$. In the case of an exhaustive service discipline, expressions for these N mean waiting times are known but their evaluation requires the solution of a set of linear equations ($O(N^2)$ or $O(N^3)$ linear equations, depending on the model variant under consideration). This motivated Bux and Truong [3] to derive some approximate formulas for the mean waiting times; we shall discuss their two best approximations in Section 3. The case of a nonexhaustive service discipline has almost completely defied exact mathematical analysis. Complete exact analyses of the case with two queues without switching times and the case with two identical queues with switching times were presented in [4] and [2], respectively (leading to waiting-time and queue-length distributions), but they required the rather complicated method of Riemann–Hilbert boundary value problems. In the more general case of an arbitrary number of queues, only the mean waiting times are known when all queues have identical characteristics. Therefore, it is not surprising that several mean waiting-time approximations have been derived in the nonexhaustive case; we mention in particular the approximation of Kuehn [8].

In this paper we consider the following problem: Suppose that in an (exhaustive or nonexhaustive)

cyclic server system as described above one would like to determine the mean waiting times at the various queues. However, one has no detailed information concerning arrival rates or service-time distributions; one can only measure the global system variable ρ = Pr(server is busy), and the lengths of the busy periods of the server (the periods during which the server is uninterruptedly serving customers); in particular, one can then estimate EP and EP^2, the first and second moments of the busy period P, respectively. Furthermore, in most of what follows, we assume that the utilizations ρ_i are also known (can be measured). The question now is: How can one use this information to give reasonable estimates of the mean waiting times Ew_i?

To get some feeling for this problem, we first consider the model of N identical queues which we call the completely symmetric case: $\lambda_i = \Lambda/N$ and $B_i(\cdot) \equiv B(\cdot)$. Then as is well known, the mean waiting time at each queue, in both the exhaustive and nonexhaustive cases, is identical to the mean waiting time in the classical M/G/1 queue with arrival rate Λ and service-time distribution $B(\cdot)$; i.e. (with $\beta_i^{(2)} \equiv \beta^{(2)}$),

$$Ew_i = \Lambda\beta^{(2)}/[2(1-\rho)], \quad i = 1, 2, \ldots, N. \tag{3}$$

Furthermore, in this case, the busy period P is distributed as the busy period in this M/G/1 model; hence,

$$EP = \rho/[\Lambda(1-\rho)], \qquad EP^2 = \beta^{(2)}/(1-\rho)^3. \tag{4}$$

Consequently, Ew_i can be expressed in ρ, EP, and EP^2 in the following way:

$$Ew_i = [EP^2/(2EP)]\rho(1-\rho), \quad i = 1, 2, \ldots, N. \tag{5}$$

In the general asymmetric exhaustive case, though, Ew_i depends in a complicated way on the $3N$ parameters λ_i, β_i, $\beta_i^{(2)}$, $i = 1, 2, \ldots, N$; in the nonexhaustive case, Ew_i depends on the complete service-time distributions [4]. In both cases, Ew_i clearly cannot be simply expressed in EP, EP^2, and ρ_i, $i = 1, 2, \ldots, N$.

To solve the problem described above, we proceed as follows. First, we present (in Section 2) an exact analysis of the busy-period distribution in both the exhaustive and nonexhaustive cases. In Sections 3 and 4, we tackle, for the exhaustive and nonexhaustive cases, respectively, the problem of obtaining mean waiting-time estimates from the values of EP, EP^2 and the utilizations ρ_i. In the exhaustive case, the known approximation (Bux and Truong [3])

$$Ew_i = \left[\sum_{j=1}^{N} \lambda_j\beta_j^{(2)}/\{2(1-\rho)\}\right]\left[(1-\rho_i)\Big/\left(1-\frac{1}{\rho}\sum_{j=1}^{N}\rho_j^2\right)\right], \quad i = 1, 2, \ldots, N, \tag{6}$$

is shown to be suitable for our proposes; in the nonexhaustive case, we derive a new, simple approximation which is more accurate than earlier approximations, and which again is suitable in the present setting:

$$Ew_i = \left[\sum_{j=1}^{N} \lambda_j\beta_j^{(2)}/\{2(1-\rho)\}\right]\left[(1-\rho+\rho_i)\Big/\left(1-\rho+\frac{1}{\rho}\sum_{j=1}^{N}\rho_j^2\right)\right], \quad i = 1, 2, \ldots, N. \tag{7}$$

In Section 5, our approximations and estimation methods are validated by extensive simulation results.

2. Analysis of the busy period

2.1. Proposition. *In both the exhaustive and nonexhaustive cases, the busy-period distribution is exactly the same as in the* M/G/1 *queue with arrival rate* Λ *and service-time distribution*

$$B^*(t) \overset{\text{def}}{=} \sum_{j=1}^{N} \frac{\lambda_j}{\Lambda} B_j(t), \quad t \geqslant 0 \tag{8}$$

(*with first and second moments*

$$\beta \overset{\text{def}}{=} \sum_{j=1}^{N} \frac{\lambda_j}{\Lambda} \beta_j = \rho/\Lambda, \qquad \beta^{(2)} \overset{\text{def}}{=} \sum_{j=1}^{N} \frac{\lambda_j}{\Lambda} \beta_j^{(2)},$$

respectively).

In particular, cf. (4),

$$EP = \rho/[\Lambda(1-\rho)], \qquad EP^2 = \beta^{(2)}/(1-\rho)^3. \tag{9}$$

Proof. In the cyclic server system with either exhaustive or nonexhaustive service, consider the process $\{V(t),\ t \geqslant 0\}$, where $V(t)$ is the total amount of work (sum of service requirements) in the system at time t. In both cases, $V(t)$ has the same distribution as $V^*(t)$, with $V^*(t)$ the amount of work in the M/G/1 queue described above (where the M/G/1 queue corresponds to a cyclic server queue with service in order of arrival). For both, $V(t)$ and $V^*(t)$, when positive, decrease at unit rates between two successive arrivals, regardless of the service discipline, and at arrival epochs they both increase by the required service time of the arriving customer—which has nothing to do with the service discipline. The probability distributions of these increments are clearly the same in the cyclic and the M/G/1 models. The fact that $V(t)$ and $V^*(t)$ are identically distributed immediately implies that the busy periods in both models are identically distributed, as these busy periods are just the uninterrupted periods during which $V(t) > 0$ and $V^*(t) > 0$, respectively. The moment expressions in (9) now immediately follow from M/G/1 theory (cf. [5]). □

2.2. Remark. In fact, a somewhat stronger result holds, which we mention here for future reference. The reasoning in the proof of Proposition 2.1 makes it immediately clear that the distribution of a busy period in the cyclic server model (in either the exhaustive or the nonexhaustive cases), given that a type-i customer starts this busy period, is the same as the distribution of a busy period in the M/G/1 model under consideration, given that the first service time of that busy period has distribution $B_i(\cdot)$. Application of the branching argument which is often used to derive the busy-period distribution in the M/G/1 model (cf. [5, p. 250]) now yields

$$E\left[e^{-sP} \mid P \text{ starts at } Q_i\right] = \beta_i(s + \Lambda - \Lambda\varphi(s)), \quad \text{Re } s \geqslant 0; \tag{10}$$

here, $\varphi(s)$ denotes the LST of the unconditional busy-period distribution. In particular,

$$E[P \mid P \text{ starts at } Q_i] = \beta_i/(1-\rho), \tag{11}$$

$$E\left[P^2 \mid P \text{ starts at } Q_i\right] = \frac{\beta_i^{(2)}}{(1-\rho)^2} + \frac{\beta_i}{(1-\rho)^3} \sum_{j=1}^{N} \lambda_j \beta_j^{(2)}, \quad i = 1, 2, \ldots, N. \tag{12}$$

Unlike busy periods, waiting times in the exhaustive service model, the nonexhaustive service model, and the related M/G/1 model are not identically distributed. In the next two sections we shall show how one can accurately estimate the mean waiting times Ew_i, using only EP, EP^2, and the utilizations ρ_i, by exploiting the simple form of (9) in which the second moments of the service-time distributions only occur in a special way.

3. The exhaustive service discipline

The cyclic service model with exhaustive service discipline (i.e., when the server visits a queue he empties its buffer completely) has been studied extensively (see survey by Takagi and Kleinrock [10]). No general expression for the mean waiting times in a model of N queues is known. Determination of these means generally requires the solution of a set of equations; for the model with zero switching times, which we consider in the present section, Cooper [6] has derived a set of $N(N-1)$ linear equations which have to be solved to determine the mean waiting times. In the performance evaluation of local-area networks, there

is a practical need for an accurate mean waiting-time formula for the model described above, a formula which can easily be evaluated for a large number of queues. With this in mind, Bux and Truong [3] derived a mean waiting-time approximation different from (6) for the general exhaustive-service model with nonzero switching times. For zero switching times, their approximation formula reduces to the following:

$$Ew_i = \frac{\lambda_i \beta_i^{(2)}}{2(1-\rho_i)} + \frac{1}{2(1-\rho)(1-\rho_i)} \sum_{k=1, k\neq i}^{N} \frac{\lambda_k \beta_k^{(2)} (1-\rho_i)^2 + \lambda_i \beta_i^{(2)} \rho_k^2}{1-\rho_i-\rho_k+2\rho_i\rho_k}, \quad i = 1, 2, \ldots, N. \tag{13}$$

By comparison with simulation results they show that the accuracy of this approximation is high. Furthermore, it has the following pleasing properties: (i) It is exact for $N = 1$ and $N = 2$, (ii) it is exact in the completely symmetrical case $\lambda_i = \Lambda/N$, $B_i(\cdot) \equiv B(\cdot)$, and (iii) summation over the mean waiting times weighted with the corresponding utilizations yields

$$EW \overset{\text{def}}{=} \sum_{i=1}^{N} \frac{\rho_i}{\rho} Ew_i = \sum_{j=1}^{N} \lambda_j \beta_j^{(2)} / [2(1-\rho)]. \tag{14}$$

This is exactly the mean waiting time in the M/G/1 model with arrival rate Λ and service-time distribution $B^*(\cdot)$ defined in (8). Indeed, the conservation law (cf. [7]) states that (14) holds for the exact mean waiting times.

Let us now turn to the problem described in Section 1: to give reasonable estimates of the mean waiting times when only the utilizations ρ_i, and the mean and second moments (or variance) of the busy period of the system are known (can be measured). From (9) and (14) it is clear how to estimate EW from measured values of ρ, EP, and EP^2, as the following exact expression holds:

$$EW = [EP^2/(2EP)]\rho(1-\rho). \tag{15}$$

Note that when $B_i(\cdot) \equiv B(\cdot)$, arrival rates not necessarily being identical, EW represents the mean waiting time averaged over all customers.

However, estimation of the separate mean waiting times Ew_i presents more problems; the approximate Ew_i's in (13) depend in a rather complicated way on the parameters λ_j, β_j, and $\beta_j^{(2)}$, $j = 1, 2, \ldots, N$. In their paper [3], Bux and Truong discuss a few alternative approximations, and the first of these is a good approximation ideally suited for our purposes. It reads

$$Ew_i = \left[\sum_{j=1}^{N} \lambda_j \beta_j^{(2)} \Big/ \{2(1-\rho)\} \right] \left[(1-\rho_i) \Big/ \left(1 - \frac{1}{\rho} \sum_{j=1}^{N} \rho_j^2 \right) \right], \quad i = 1, 2, \ldots, N. \tag{16}$$

Hence we can write

$$Ew_i = EW(1-\rho_i) \Big/ \left(1 - \frac{1}{\rho} \sum_{j=1}^{N} \rho_j^2 \right), \quad i = 1, 2, \ldots, N, \tag{17}$$

which yields the following estimation for Ew_i, only using EP, EP^2, and the utilizations ρ_i (cf. (15)):

$$Ew_i = \frac{EP^2}{2EP} \rho(1-\rho)(1-\rho_i) \Big/ \left(1 - \frac{1}{\rho} \sum_{j=1}^{N} \rho_j^2 \right), \quad i = 1, 2, \ldots, N. \tag{18}$$

Clearly, (18) yields a good estimation when (16) is a good approximation. Equation (16) is not quite as good as (13) but, as remarked by Bux and Truong [3], comparison with simulation shows that it can serve as a reasonable approximation over the entire range of parameters. Like (13), it is exact for $N = 1$, it is exact in the completely symmetric case, and it yields the exact formula (14) for the utilization weighted mean EW (it was specifically constructed that way); but it is not exact for $N = 2$. Numerical results, presented in Section 5, confirm the usefulness of approximation (16), and at the same time display the main reason for its success: the robustness of the mean waiting times with respect to the traffic pattern.

While the ordinary M/G/1 queue is very sensitive to changes in, say, its arrival rate, in the cyclic service model two queues with identical service-time distributions but strongly differing arrival rates generally still have rather similar mean waiting times.

A pleasant global asset of (16) is that it reflects the property that customers in light-traffic queues usually experience a longer waiting time than customers in heavy-traffic queues. Bux and Truong explain this property by pointing out that customers arriving at a heavy-traffic queue have a better chance that their queue is currently being served than those arriving at a light-traffic queue. We return to this matter in Section 4 in our discussion of the nonexhaustive service discipline, which turns out to give rise to just the opposite phenomenon.

3.1. Remark. In Remark 2.2, expressions for the first two busy-period moments were given, under the condition that the busy period started at Q_i. When these conditional moments can be estimated by monitoring the system, it is easy to obtain estimates for the first and second moments of the individual service-time distributions. In that case, one is able to use the somewhat more sophisticated approximation (13).

4. The nonexhaustive service discipline

The cyclic service model with an (ordinary cyclic) nonexhaustive service discipline (i.e., when the server visits a queue, he will only serve one customer—if any is present) has extensively been studied, like its 'exhaustive' counterpart. Again, we refer to the survey of Takagi and Kleinrock [10]. The nonexhaustive service discipline is more important than the exhaustive one, being fair to small users instead of heavy users. The nonexhaustive service discipline gives rise to a notoriously difficult mathematical model, and one generally has to take recourse to simulation or approximations. The goal of this section is to present a new approximation for the mean waiting times in the case of zero switching times. In deriving this approximation, we aimed at fulfilling the following criteria:

(i) It is exact for $N = 1$.

(ii) It is exact in the completely symmetric case.

(iii) Summation over the mean waiting times Ew_i, weighted with the corresponding utilizations, yields the exact result for EW displayed in (14) (as required by the conservation law).

(iv) It can be used to estimate mean waiting times when only the utilizations ρ_i and the mean and second moments (or variance) of the busy period of the system are known (or can be measured).

In essence, we have aimed at obtaining an approximation which has the same properties as approximation (16) in the exhaustive case.

Although several mean waiting-time approximations for the present model are known (cf. [8] for a well-known approximation and further references; see also [9] for a recent approximation for the case that the system has multiple cyclic servers), they do not fulfill the above criteria. Most of these approximations have been devised for the general case of nonzero switching times, and have the property that their accuracy worsens for decreasing switching times (often leading to useless results when switching times reduce to zero). Indeed, while Kuehn's [8] approximation is exact for $N = 1$ and leads to excellent results in a great variety of cyclic models with nonexhaustive service, the case of zero switching times represents the worst case with respect to the approximation accuracy. In this case, it reads as follows:

$$Ew_i = \left[\sum_{j=1}^{N} \lambda_j \beta_j^{(2)} / \{2(1-\rho)\}\right](1-\rho+\rho_i) + \frac{\rho_i \beta_i}{2(1-\rho)(1-\rho+\rho_i)}\left[2\rho - 2\rho_i + 2\rho\rho_i - \rho^2 - \sum_{j=1}^{N} \rho_j^2\right], \quad i = 1, 2, \ldots, N. \tag{19}$$

It can easily be verified that this approximation does not fulfill criteria (ii), (iii), and (iv) above.

We propose the following approximation for the mean waiting times Ew_i in the case of zero switching time:

$$Ew_i = \left[\sum_{j=1}^{N} \lambda_j \beta_j^{(2)} / \{2(1-\rho)\}\right]\left[(1-\rho+\rho_i) \Big/ \left(1-\rho+\frac{1}{\rho}\sum_{j=1}^{N}\rho_j^2\right)\right]$$

$$= EW(1-\rho+\rho_i) \Big/ \left(1-\rho+\frac{1}{\rho}\sum_{j=1}^{N}\rho_j^2\right), \quad i=1,2,\ldots,N. \tag{20}$$

Note that (20) does fulfill criteria (i) through (iv), and that it is very similar to approximation (16) for the exhaustive case. Also note the difference from (16): equation (20) reflects the property that customers in light-traffic queues experience a shorter waiting time than customers in heavy-traffic queues—which is usually indeed the case in this model, as opposed to the exhaustive-service model.

4.1. Derivation of approximation (20)

First, we give some definitions: x_i denotes the queue length in Q_i just before the arrival of a type-i customer, c_i denotes the length of a cycle of the server which starts with a service at Q_i and which ends when the server returns to Q_i (an 'i-cycle'), and rc_i denotes a residual i-cycle, i.e., the time from the arrival of a type-i customer until the server returns to Q_i.

The waiting time w_i at Q_i consists of two parts: a residual cycle rc_i and just as many complete i-cylces as there are type-i customers waiting; hence,

$$Ew_i = Erc_i + Ex_i Ec_i. \tag{21}$$

Since Poisson arrivals see time averages [11], Ex_i equals the mean number of waiting customers at an arbitrary instant of time. This permits the use of Little's formula $Ex_i = \lambda_i Ew_i$, and it follows that

$$Ew_i = Erc_i/(1-\lambda_i Ec_i). \tag{22}$$

We introduce two approximation assumptions to estimate the unknown Ec_i and Erc_i.

Approximation Assumption A

$$Ec_i = \beta_i/(1-\rho+\rho_i), \quad i=1,2,\ldots,N. \tag{23}$$

Kuehn [8] has also used (23) in his approximation. His motivation is as follows: An i-cycle consists of a type-i service and—possibly—services of customers of other types:

$$Ec_i = \beta_i + \sum_{j\neq i} \alpha_{ij}\beta_j, \tag{24}$$

with

$$\alpha_{ij} \overset{\text{def}}{=} \Pr(i\text{-cycle contains a type-}j\text{ service}) = E(\text{number of type-}j\text{ services in an }i\text{-cycle})$$
$$\approx E(\text{number of type-}j\text{ arrivals in an }i\text{-cycle}) = \lambda_j Ec_i, \quad j\neq i. \tag{25}$$

Equation (23) immediately follows from (24) and (25). The last equality in (25) is based on a balance-of-flow argument. Equation (23) is obviously exact for $N=1$; it should also be very accurate when traffic is low and in the completely symmetric case, but not when traffic is heavy with highly asymmetric arrival rates and service demands (see also Remark 5.2).

Approximation Assumption B. Erc_i is independent of i.

This assumption is trivially exact for $N=1$ and in the completely symmetric case. We now show that it is very accurate when traffic is low. In a low-traffic situation, it is highly probable that at most one

customer is present when a type-i customer arrives, and that, if one is present, no customer arrives at an intermediate queue before the i-service has started. Hence, in low traffic, Erc_i is closely approximated by

$$Erc_i = (1-\rho)0 + \rho \sum_{k=1}^{N} \Pr(\text{customer in service is type-}k)\frac{\beta_k^{(2)}}{2\beta_k} = \frac{1}{2}\sum_{k=1}^{N}\lambda_k\beta_k^{(2)}, \tag{26}$$

which is independent of i.

In heavy traffic, Erc_i will depend on i but, in general, the dependence will be small—in particular when the number of queues is large and traffic is not highly asymmetric. Moreover, for heavy-traffic queues, the contribution of Erc_i to Ew_i is generally small.

Now, we are going to determine an estimate for $Erc \equiv Erc_i$ by demanding that criterion (iii) above be fulfilled, i.e., the mean waiting-time approximation should correctly reflect the fact that the system obeys the mean waiting-time conservation law. From (22) and Approximation Assumptions A and B,

$$Ew_i = [Erc/(1-\rho)](1-\rho+\rho_i), \quad i = 1, 2, \ldots, N, \tag{27}$$

and hence

$$EW = \sum_{i=1}^{N}\frac{\rho_i}{\rho}Ew_i = \frac{Erc}{1-\rho}\sum_{i=1}^{N}\frac{\rho_i}{\rho}(1-\rho+\rho_i) = \frac{Erc}{1-\rho}\left(1-\rho+\frac{1}{\rho}\sum_{j=1}^{N}\rho_j^2\right).$$

This implies that

$$Erc = EW(1-\rho)\Big/\left(1-\rho+\frac{1}{\rho}\sum_{j=1}^{N}\rho_j^2\right) \tag{28}$$

(cf. (26) for $\rho \downarrow 0$). Finally, the proposed mean waiting-time approximation (20) follows from (27) and (28).

4.1. Remark. Recently, Arndt and Sulanke [1] announced an approximation of the case of $N = 2$ queues which satisfies our first three criteria. It reads

$$\begin{aligned} Ew_1 &= EW[1-\min\{1, \lambda_2/\lambda_1\}(\rho_1+\rho_2)]/[1-\min\{1, \lambda_2/\lambda_1\}(\rho_1+\lambda_1/\beta_2)], \\ Ew_2 &= EW[1-\min\{1, \lambda_1/\lambda_2\}(\rho_1+\rho_2)]/[1-\min\{1, \lambda_2/\lambda_1\}(\rho_1+\lambda_1/\beta_2)]. \end{aligned} \tag{29}$$

Some preliminary tests show that this is a very good approximation. It would be interesting to investigate the possibility of extending it to the case of an arbitrary number of queues.

4.2. Remark. In the foregoing, we have restricted ourselves to the case of Poisson arrivals. Kuehn [8] stresses the importance of analytic studies on cyclic queueing systems with more general arrival processes, as his simulation reveals that mean waiting times in cyclic server queues can be very sensitive to the arrival processes. Our approximation approach can be extended in the following way to hold for general arrival processes at the various queues: Formula (22) remains valid; Approximation Assumptions A and B can still be used; the conservation law for the weighted sum of the mean waiting times of type-i customers (see below (14)) can be extended to the case of general arrival processes (see [7, p. 117, formula (3.22)]), but the resulting expression for this weighted sum, which can be shown to equal the mean waiting time in the corresponding G/G/1 queue, is in general unknown. For specific choices of the arrival and service-time processes, it can be evaluated, or approximated otherwise. Once an (exact or approximate) expression for this weighted sum has been obtained, Erc_i and Ew_i are obtained similarly as above, and Ew_i is again approximated by the expression in the right-hand side of (20). It is tempting to propose that (17) can also be extended as approximation for G/G/1, but this requires further investigation.

5. Numerical results and conclusions

In this section we present comparisons of our approximations with exact results for the exhaus-

Table 1
Comparison of the mean waiting-time approximation (6) (Bux and Truong, exhaustive) with exact results, and of the mean waiting-time approximation (7) (nonexhaustive) with simulation results; $N = 3$ queues, $\Lambda = 1$, $\lambda_1 = 0.6$, $\lambda_2 = \lambda_3 = 0.2$; all service-time distributions negative exponential with identical means

	Exhaustive			Nonexhaustive		
ρ	0.3	0.5	0.8	0.3	0.5	0.8
Ew_1 exact/simulated	0.121	0.442	2.505	0.135	0.553	4.144
Ew_1 approximation	0.122	0.439	2.568	0.136	0.556	3.942
Error %	0.8	−0.7	2.5	0.7	0.5	−4.9
Ew_2 exact/simulated	0.140	0.579	4.157	0.115	0.389	1.464
Ew_2 approximation	0.140	0.577	4.148	0.118	0.417	2.087
Error %	0	−0.3	−0.2	2.6	7.2	42.6
Ew_3 exact/simulated	0.141	0.595	4.328	0.115	0.396	1.490
Ew_3 approximation	0.140	0.577	4.148	0.118	0.417	2.087
Error %	−0.7	−3.0	−4.2	2.6	5.3	40.1
EW	0.129	0.5	3.2	0.129	0.5	3.2
EW in approximation of Kuehn				0.116	0.416	2.397

tive case (obtained by solving Cooper's [6] set of equations), and with simulation results for both the exhaustive and nonexhaustive cases. The models under consideration contain so many parameters that we can only present a relatively small number of representative examples. In particular, we present results for $N = 3$ stations (Tables 1 and 2) and $N = 16$ stations (Tables 3, 4, 5, and 6). The simulation was performed using the IBM RESQ2 package. Every simulation run consisted of five independent replications, in each of which 100 000 customers were generated for the case of 3 queues, and 500 000 for the case of 16 queues. Still, the simulated values of the mean waiting times may have errors in the order of 10% when traffic is heavy and pronouncedly asymmetric. This is, in fact, exactly the situation in which approximations (6) and (7) are least accurate.

Approximation (6) of Bux and Truong for the exhaustive case has a rather small error, usually just a few percent. Our approximation (7) for the —more complicated—nonexhaustive case is just

Table 2
Comparison of the mean waiting-time approximation (6) (Bux and Truong, exhaustive) with exact results, and of the mean waiting-time approximation (7) (nonexhaustive) with simulation results; $N = 3$ queues, $\Lambda = 1$, $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$; all service-time distributions negative exponential with $\beta_2 = \beta_3 = \frac{1}{3}\beta_1$

	Exhaustive			Nonexhaustive		
ρ	0.3	0.5	0.8	0.3	0.5	0.8
Ew_1 exact/simulated	0.154	0.552	3.153	0.175	0.677	4.473
Ew_1 approximation	0.161	0.592	3.390	0.180	0.733	5.203
Error %	4.5	7.2	7.5	2.9	8.3	16.3
Ew_2 exact/simulated	0.187	0.777	5.510	0.153	0.559	3.534
Ew_2 approximation	0.184	0.762	5.476	0.155	0.550	2.755
Error %	−1.6	−1.9	−0.6	1.3	−1.6	−22.0
Ew_3 exact/simulated	0.201	0.867	6.151	0.159	0.578	3.606
Ew_3 approximation	0.184	0.762	5.476	0.155	0.550	2.755
Error %	−8.5	−12.1	−11.0	−2.5	−4.8	−23.6
EW	0.170	0.66	4.224	0.170	0.66	4.224
EW in approximation of Kuehn				0.153	0.555	3.204

Table 3
Comparison of the mean waiting-time approximations (6) (Bux and Truong, exhaustive) and (7) (nonexhaustive) with simulation results: $N=16$ queues, $\Lambda=1$, $\lambda_1=\cdots=\lambda_4=0.16$, $\lambda_5=\cdots=\lambda_{16}=0.03$; all service-time distributions negative exponential with identical means

	Exhaustive			Nonexhaustive		
ρ	0.3	0.5	0.8	0.3	0.5	0.8
Ew_{1-4} simulation [a]	0.126	0.487	3.051	0.131	0.532	3.905
Ew_1 approximation	0.127	0.488	3.068	0.131	0.521	3.612
Error %	0.8	0.2	0.6	0	−2.1	−7.5
Ew_{5-16} simulation [a]	0.132	0.520	3.412	0.123	0.439	1.896
Ew_5 approximation	0.132	0.522	3.434	0.124	0.463	2.467
Error %	0	0.4	0.6	0.8	5.5	30.1
EW	0.129	0.5	3.2	0.129	0.5	3.2
EW in approximation of Kuehn				0.099	0.310	1.415

[a] The results represent mean waiting times averaged over the corresponding group of queues.

as good for server utilizations of 0.3 and 0.5; for utilization of 0.8, the approximation becomes worse for queues with relatively low traffic but is still good for heavy-traffic queues (cf. Remark 5.2).

The tables reflect the property mentioned in Sections 3 and 4 that a queue with high traffic compared to the other queues, e.g., Q_1 in all our examples, has relatively short mean waiting times for exhaustive service and relatively long mean waiting times for nonexhaustive service.

Comparison of the results in Tables 1 and 2 with those in the somewhat related Tables 3 and 6 suggests that the approximation for comparable cases becomes more accurate for increasing N. The comparison also reveals that the mean waiting times tend to 'average out' for increasing N: Differences between mean waiting times at the various queues become quite small when traffic is not highly asymmetric—as predicted by the approximations. This 'averaging-out' effect is one of the main reasons why approximations (6) and (7) perform well for large numbers of stations and for a wide range of realistic traffic parameters.

5.1. Remark. In connection with the 'averaging-out' effect we mention the following: The exact results of Cooper [6] for the exhaustive case show that mean waiting times in two queues with identical traffic characteristics can differ as a function of the positions of these queues with respect to queues with other characteristics; however, these differences are generally rather small. The largest differences in mean waiting times for queues with

Table 4
Comparison of the mean waiting-time approximations (6) (Bux and Truong, exhaustive) and (7) (nonexhaustive) with simulation results; $N=16$ queues, $\Lambda=1$, $\lambda_1=0.6$, $\lambda_2=\cdots=\lambda_{16}=\frac{2}{75}$; all service-time distributions negative exponential with identical means

	Exhaustive			Nonexhaustive		
ρ	0.3	0.5	0.8	0.3	0.5	0.8
Ew_1 simulation	0.118	0.423	2.301	0.140	0.595	4.538
Ew_1 approximation	0.119	0.430	2.365	0.140	0.584	4.383
Error %	0.8	1.7	2.8	0	−1.8	−3.4
Ew_{2-16} simulation [a]	0.144	0.613	4.503	0.110	0.355	1.149
Ew_2 approximation	0.144	0.606	4.452	0.113	0.375	1.427
Error %	0	−1.1	−1.1	2.8	5.6	24.2
EW	0.129	0.5	3.2	0.129	0.5	3.2
EW in approximation of Kuehn				0.111	0.386	2.685

[a] The results represent mean waiting times averaged over the corresponding group of queues.

Table 5
Comparison of the mean waiting-time approximations (6) (Bux and Truong, exhaustive) and (7) (nonexhaustive) with simulation results; $N=16$ queues, $\Lambda=1$, $\lambda_1=\lambda_7=0.15$, $\lambda_2=\cdots=\lambda_6=\lambda_8=\cdots=\lambda_{16}=0.05$; service-time distributions at $Q_2,\ldots,Q_6$, $Q_8,\ldots,Q_{16}$ negative exponential with identical means; service-time distribution at Q_1 Erlang-4 with $\beta_1=6\beta_2$; service-time distribution at Q_7 two-stage hyperexponential $q(1-e^{-t/m_1})+(1-q)(1-e^{-t/m_2})$ with $q=0.8873$, $m_1=0.5635\times\beta_7$, $m_2=4.4365\times\beta_7$, $\beta_7=6\beta_2$, $\beta_7^{(2)}=5\beta_7^2$

	Exhaustive			Nonexhaustive		
ρ	0.3	0.5	0.8	0.3	0.5	0.8
Ew_1 simulation	0.378	1.448	8.890	0.377	1.479	10.748
Ew_1 approximation	0.351	1.329	8.129	0.375	1.512	10.662
Error %	−7.1	−8.2	−8.6	−0.5	2.2	−0.8
Ew_{2-6} simulation [a]	0.433	1.818	12.689	0.332	1.107	4.128
Ew_2 approximation	0.391	1.604	11.234	0.328	1.134	4.719
Error %	−9.7	−11.8	−11.5	−1.2	2.4	14.3
Ew_7 simulation	0.318	1.151	7.416	0.385	1.547	11.105
Ew_7 approximation	0.351	1.329	8.129	0.375	1.512	10.662
Error %	10.4	15.4	9.6	−2.6	−2.3	−4.0
Ew_{8-16} simulation [a]	0.368	1.437	10.728	0.307	1.015	3.888
Ew_8 approximation	0.391	1.604	11.234	0.328	1.134	4.719
Error %	6.2	11.6	4.7	6.8	11.7	21.4
EW	0.362	1.406	8.998	0.362	1.406	8.998
EW in approximation of Kuehn				0.299	1.005	5.097

[a] The results represent mean waiting times averaged over the corresponding group of queues.

identical load have been observed in Table 5 (compare Ew_1 with Ew_7 and Ew_{2-6} with Ew_{8-16}). Even here the differences within the groups of queues $(Q_2,\ldots,Q_6)$ and $(Q_8,\ldots,Q_{16})$ are not very significant, and we have only presented their averages. The mean waiting times tend to increase

Table 6
Comparison of the mean waiting-time approximations (6) (Bux and Truong, exhaustive) and (7) (nonexhaustive) with simulation results; $N=16$ queues, $\Lambda=1$, $\lambda_1=\cdots=\lambda_{16}=\frac{1}{16}$; all service-time distributions negative exponential with $\beta_1=\beta_7$, $\beta_2=\cdots=\beta_6=\beta_8=\cdots=\beta_{16}=\frac{1}{3}\beta_1$

	Exhaustive			Nonexhaustive		
ρ	0.3	0.5	0.8	0.3	0.5	0.8
Ew_1 simulation	0.157	0.602	3.816	0.175	0.679	4.490
Ew_1 approximation	0.161	0.617	3.851	0.170	0.681	4.965
Error %	2.5	2.5	0.9	−2.9	0.3	10.6
Ew_{2-6} simulation [a]	0.167	0.658	4.282	0.163	0.602	3.891
Ew_2 approximation	0.167	0.650	4.201	0.163	0.622	3.724
Error %	0	−1.2	−1.9	0	3.3	−4.3
Ew_7 simulation	0.154	0.584	3.747	0.175	0.675	4.468
Ew_7 approximation	0.161	0.617	3.851	0.170	0.681	4.965
Error %	4.5	5.7	2.8	−2.9	0.9	11.1
Ew_{8-16} simulation [a]	0.167	0.654	4.253	0.161	0.620	3.869
Ew_8 approximation	0.167	0.650	4.201	0.163	0.622	3.724
Error %	0	−0.6	−1.2	1.2	0.3	−3.7
EW	0.165	0.64	4.096	0.165	0.64	4.096
EW in approximation of Kuehn				0.125	0.386	1.713

[a] The results represent mean waiting times averaged over the corresponding group of queues.

slightly for increasing distance from the preceding more heavily loaded queue. In one of the most extreme cases, Ew_{8-16} in Table 5, the following ranges (Ew_8, Ew_{16}) have been observed:

(0.353, 0.380), (1.396, 1.535),
(10.310, 11.099), (0.299, 0.317),
(0.980, 1.048), (3.809, 3.991).

5.2. Remark. The tables confirm that the approximations are less useful when traffic is heavy and at the same time highly asymmetric. A remedy might be to combine the present approximation with a heavy-traffic approximation; we have not pursued this here, as it would interfere with our objective of obtaining mean waiting-time estimates when only utilizations and busy-period lengths can be measured. However, in the nonexhaustive case, we should repeat Kuehn's [8] remark that the estimate of α_{ij} (a probability—see (25)) can become larger than one when traffic is heavy and highly asymmetric. Following Kuehn's suggestion, one can replace α_{ij} by one in such a case. The problem arises (for $\rho = 0.8$) in Tables 1 and 4; the above-mentioned remedy yields the following new mean waiting-time estimates: in Table 1,

$Ew_1 = 4.050, \quad Ew_2 = Ew_3 = 1.924;$

in Table 4,

$Ew_1 = 4.412, \quad Ew_2 = \cdots = Ew_{16} = 1.382.$

Acknowledgment

The authors are indebted to H. David Maxey for suggesting this problem.

References

[1] K. Arndt and H. Sulanke, A queueing system with relative and cyclic priorities, *Elektronische Informationsverarbeitung und Kybernetik* **20** (1984) 423–425.
[2] O.J. Boxma, Two symmetric queues with alternating service and switching times, in: E. Gelenbe, ed., *Proc. Performance '84* (North-Holland, Amsterdam, 1984) 475–490.
[3] W. Bux and H.L. Truong, Mean-delay approximations for cyclic-service queueing systems, *Performance Evaluation* **3** (1983) 187–196.
[4] J.W. Cohen and O.J. Boxma, *Boundary Value Problems in Queueing System Analysis* (North-Holland, Amsterdam, 1983).
[5] J.W. Cohen, *The Single Server Queue* (North-Holland, Amsterdam, 2nd ed., 1982).
[6] R.B. Cooper, Queues served in cyclic order: Waiting times, *Bell Syst. Tech. J.* **49** (1970) 399–413.
[7] L. Kleinrock, *Queueing Systems, Vol. II* (Wiley, New York, 1976).
[8] P.J. Kuehn, Multi-queue systems with nonexhaustive cyclic service, *Bell Syst. Tech. J.* **58** (1979) 671–698.
[9] R.J.T. Morris and Y.T. Wang, Some results for multi-queue systems with multiple cyclic servers, in: H. Rudin and W. Bux, eds., *Performance of Computer-Communications Systems* (North-Holland, Amsterdam, 1984) 245–258.
[10] H. Takagi and L. Kleinrock, Analysis of polling systems, JSI Res. Rept. No. TR87-0002, IBM Japan, January 1985.
[11] R.W. Wolff, Poisson arrivals see time averages, *Oper. Res.* **30** (1982) 223–231.