Waiting-Time Approximations for Cyclic-Service Systems with Switchover Times *

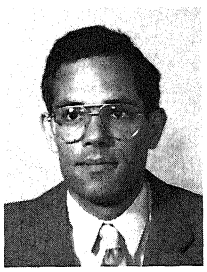
Onno J. Boxma
Centre for Mathematics and Computer Science, P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

Bernd Werner Meister
IBM Research Division, Zürich Research Laboratory, 8803 Rüschlikon, Switzerland

Received 22 December 1985
Revised 7 July 1986

Mean waiting-time approximations are derived for a single-server multi-queue system with nonexhaustive cyclic service. Nonzero switchover times of the server between consecutive queues are assumed. The main tool used in the derivation is a pseudo-conservation law recently found by Watson (1984). The approximation is simpler and, as extensive simulations show, more accurate than existing approximations. Moreover, it gives very good insight into the qualitative behaviour of cyclic-service queueing systems.

Keywords: Waiting-Time Approximation, Cyclic-Service System, Switchover Time, Nonexhaustive Service.



Onno J. Boxma received his Master's degree from Delft Technology 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.

* This paper was presented at Performance '86, the proceedings of which appeared in Performance Evaluation Review **14** (1986) 254–262.

North-Holland
Performance Evaluation 7 (1987) 299–308

1. Introduction

In local area networks with ring or bus topology, medium access control protocols based on token passing have become increasingly popular. Such networks can be modelled as single-server multi-queue systems with a cyclic-service discipline, for example exhaustive, gated or nonexhaustive service. When token rings or buses become longer and/or faster such that propagation delays are noticeable, it becomes important to model the times for passing the token, subsequently called switchover times. In real-time applications, a number of stations, for example measurement devices, is often scanned in a fixed order. Again, switchover times of the server may have an impact on system performance, especially when task switching takes place between the service of two (consecutive) stations.

A basic queueing model for the performance evaluation of such cyclic-service systems with switchover times is investigated in this paper; it will now be described in detail.

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.

1.1. Model description

A single service facility serves N queues $Q_1, Q_2, \ldots, Q_N$ (with infinite buffer capacities) in a cyclic manner. The service discipline considered is ordinary cyclic or nonexhaustive service (sometimes also called chaining, polling, or alternating service): When the server visits a queue, it only serves one customer (if any is present). The switchover times of the server between the ith and $(i+1)$st queue are independent, identically distributed stochastic variables S_i with first moment s_i and variance Ψ_i^2. The mean of the total switchover time during a cycle of the server, s, is given by

$$s = \sum_{i=1}^{N} s_i. \tag{1}$$

Customers arrive at all queues according to independent Poisson processes with rates $\lambda_1, \lambda_2, \ldots, \lambda_N$; the total arrival rate is Λ. Customers which arrive at Q_i are called type-i customers. The service times of type-i customers are independent, identically distributed stochastic variables with distribution $B_i(\cdot)$, with first and second moments β_i and $\beta_i^{(2)}$; the service process is also independent of the arrival process and of the switchover process. The utilization at Q_i, ρ_i, is defined as

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

The total utilization of the server, ρ, is defined as

$$\rho = \sum_{i=1}^{N} \rho_i. \tag{3}$$

It was shown by Kuehn [8] that the following conditions are necessary and sufficient for stability of the system:

$$\rho < 1 \quad \text{and} \quad \max(\lambda_i) s < 1 - \rho. \tag{4}$$

In fact, it is easily shown (cf. [8]) in the stationary situation that the mean cycle time for Q_i, i.e., the mean interarrival time of the server at Q_i, is independent of i, and is given by

$$Ec = \frac{s}{1-\rho}, \tag{5}$$

which immediately implies the necessity of the above stability conditions.

Important performance measures in multi-queue systems are the mean waiting times Ew_i at the individual queues Q_i, $i = 1, 2, \ldots, N$. In the case of nonexhaustive service, which is considered here, the determination of exact values of the mean waiting times is an extremely complicated mathematical problem which could not be solved so far except for a few special cases. A complete exact analysis of the case of $N = 2$ queues without switchover times and of the case of two queues with identical characteristics with switchover times has been presented in [6,5] and [1], respectively (also leading to waiting-time and queue-length distributions). The solution method transforms the problem into a Riemann–Hilbert boundary value problem, and it is not yet clear how it can be generalized to solve the model with more than two queues. Using a different method, Nomura and Tsukamoto [10] and Takagi (cf. [13]) have obtained the exact mean waiting times for a system with an arbitrary number of queues which all have identical characteristics. A heuristic approximation for the case where one queue has a much higher arrival rate than the other queues (which have identical characteristics) can be found in [15]. The excellent survey of Takagi [13] contains several further references. (*Note*: The case in which switchover times, arrival rates, and service time distributions are the same for each queue, will be denoted in the sequel as the completely symmetric case.)

The intractability of the general model has led several authors to the development of mean waiting-time approximations. An important approximation is due to Kuehn [8], who obtains mean waiting-time approximations for nonexhaustive cyclic-service systems with and without switchover times and with batch Poisson input. Earlier references for mean waiting-time approximations can also be found in [8]. An approximation for systems with multiple cyclic servers is given in [9]; the case of cyclic systems with finite-capacity queues has been considered in [14].

In the present paper, the method used in [3] for cyclic service systems *without* switchover times is generalized to obtain simple yet accurate mean waiting-time approximations for the model *with* switchover times. This generalization is made possible by means of a pseudo-conservation law, recently obtained for this model by Watson [16]. The approximation is derived in Section 2. In Section 3, the accuracy of the approximation is assessed. Some conclusions are presented in Section 4.

2. The approximation

We first need some definitions:

- x_i denotes the queue length at 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 ends when the server returns to Q_i (an 'i-cycle');
- 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. An arriving type-i customer first has to wait until the server returns to Q_i and subsequently he has to wait until all customers in front of him have been served. Therefore, the mean waiting time of this customer consists of two parts: a residual cycle rc_i and just as many i-cycles as there are type-i customers waiting; approximately (ignoring dependencies)

$$Ew_i = Erc_i + Ex_i Ec_i. \tag{6}$$

Owing to the fact that Poisson arrivals see time averages (cf. [17]), Ex_i equals the mean number of waiting customers at Q_i at an arbitrary instant of time. This permits the use of Little's formula, yielding

$$Ew_i = \frac{Erc_i}{1-\lambda_i Ec_i}. \tag{7}$$

Similarly to [3], we introduce two approximation assumptions to estimate the two unknowns Ec_i and Erc_i.

Assumption A

$$Ec_i = \frac{\beta_i + s}{1-\rho+\rho_i}, \quad i = 1, 2, \ldots, N. \tag{8}$$

This approximation, which is due to Kuehn [8], can be motivated as follows. An i-cycle consists of a type-i service and, possibly, services of customers of other types, plus the sum of N switchover times.

Define

$$\begin{aligned}\alpha_{ij} &= \Pr(i\text{-cycle contains a type-}j\text{ service})\\ &= E[\text{number of type-}j\text{ services in an }i\text{-cycle}]\\ &\approx \lambda_j Ec_i, \quad j \neq i;\end{aligned} \tag{9}$$

the second equality holds because an i-cycle contains at most one type-i service.

Hence,

$$Ec_i = \beta_i + s + \sum_{j\neq i} \alpha_{ij}\beta_j, \tag{10}$$

which together with (9) yields our assumption (8). Equation (8) is trivially exact for $N = 1$. The approximation in (9) is based on a balance-of-flow argument. It should be very accurate in the completely symmetric case; it should also be very accurate for light traffic, but not for heavy traffic with highly asymmetric arrival rates and/or service demands.

Assumption B. Erc_i is independent of i.

This assumption is trivially exact for $N = 1$ and in the completely symmetric case. In the limiting case $\rho = 0$ it is also true, as can be seen in the following way. Consider, for example, Erc_1:

$$Erc_1 = \sum_{j=1}^{N} \frac{s_j}{s}\left(\tilde{s}_j + s_{j+1} + \cdots + s_N\right),$$

where $\tilde{s}_j$ is the mean residual switchover time between Q_j and Q_{j+1}. Using

$$\tilde{s}_j = \frac{ES_j^2}{2ES_j} = \frac{\Psi_j^2 + s_j^2}{2s_j},$$

it easily follows that

$$Erc_1 = \sum_{j=1}^{N} \frac{\Psi_j^2}{2s} + \tfrac{1}{2}s.$$

This last expression equals the mean residual lifetime of S. For symmetry reasons, hence also $Erc_2, \ldots, Erc_N$ equal the same expression (which could also have been derived from a simple probabilistic argument).

For small values of ρ, the probability that a type-i customer finds other customers present upon his arrival (anywhere in the system) is $O(\rho)$. Furthermore, the mean contribution to Erc_i of work of other customers, arriving between his arrival and the moment at which the server reaches Q_i, is also $O(\rho)$. Hence,

$$Erc_i = \sum_{j=1}^{N} \frac{\Psi_j^2}{2s} + \tfrac{1}{2}s + O(\rho), \quad \rho \to 0.$$

Unlike the case of zero switchover times [3], the $O(\rho)$ term is not completely independent of i but its influence is negligible for small values of ρ, because of the domination of the $O(1)$ term. Therefore, Assumption B should be accurate for low traffic.

The only unknown in expression (7) for Ew_i is

$Erc \equiv Erc_i$, which will be determined by means of the following pseudo-conservation law, due to Watson [16]:

$$\sum_{i=1}^{N} \rho_i(1-\alpha_i)Ew_i = \frac{\rho}{2(1-\rho)} \sum_{j=1}^{N} \lambda_j \beta_j^{(2)} + \frac{\rho}{2s} \sum_{j=1}^{N} \Psi_j^2 + \frac{s}{2(1-\rho)} \sum_{j=1}^{N} \rho_j(1+\rho_j), \tag{11}$$

with α_i defined as

$$\alpha_i = \lambda_i \frac{s}{1-\rho}; \tag{12}$$

α_i is the probability that the server, upon arrival at Q_i, finds at least one customer present.

2.1. Remark. Watson has derived formula (11) by first writing down a set of N recurrence relations for the N generating functions of the joint stationary queue-length distributions at arrival instants of the server at the various queues, and subsequently differentiating these relations twice, after each differentiation taking all generating function parameters equal to one. Finally, he arrives at equation (11) by cleverly eliminating all but N unknowns, which are simply expressed in the Ew_i.

If all switchover times are zero, (11) reduces to Kleinrock's [7] conservation law for M/G/1-type queues: the right-hand side of (11) in this case constitutes the mean waiting time in an M/G/1 queue with arrival rate Λ and with service-time distribution being a weighted sum of the individual service-time distributions. In the present paper, relation (11) is called a pseudo-conservation law because it is an extension of Kleinrock's conservation law, based on the principle of work conservation, to a situation in which work is no longer conserved (see [2] for a probabilistic proof of (11) which also yields an interpretation for the terms in the right-hand side).

Note that the expression in the right-hand side of (11) only involves the first two moments of the service- and switch-over times, and is independent of the polling order of the queues.

An estimate for $Erc \equiv Erc_i$ will be obtained by demanding that the mean waiting-time approximation fulfills the pseudo-conservation law of Watson (note that this immediately implies that the approximation also has the desirable properties of being exact for $N = 1$ and in the completely symmetric case). From (7) and the two above assumptions,

$$Ew_i = Erc \frac{1-\rho+\rho_i}{1-\rho-\lambda_i s}, \quad i = 1, 2, \ldots, N. \tag{13}$$

Substituting these Ew_i into (11) yields

$$Erc = \frac{1-\rho}{(1-\rho)\rho + \sum_{j=1}^{N} \rho_j^2} \times \left[\frac{\rho}{2(1-\rho)} \sum_{j=1}^{N} \lambda_j \beta_j^{(2)} + \frac{\rho}{2s} \sum_{j=1}^{N} \Psi_j^2 + \frac{s}{2(1-\rho)} \sum_{j=1}^{N} \rho_j(1+\rho_j) \right]. \tag{14}$$

Finally, this yields our main result:

$$Ew_i \approx \frac{1-\rho+\rho_i}{1-\rho-\lambda_i s} \frac{1-\rho}{(1-\rho)\rho + \sum_{j=1}^{N} \rho_j^2} \times \left[\frac{\rho}{2(1-\rho)} \sum_{j=1}^{N} \lambda_j \beta_j^{(2)} + \frac{\rho}{2s} \sum_{j=1}^{N} \Psi_j^2 + \frac{s}{2(1-\rho)} \sum_{j=1}^{N} \rho_j(1+\rho_j) \right], \quad i = 1, 2, \ldots, N. \tag{15}$$

2.2. Remark. In the special case of zero switchover times, approximation (15) reduces to the approximation given in [3]:

$$Ew_i \approx \frac{1-\rho+\rho_i}{(1-\rho)\rho + \sum_{j=1}^{N} \rho_j^2} \frac{\rho}{2(1-\rho)} \sum_{j=1}^{N} \lambda_j \beta_j^{(2)}. \tag{16}$$

In the case $N = 1$, it reduces to the exact mean waiting time for an M/G/1 model with vacations (see [12]); in the completely symmetric case, it reduces to the exact result which has been derived in [10,13].

2.3. Remark. According to (15),

$$\frac{Ew_i}{Ew_j} \approx \frac{1-\rho+\rho_i}{1-\rho+\rho_j} \frac{1-\rho-\lambda_j s}{1-\rho-\lambda_i s}. \tag{17}$$

Formulas (15) and (17) suggest that the mean waiting time at a queue is much more sensitive to a change of arrival rate than to a change of mean service time. In particular, two queues in heavy traffic with the same service-time distribution but with slightly different arrival rates may have quite different mean waiting times.

Formula (15) also suggests that the mean switchover time, s, can have a strong influence on the mean waiting times, whereas the means and variances of the individual switchover times are not very critical. These observations will be confirmed by the simulation results presented in Section 3.

2.4. Remark. It is interesting to compare approximation (15) with the mean waiting-time approximation of Bux and Truong [4] for the case of exhaustive service:

$$EW_i \approx \frac{1-\rho_i}{\rho - \sum_{j=1}^{N} \rho_j^2} \times \left[\frac{\rho}{2(1-\rho)} \sum_{j=1}^{N} \lambda_j \beta_j^{(2)} + \frac{s}{2(1-\rho)} \sum_{j=1}^{N} \rho_j (1-\rho_j) \right]. \tag{18}$$

This formula was derived for the case of constant switchover times, and it turns out to satisfy Watson's pseudo-conservation law for the exhaustive service discipline [16]. If the term

$$\frac{\rho}{2s} \sum_{j=1}^{N} \Psi_j^2$$

is added to the expression within brackets in the right-hand side of (18), to take random switchover times into account, then a mean waiting-time approximation will result which is very similar in structure to the approximation (15) for the nonexhaustive service discipline.

Formula (18) reflects the property of the exhaustive service discipline that customers in light-traffic queues usually experience a longer waiting time than customers in heavy-traffic queues: 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. The nonexhaustive service discipline without switchover times, on the other hand, leads to relatively small waiting times at light-traffic queues, as can be seen from (16). For nonzero switchover times, such a general statement cannot be made, but in most cases the behaviour is similar to that for zero switchover times (cf. (15) and (17)).

The derivation of approximation (15) suggests that it will be least accurate in heavy, very asymmetric, traffic. Numerical experiments confirm this (cf. Section 3), disclosing the most sensitive heavy-traffic case: if one or more queues $Q_{i_1}, \ldots, Q_{i_l}$ have relatively large arrival rates, so that these queues become nearly unstable (cf. (4)), approximation (15) has difficulties predicting the mean waiting times at the *other* queues accurately.

Below, a modification of the approximation for the latter queues is suggested. In the original version of this paper, which was presented at Performance '86, a rather complicated rule of thumb was suggested for the application of this modification. The following rule is generally equivalent, but simpler and unambiguous: apply the modification to those queues Q_i for which (cf. (9))

$$\lambda_i \frac{\beta_k + s}{1 - \rho + \rho_k} > 1 \quad \text{for at least one k.}$$

The basic idea of the modification is the following. Remove the queues with a relatively high arrival rate from the system, and enlarge the switchover times to compensate for the service times at the removed queues. The resulting system has a lower and more symmetric traffic load and, hence, approximation (15) becomes much more accurate.

We now present the argument in some more detail. Suppose Q_i is a queue with a relatively high arrival rate λ_i (and hence relatively high α_i). Consider a cyclic queueing system consisting of the $N-1$ queues $Q_1, \ldots, Q_{i-1}, Q_{i+1}, \ldots, Q_N$, with all queues having the same characteristics as in the original model, and with switchover time SW_{i-1} from Q_{i-1} to Q_{i+1} being defined as

$$SW_{i-1} = S_{i-1} + \tau_i + S_i, \tag{19}$$

where

$$\Pr(\tau_i = 0) = 1 - \alpha_i,$$

$$\Pr(\tau_i < t) = 1 - \alpha_i + \alpha_i B_i(t), \quad t > 0,$$

with

$$E\tau_i = \alpha_i \beta_i, \qquad E\tau_i^2 = \alpha_i \beta_i^{(2)}.$$

So, the switchover time from Q_{i-1} to Q_{i+1} is composed of the switchover times from Q_{i-1} to Q_i and from Q_i to Q_{i+1} in the original model, plus a stochastic variable τ_i which takes account of a possible service time in Q_i in the original model. Clearly, when α_i is close to one, $Q_1,\ldots,Q_{i-1}$, $Q_{i+1},\ldots,Q_N$ should behave very similarly in both models. Setting $\alpha_i = 1$ will in most cases yield an upper bound for EW_j, $j \neq i$.

If another queue also has a relatively high arrival rate, the same reasoning should be applied once more, etc. In the finally resulting model, the

Table 1
Comparison of the mean waiting-time approximation (15) with simulation and with Kuehn's approximation; $N = 3$ queues, $\Lambda = 1$, $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$; all service-time distributions negative exponential with $\beta_2 = \beta_3 = \frac{1}{3}\beta_1$

(a) All switchover times equal to 0.05			
ρ	0.3	0.5	0.8
Ew_1 simulation	0.333	1.003	6.80
Ew_1 approximation (15)	0.331	1.041	7.77
Error %	−0.6	3.8	14.3
Ew_1 approximation (Kuehn)	0.317	0.939	6.29
Ew_{2-3} simulation [a]	0.289	0.830	5.38
Ew_2 approximation (15)	0.286	0.780	4.11
Error %	−1.0	−6.0	−23.6
Ew_2 approximation (Kuehn)	0.263	0.645	3.00
(b) All switchover times equal to 0.10			
ρ	0.3	0.5	0.8
Ew_1 simulation	0.506	1.381	10.72
Ew_1 approximation (15)	0.509	1.425	12.90
Error %	0.6	3.2	20.3
Ew_1 approximation (Kuehn)	0.493	1.309	10.64
Ew_{2-3} simulation [a]	0.444	1.155	8.30
Ew_2 approximation (15)	0.439	1.069	6.83
Error %	−1.1	−7.4	−17.7
Ew_2 approximation (Kuehn)	0.415	0.922	5.31
(c) All switchover times negative exponentially distributed with mean 0.05			
ρ	0.3	0.5	0.8
Ew_1 simulation	0.356	1.056	6.90
Ew_1 approximation (15)	0.360	1.071	7.81
Error %	1.1	1.4	13.2
Ew_1 approximation (Kuehn)	0.341	0.961	6.31
Ew_{2-3} simulation [a]	0.314	0.869	5.59
Ew_2 approximation (15)	0.311	0.804	4.13
Error %	−1.0	−7.5	−26.1
Ew_2 approximation (Kuehn)	0.284	0.662	3.01
(d) All switchover times negative exponentially distributed with mean 0.10			
ρ	0.3	0.5	0.8
Ew_1 simulation	0.570	1.394	11.26
Ew_1 approximation (15)	0.570	1.494	13.02
Error %	0.0	7.2	15.6
Ew_1 approximation (Kuehn)	0.545	1.359	10.71
Ew_{2-3} simulation [a]	0.502	1.196	8.60
Ew_2 approximation (15)	0.493	1.121	6.89
Error %	−1.8	−6.3	−19.9
Ew_2 approximation (Kuehn)	0.459	0.960	5.34

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

total utilization will not be very high while the traffic is less asymmetric than in the original system; and Assumptions A and B, in combination with the pseudo-conservation law (11) for this model, will lead to satisfactory mean waiting-time approximations for the queues of the modified model—and hence also for the queues with relatively low arrival rates of the original model.

Summarizing, in the modified approximation the mean waiting times in the queues with rela-

Table 2
Comparison of the mean waiting-time approximation (15) with simulation and with Kuehn's approximation; $N = 3$ queues, $\Lambda = 1$, $\lambda_1 = 0.6$, $\lambda_2 = \lambda_3 = 0.2$; all service-time distributions negative exponential with identical means

(a) All switchover times equal to 0.05			
ρ	0.3	0.5	0.8
Ew_1 simulation	0.304	0.937	9.34
Ew_1 approximation (15)	0.303	0.925	8.30
Error %	−0.3	−1.3	−11.1
Ew_1 approximation (Kuehn)	0.288	0.812	6.31
Ew_{2-3} simulation [a]	0.236	0.581	1.89
Ew_2 approximation (15)	0.238	0.605	1.47
Error %	0.8	4.1	−22.2
Ew_2 approximation (Kuehn)	0.225	0.535	2.47
(b) All switchover times equal to 0.10			
ρ	0.3	0.5	0.8
Ew_1 simulation	0.525	1.510	55.70
Ew_1 approximation (15)	0.528	1.503	51.91
Error %	0.6	−0.5	−6.8
Ew_1 approximation (Kuehn)	0.510	1.356	40.77
Ew_{2-3} simulation [a]	0.370	0.775	2.31
Ew_2 approximation (15)	0.371	0.820	2.22
Error %	0.3	5.8	−3.9
Ew_2 approximation (Kuehn)	0.358	0.750	3.58
(c) All switchover times negative exponentially distributed with mean 0.05			
ρ	0.3	0.5	0.8
Ew_1 simulation	0.333	0.976	9.09
Ew_1 approximation (15)	0.334	0.959	8.36
Error %	0.3	−1.7	−8.0
Ew_1 approximation (Kuehn)	0.313	0.836	6.34
Ew_{2-3} simulation [a]	0.261	0.599	1.92
Ew_2 approximation (15)	0.262	0.628	1.48
Error %	0.4	4.8	−22.9
Ew_2 approximation (Kuehn)	0.245	0.551	2.48
(d) All switchover times negative exponentially distributed with mean 0.10			
ρ	0.3	0.5	0.8
Ew_1 simulation	0.600	1.625	53.87
Ew_1 approximation (15)	0.600	1.590	52.53
Error %	0.0	−2.2	−2.5
Ew_1 approximation (Kuehn)	0.569	1.419	41.19
Ew_{2-3} simulation [a]	0.418	0.825	2.37
Ew_2 approximation (15)	0.421	0.867	2.25
Error %	0.7	5.1	−5.1
Ew_2 approximation (Kuehn)	0.399	0.784	3.60

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

tively low arrival rates are approximated by using (15) in the modified model with fewer queues and different total utilization and switchover times.

3. Comparison with simulation

This section presents a comparison of the mean waiting-time approximation with simulation results, generated with the IBM RESQ2 package [11], and with the well-known approximation of Kuehn [8], together with some general observations. The numerical results are collected in Tables 1 to 6. Representative examples have been chosen to estimate the accuracy of the approximation for different parameter combinations and service-time distributions. Watson's pseudo-conservation law permits a convenient additional validation of the accuracy of the simulation. The simulation results 'fulfill' this law with an error of about 2% for ρ up to 0.5 and an error of about 5% for $\rho = 0.8$.

The relative error of approximation (15) given in the tables, is defined as

$$100\% \frac{(\text{approximation result} - \text{simulation result})}{\text{simulation result}}.$$

Table 3
Comparison of the mean waiting-time approximation (15) with simulation and with Kuehn's approximation; $N = 16$ queues, $\Lambda = 1$, $\lambda_1 = \cdots = \lambda_{16} = \frac{1}{16}$, all service-time distributions negative exponential with $\beta_1 = \beta_7$, $\beta_2 = \cdots = \beta_6 = \beta_8 = \cdots = \beta_{16} = \frac{1}{3}\beta_1$

All switchover times equal to 0.05			
ρ	0.3	0.5	0.8
Ew_1 simulation	0.823	1.697	8.78
Ew_1 approximation (15)	0.831	1.742	10.06
Error %	1.0	2.7	14.6
Ew_1 approximation (Kuehn)	0.796	1.513	7.35
Ew_{2-6} simulation [a]	0.793	1.591	7.98
Ew_2 approximation (15)	0.797	1.590	7.54
Error %	0.5	−0.1	−5.5
Ew_2 approximation (Kuehn)	0.752	1.301	4.58
Ew_7 simulation	0.833	1.720	8.90
Ew_7 approximation (15)	0.831	1.742	10.06
Error %	−0.2	1.3	11.8
Ew_7 approximation (Kuehn)	0.796	1.513	7.35
Ew_{8-16} simulation [a]	0.793	1.591	7.91
Ew_8 approximation (15)	0.797	1.590	7.54
Error %	0.5	−0.1	−4.6
Ew_8 approximation (Kuehn)	0.752	1.301	4.58

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

A more detailed discussion of the results follows. Tables 1 and 2 show results for $N = 3$ queues. In Table 1, the arrival rates are equal but the service times different whereas, in Table 2, different arrival rates but equal service times have been chosen. The tables show that the effect of a higher arrival rate is much stronger than that of a higher mean service time. In Table 1, the mean waiting times at all queues are roughly the same, although the mean service times differ by a factor of three. In Table 2, where the arrival rates differ by a factor of three, this is no longer true in heavy traffic: the mean waiting times at the heavy-traffic queue are much larger than those at the other queues.

Comparing mean waiting times at the low-traffic queues in Tables 1 and 2 (which have the same utilization in both tables) it can be seen that, although the mean service times at Q_2 and Q_3 in Table 1 are *smaller* than those in Table 2, the mean waiting times at Q_2 and Q_3 in Table 1 are *larger*—due to the fact that arrival rates are higher.

Tables 1 and 2 also reveal that the influence of random switchover times in comparison to constant switchover times is only marginal. All the above-mentioned phenomena are correctly predicted by the form of (15) (cf. also Remark 2.3).

Stability condition (4) indicates that $\gamma_i := \rho + \lambda_i s$ must be smaller than one. If γ_i is nearly one, the mean waiting time at Q_i becomes very large even if ρ is considerably smaller than one. An example is the case $\rho = 0.8$ and $s_i = 0.1$ in Table 2, for which $\gamma_1 = 0.98$. The original approximation

Table 4
Comparison of the mean waiting-time approximation (15) with simulation and with Kuehn's approximation; $N = 16$ 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

All switchover times equal to 0.05			
ρ	0.3	0.5	0.8
Ew_{1-4} simulation [a]	0.898	1.929	17.66
Ew_1 approximation (15)	0.897	1.884	16.87
Error %	−0.1	−2.3	−4.2
Ew_1 approximation (Kuehn)	0.863	1.646	12.02
Ew_{5-16} simulation [a]	0.717	1.267	3.57
Ew_5 approximation (15)	0.720	1.307	3.14
Error %	0.4	3.2	−12.0
Ew_5 approximation (Kuehn)	0.689	1.122	3.36

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

(15) yields an error of about 50% for the low-traffic queues. In Table 2, for $\rho = 0.8$, the modified approximation for the low-traffic queues has been used; this way, good results have also been obtained for this extreme case.

As already mentioned in [3], the mean waiting times at different queues with identical characteristics need not be the same, as they depend slightly on the locations of these queues with respect to queues with other traffic patterns. Our approximation does not take this effect into account (and neither does Kuehn's approximation). In our simulations, these differences have been very small. Therefore, mean waiting times at consecutive queues with identical characteristics are only represented by their average in the tables.

Tables 3–6 give results for $N = 16$ queues. Only constant switchover times are considered, as the choice of the switchover time distributions has little bearing on the results. Table 3 is similar to Table 1, but now Q_1 and Q_7 have relatively long mean service times; equation (15) gives a good approximation. In Tables 4 and 5, as in Table 2, different arrival rates are considered. $\gamma_1 = \cdots = \gamma_4 = 0.928$ and $\gamma_1 = 0.896$ in these respective tables if $\rho = 0.8$. The modified approximation has been used for the low-traffic queues in these cases, removing $Q_1, \ldots, Q_4$ and Q_1, respectively.

The combined effect of different service-time distributions and different arrival rates is shown in Table 6. This case is very asymmetric, ρ_1 and ρ_7 being 18 times as large as the other ρ_i. Here, too, in the heavy-traffic case (with $\gamma_1 = \gamma_7 = 0.92$), the modified approximation has been used, but here the improvement is negligible.

Table 5
Comparison of the mean waiting-time approximation (15) with simulation and with Kuehn's approximation; $N = 16$ queues, $\Lambda = 1$, $\lambda_1 = 0.6$, $\lambda_2 = \cdots = \lambda_{16} = \frac{2}{75}$; all service-time distributions negative exponential with identical means

All switchover times equal to 0.01			
ρ	0.3	0.5	0.8
Ew_1 simulation	0.330	1.015	9.71
Ew_1 approximation (15)	0.321	0.996	9.79
Error %	−2.7	−1.9	0.9
Ew_1 approximation (Kuehn)	0.302	0.850	6.93
Ew_{2-16} simulation [a]	0.222	0.495	1.35
Ew_2 approximation (15)	0.224	0.521	1.24
Error %	0.9	5.3	−8.1
Ew_2 approximation (Kuehn)	0.205	0.418	1.21

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

Table 6
Comparison of the mean waiting-time approximation (15) with simulation and with Kuehn's approximation; $N = 16$ 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 $q(1-\exp\{-t/m_1\})+(1-q)(1-\exp\{-t/m_2\})$ with $q = 0.8873$, $m_1 = 0.5635 \times \beta_7$, $m_2 = 4.4365 \times \beta_7$, $\beta_7 = 6\beta_2$, $\beta_7^{(2)} = 5\beta_7^2$

All switchover times equal to 0.05			
ρ	0.3	0.5	0.8
Ew_1 simulation	1.198	3.253	41.26
Ew_1 approximation (15)	1.224	3.271	33.84
Error %	2.2	0.6	−18.0
Ew_1 approximation (Kuehn)	1.153	2.755	23.02
Ew_{2-6} simulation [a]	0.946	2.011	6.27
Ew_2 approximation (15)	0.940	2.027	4.90
Error %	−0.6	0.8	−21.9
Ew_2 approximation (Kuehn)	0.868	1.610	4.72
Ew_7 simulation	1.247	3.335	39.21
Ew_7 approximation (15)	1.224	3.271	33.84
Error %	−1.8	−1.9	−13.7
Ew_7 approximation (Kuehn)	1.153	2.755	23.02
Ew_{8-16} simulation [a]	0.922	1.902	6.17
Ew_8 approximation (15)	0.940	2.027	4.90
Error %	2.0	6.6	−20.6
Ew_8 approximation (Kuehn)	0.868	1.610	4.72

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

4. Conclusions

A simple mean waiting-time approximation for nonexhaustive cyclic-service systems with switchover times has been derived and investigated. The results can be summarized as follows.

- Approximation (15) is constructed in such a way that it fulfills Watson's pseudo-conservation law and, hence, it is in particular exact for the completely symmetric case.
- The approximation gives considerable insight into both the qualitative and quantitative behaviour of the mean waiting times.
- The approximation yields generally better results than known approximations.
- The relative error of the approximation is a few percent for utilizations up to 0.5 in all of our

examples. For traffic patterns which are not too asymmetric, the error is rather small for a utilization of 0.8 (cf. Table 3). The error in (15) becomes larger in cases of strong asymmetry, when some of the queues become nearly unstable. In such cases, the modified approximation described at the end of Section 2 usually leads to considerable improvements.

- The approximation accuracy generally improves with an increasing number of queues, a property which seems to hold for all approximations known. This might be explained by an 'averaging out' effect which stabilizes systems with a large number of queues.
- The error is very insensitive to changes, in mean or distribution, of the switchover times.

Acknowledgement

The authors would like to express their thanks to W. Bux for reviewing the manuscript.

References

[1] O.J. Boxma, Two symmetric queues with alternating service and switching times, in: E. Gelenbe, ed., *Performance '84* (North-Holland, Amsterdam, 1984) 409–431.

[2] O.J. Boxma and W.P. Groenendijk, Pseudo-conservation laws in cyclic-service systems, *Adv. Appl. Probab.* **19** (1987) to appear.

[3] O.J. Boxma and B.W. Meister, Waiting-time approximations in multi-queue systems with cyclic service, *Performance Evaluation* **7** (1) (1987) 59–70.

[4] W. Bux and H.L. Truong, Mean-delay approximations for cyclic-service queueing systems, *Performance Evaluation* **3** (3) (1983) 187–196.

[5] J.W. Cohen and O.J. Boxma, *Boundary Value Problems in Queueing System Analysis* (North-Holland, Amsterdam, 1983).

[6] M. Eisenberg, Two queues with alternating service, *SIAM J. Appl. Math.* **36** (1979) 287–303.

[7] L. Kleinrock, *Queueing Systems, Vol. II* (Wiley, New York, 1976).

[8] P.J. Kuehn, Multi-queue systems with non-exhaustive 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] M. Nomura and K. Tsukamoto, Traffic analysis on polling systems, *Trans. Inst. Electronics & Communication Engineers Japan* **J61-B** (7) (1978) 600–607 (in Japanese).

[11] C.H. Sauer and E.A. MacNair, *Simulation of Computer Communication Systems* (Prentice-Hall, Englewood Cliffs, NJ, 1983).

[12] C.E. Skinner, A priority queueing system with server-walking time, *Oper. Res.* **15** (1967) 278–285.

[13] H. Takagi, *Analysis of Polling Systems* (MIT Press, Cambridge, MA, 1986).

[14] P. Tran-Gia and T. Raith, Multiqueue systems with finite capacity and nonexhaustive cyclic service, in: T. Hasegawa, H. Takagi and Y. Takahashi, eds., *Proc. Computer Networking and Performance Evaluation* (North-Holland, Amsterdam, 1986) 213–227.

[15] Ding-gong Wang, *Analysis of Cyclic Service Multi-Queue Systems for Ring Type Local Area Networks*, Thesis, Dept. of Electrical and Computer Engineering, North Carolina State Univ., Raleigh, NC, 1986.

[16] K.S. Watson, Performance evaluation of cyclic service strategies—A survey, in: E. Gelenbe, ed., *Performance '84* (North-Holland, Amsterdam, 1984) 521–533.

[17] R.W. Wolff, Poisson arrivals see time averages, *Oper. Res.* **30** (1982) 223–231.