Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

W.P. Groenendijk

Waiting-time approximations for cyclic-service systems with mixed service strategies

Department of Operations Research and System Theory Report OS-R8802 January

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

Waiting-Time Approximations for Cyclic-Service Systems with Mixed Service Strategies

W.P. Groenendijk
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

An approximation method for mean waiting times in cyclic-service systems with a mixture of exhaustive/gated/1-limited service strategies is presented. The method provides a unification and generalization of some known good approximations for single strategies. The results of an exact analysis of the two-queue case where one queue is served exhaustively and the other queue 1-limited, support the approximation idea.

1980 Mathematics Subject Classification: 60K25, 68M20. 69C40
Key Words & Phrases: cyclic service, mixed service strategies, (pseudo-) conservation law.
Note: This report will be submitted for publication elsewhere.

1. Introduction

Token-passing protocols are becoming increasingly popular for application in Local Area Networks (LAN's) with a ring or bus topology. With these protocols, rather than to attach all stations to a single subnetwork such as a ring, for various reasons it is preferred to interconnect several subnetworks. When there is no need for protocol conversion in the interconnection a *bridge* is used: a dedicated station in a LAN providing interconnection between subnetworks, at a very low level of architecture. When a LAN is to be connected to a network of an other type, a *gateway* must be used: from the point of view we shall adopt here, a gateway is a dedicated station in a LAN providing the interconnection between networks, where there may be a need for protocol conversion. Of course a gateway has to be substantially more complicated and operates at a higher level of architecture than a bridge does, and hence is usually much slower compared to a bridge. So, essentially, we distinguish between three types of stations in a token-passing (sub-)network: ordinary stations, bridge stations, and gateway stations. It is clear that, since these types of stations each have different characteristics, it may be advantageous to assign them different priorities with respect to the communication protocol.

In general, the performance of polling schemes, of which the token-passing protocol is an example, can be analyzed by studying single-server, multi-queue queueing systems. For example in a token ring LAN the common transmission medium may be represented by the single server, and the workstations attached to the ring by the queues. The circulation of the token along the ring implies that the stations are polled in a cyclic order.

This paper is concerned with the waiting-time process at the various queues of a polling system as described above. Let us first present a more detailed *model description.* We consider a system of N queues, $Q_1,...,Q_N$, served sequentially in cyclic order by a single server S. Messages arrive at all queues according to independent Poisson processes with arrival intensities $\lambda_1,\lambda_2,\ldots,\lambda_N$. The switch-over times of the server between the i-th and $(i+1)$-th queue are independent, identically distributed random variables, with first moment s_i, second moment $s_i^{(2)}$ and Laplace Stieltjes Transform (LST) $\sigma_i(\cdot)$. The first moment of the *total* switch-over time during a cycle of the server is denoted by s, its second moment by $s^{(2)}$. The service times of type-i messages (messages enqueued in Q_i) are independent, identically distributed random variables with first moment β_i, second moment $\beta_i^{(2)}$ and LST $\beta_i(\cdot)$. We assume that the arrival process, the service process and the switch-over process are mutually independent. The offered traffic at Q_i, ρ_i, and the total offered traffic, ρ, are defined as:

$$\rho_i := \lambda_i \beta_i, \quad i = 1,...,N; \qquad \rho := \rho_1 + \rho_2 + ... + \rho_N.$$

For the service strategies at the queues we consider three possibilities, which differ in the number of messages which may be served in a queue during a visit of server S to that queue. Assume that S visits Q_i. When Q_i is empty, S immediately begins to switch to Q_{i+1}. Otherwise S acts as follows, depending on the service strategy at Q_i:

1) Exhaustive service: S serves type-i messages until Q_i is empty,
2) 1-Limited service: S serves one type-i message,
3) Gated service: S serves exactly those type-i messages present upon his arrival at Q_i (a gate closes upon his arrival).

In a queueing model of a token ring LAN where some of the stations act as bridges or gateways, it may be natural to assign a higher priority to the queues which represent these dedicated stations than to the other queues at the ring. In order to incorporate this in our model, we shall allow mixed cyclic-service strategies (e.g., exhaustive at Q_1 and Q_3, gated at Q_4 and Q_5 and 1-limited at Q_2 and $Q_6,...,Q_N$). The service strategy at the ordinary queues usually is 1-limited, but at the dedicated queues one may use the exhaustive or gated service strategy to model the preferential treatment received by these queues.

The main performance measures of interest in cyclically served queueing systems are the mean waiting times of messages at the various queues. Unfortunately, explicit analytical results for even mean waiting times in these systems are only available in some exceptional cases. Furthermore, even in the case that it is possible to calculate the mean waiting times explicitly, as in a system with (a mixture of) exhaustive or gated service at all queues, the determination of the N mean waiting times requires the solution of at least $O(N^2)$ linear equations, cf. [8], [14]. For a system of queues with 1-limited service, only the two-queue case has been explicitly solved ([3]), requiring the solution of a Riemann-Hilbert boundary value problem of mathematical physics. So there definitely is a need for approximations, and already a vast literature has appeared on this subject, cf. [4], [5], [7], [9], [11] and [13]. Most of the recent approximations are based on the recently discovered *pseudo-conservation laws,* exact expressions for weighted sums of the mean waiting times. In [8] such a pseudo-conservation law has been proven for a system of queues with either exhaustive, or gated service at all queues, and in [15] also for a system of queues with 1-limited service. In [1], [2], these laws have been generalized by allowing a mixture of (four) different service strategies at different queues. For the model described above, the pseudo-conservation law derived in [1] reduces to the following expression. Denote by e the group of e(xhaustive) queues, by g the group of g(ated) queues, and by $1l$ the group of 1-l(imited) queues. Assume that $\rho<1$, and that, for all $j \in 1l$, $\rho+\lambda_j s<1$. This ensures that the stationary distributions of the waiting times exist. Denote by EW_i the mean waiting time at Q_i. Then

$$\sum_{i\in e}\rho_i EW_i + \sum_{i\in g}\rho_i EW_i + \sum_{i\in 1l}\rho_i[1-\frac{\lambda_i s}{1-\rho}]EW_i = \tag{1.1}$$

$$\rho\sum_i \frac{\lambda_i \beta_i^{(2)}}{2(1-\rho)} + \rho\frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)}[\rho^2 - \sum_{i\in e}\rho_i^2 + \sum_{i\in g,\,1l}\rho_i^2].$$

In this paper we shall present a very straightforward approximation for the mean waiting times in cyclic-service systems with a mixture of exhaustive/gated/1-limited service strategies. The approximation unifies and generalizes existing approximations for single strategies. The paper is organized as follows: in Section 2 the results of an exact analysis of the two-queue E/1L model are presented; in Section 3 the approximation is introduced; it is compared with simulation. The results from Section 2 support the concept. The approximation appears to be very accurate for mixtures of the gated and exhaustive service strategies. If also the 1-limited strategy is allowed, the approximation becomes worse when one or more of the 1-limited queues becomes heavily loaded ($\rho+\lambda_i s$ close to one). In such a "heavy-traffic" case, the approximation should be refined. Such a refinement is suggested in Section 4; it will be discussed in detail in a forth-coming paper [10].

2. Exact mean waiting-time results for the two-queue E/1L model

In this section we consider a model consisting of two queues, Q_1 and Q_2, where the service strategy is exhaustive at Q_1 and 1-limited at Q_2. For a description of the model and the model parameters we refer to Section 1. Note that in case the switch-over times are zero the model may be identified with the non-preemptive priority M/G/1 queue with two types of customers.

As a notational convention we shall write $\alpha\beta(s)$ for the product of the Laplace Stieltjes Transforms $\alpha(s)$ and $\beta(s)$. Some further notation:

$$r_i := \lambda_i / \lambda;$$

$$x := \lambda(1-r_1 z_1 - r_2 z_2).$$

Let $F_j(\cdot,\cdot)$ denote the joint generating function of the queue-length stationary state distribution at arrival instants of the server at Q_j, $j=1,2$. It is easily found, that

$$F_1(z_1,z_2) = \frac{\sigma_2\beta_2(x)}{z_2} F_2(z_1,z_2) + \sigma_2(x)\frac{z_2-\beta_2(x)}{z_2} F_2(z_1,0); \tag{2.1}$$

$$F_2(z_1,z_2) = F_1(\gamma_1(\lambda_2(1-z_2)),z_2)\sigma_1(x), \tag{2.2}$$

where $\gamma_1(\cdot)$ is the LST of the length of the busy period at Q_1 starting with one customer present. For notational convenience we shall write

$$\delta_1(z_2) := \gamma_1(\lambda_2(1-z_2)),$$

and

$$\tilde{x} := \lambda(1-r_1\delta_1(z_2)-r_2 z_2).$$

Taking $z_1 = \delta_1(z_2)$ in (2.1) yields:

$$F_1(\delta_1(z_2),z_2) = \frac{\sigma_2\beta_2(\tilde{x})}{z_2} F_2(\delta_1(z_2),z_2) + \sigma_2(\tilde{x})\frac{z_2-\beta_2(\tilde{x})}{z_2} F_2(\delta_1(z_2),0). \tag{2.3}$$

From (2.2) and (2.3) we obtain

$$F_2(z_1,z_2) = \sigma_1(x)\frac{\sigma_2\beta_2(\tilde{x})}{z_2} F_2(\delta_1(z_2),z_2) + \sigma_1(x)\sigma_2(\tilde{x})\frac{z_2-\beta_2(\tilde{x})}{z_2} F_2(\delta_1(z_2),0). \tag{2.4}$$

Again taking $z_1 = \delta_1(z_2)$, now in (2.4), yields:

$$F_2(\delta_1(z_2),z_2) = \frac{\sigma_1\sigma_2\beta_2(\tilde{x})}{z_2} F_2(\delta_1(z_2),z_2) + \sigma_1\sigma_2(\tilde{x})\frac{z_2-\beta_2(\tilde{x})}{z_2} F_2(\delta_1(z_2),0), \tag{2.5}$$

which may be written as

$$F_2(\delta_1(z_2),z_2) = \sigma_1\sigma_2(\tilde{x})\frac{z_2-\beta_2(\tilde{x})}{z_2-\sigma_1\sigma_2\beta_2(\tilde{x})} F_2(\delta_1(z_2),0). \tag{2.6}$$

Note that, cf. (2.2):

$$F_2(z_1,0) = F_1(\gamma_1(\lambda_2),0)\sigma_1(\lambda(1-r_1 z_1)). \tag{2.7}$$

Of course we have, cf. for instance [1],

$$F_2(1,0) = \frac{1-\rho-\lambda_2 s}{1-\rho}. \tag{2.8}$$

Hence, from (2.7) and (2.8),

$$F_2(\delta_1(z_2),0) = \frac{1-\rho-\lambda_2 s}{1-\rho}\frac{\sigma_1(\lambda(1-r_1\delta_1(z_2)))}{\sigma_1(\lambda_2)}. \tag{2.9}$$

Finally, from (2.4), (2.6) and (2.9):

$$F_2(z_1,z_2) = \sigma_1(x)\sigma_1(\lambda(1-r_1\delta_1(z_2)))\sigma_2(\tilde{x})\frac{z_2-\beta_2(\tilde{x})}{z_2-\sigma_1\sigma_2\beta_2(\tilde{x})}\frac{1-\rho-\lambda_2 s}{1-\rho}\frac{1}{\sigma_1(\lambda_2)}. \tag{2.10}$$

The generating functions of the queue lengths at polling instants and the Laplace Stieltjes Transforms of the waiting time are related as follows (cf. Watson [15]):

$$E[e^{-\lambda_1(1-z_1)W_1}] = \frac{1-\lambda_1\beta_1}{\frac{d}{dz_1}F_1(z_1,1)|_{z_1=1}}\frac{1-F_1(z_1,1)}{\beta_1(\lambda_1(1-z_1))-z_1}, \tag{2.11}$$

$$E[e^{-\lambda_2(1-z_2)W_2}] = \frac{F_2(1,z_2)-F_2(1,0)}{z(1-F_2(1,0))}.$$

A straightforward calculation now yields the Laplace Stieltjes Transforms of the waiting times at Q_1 and Q_2 respectively:

$$E[e^{-vW_1}] = \frac{\sigma_1\sigma_2\beta_2(v)-1}{\lambda_1-v-\lambda_1\beta_1(v)}\frac{1-\rho}{s} + \tag{2.12}$$

$$\frac{1-\rho-\lambda_2 s}{s}\frac{\sigma_1(\lambda_2+v)\sigma_2(v)}{\sigma_1(\lambda_2)}\frac{1-\beta_2(v)}{\lambda_1-v-\lambda_1\beta_1(v)}.$$

$$E[e^{-vW_2}] = \frac{1-\rho-\lambda_2 s}{s}\sigma_1(v)\frac{\sigma_1(\lambda(1-r_1\gamma_1(v)))}{\sigma_1(\lambda_2)}\sigma_2(\lambda(1-r_1\gamma_1(v)-r_2(1-\frac{v}{\lambda_2})))\times$$

$$\frac{\lambda_2-v-\lambda_2\beta_2(\lambda(1-r_1\gamma_1(v)-r_2(1-\frac{v}{\lambda_2})))}{\lambda_2-v-\lambda_2\sigma_1\sigma_2\beta_2(\lambda(1-r_1\gamma_1(v)-r_2(1-\frac{v}{\lambda_2})))}\frac{1}{\lambda_2-v} - \frac{1}{\lambda_2-v}\frac{1-\rho-\lambda_2 s}{s}.$$

The mean waiting times are given by:

$$EW_1 = \frac{\lambda_1\beta_1^{(2)}+\lambda_2\beta_2^{(2)}}{2(1-\rho_1)} + \frac{1-\rho}{1-\rho_1}(\frac{s^{(2)}}{2s}+\beta_2) + \frac{1-\rho-\lambda_2 s}{s(1-\rho_1)}(\frac{\sigma_1'(\lambda_2)}{\sigma_1(\lambda_2)}-s_2)\beta_2; \tag{2.13}$$

$$EW_2 = \frac{\lambda_1\beta_1^{(2)}+\lambda_2\beta_2^{(2)}}{2(1-\rho_1)(1-\rho-\lambda_2 s)} + \frac{1-\rho}{1-\rho_1}\frac{1}{1-\rho-\lambda_2 s}(\frac{s^{(2)}}{2s}+\beta_2) -$$

$$\frac{\rho_1(1-\rho)}{\lambda_2 s(1-\rho_1)}(\frac{\sigma_1'(\lambda_2)}{\sigma_1(\lambda_2)}-s_2) - \frac{\rho}{\lambda_2}.$$

If we take $\sigma_1(v)=\exp[-s_1 v]$, we obtain a special case of a model previously studied by Skinner [12]. In this particular case

$$EW_1 = (1-\rho-\lambda_2 s)EW_2, \tag{2.14}$$

while in general we have

$$EW_1 \geqslant (1-\rho-\lambda_2 s)EW_2. \tag{2.15}$$

We shall come back to equality (2.14) later.

3. The approximation method

The approximation consists of two steps. The first step is to express the mean waiting time at Q_i in $E\mathbf{rc}_i$: the expected time until the next arrival of S at Q_i. Below we will show that for Q_i gated:

$$E\mathbf{W}_i = (1+\rho_i)E\mathbf{rc}_i; \tag{3.1}$$

while for Q_i exhaustive:

$$E\mathbf{W}_i \approx (1-\rho_i)E\mathbf{rc}_i. \tag{3.2}$$

For Q_i 1-limited we take:

$$E\mathbf{W}_i \approx \frac{E\mathbf{rc}_i}{1-\lambda_i E\mathbf{C}_{b,i}}, \tag{3.3}$$

where $E\mathbf{C}_{b,i}$ denotes the mean cycle time at Q_i, given the cycle contains a service at Q_i. An approximation for $E\mathbf{C}_{b,i}$ will be presented in the sequel.

The second step is to assume $E\mathbf{rc}_i$ is the same for all i: $E\mathbf{rc}_i = E\mathbf{rc}$. We then substitute the expressions for the mean waiting times into the pseudo-conservation law (1.1) and thus obtain the following approximation for the mean residual cycle time:

$$E\mathbf{rc} \approx \left[\rho\sum_i \frac{\lambda_i\beta_i^{(2)}}{2(1-\rho)} + \rho\frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)}(\rho^2 - \sum_{i\in e}\rho_i^2 + \sum_{i\in g,\,1l}\rho_i^2) \right] \times \tag{3.4}$$

$$\times \left[\sum_{i\in e}\rho_i(1-\rho_i) + \sum_{i\in g}\rho_i(1+\rho_i) + \sum_{i\in 1l}\rho_i\left(1-\frac{\lambda_i s}{1-\rho}\right)\frac{1}{1-\lambda_i E\mathbf{C}_{b,i}} \right]^{-1}.$$

Substitution of $E\mathbf{rc}$ in the above formulas for $E\mathbf{W}_i$ yields the approximation for the mean waiting times at the various queues.

Now that we have sketched the global idea we shall take a somewhat more detailed look at the approximation. We define the cycle time $\mathbf{C}_i$ for Q_i as the time between two successive arrivals of S at Q_i. It is easily seen that $E\mathbf{C}_i$ is independent of i. The visit time $\mathbf{V}_i$ of S for Q_i is the time between the arrival of S at Q_i and his subsequent departure from that queue. The intervisit time, $\mathbf{I}_i$, for Q_i is defined as: $\mathbf{I}_i = \mathbf{C}_i - \mathbf{V}_i$. Below we will discuss how we can express the mean waiting times at the various queues in $E\mathbf{rc}_i$, the mean residual cycle time for Q_i.

1) *Gated service strategy at Q_i.* In this case the mean waiting time of an arriving tagged type-i message consists of two components. First a mean residual type-i cycle time, because due to the gating mechanism a message is never served in the cycle in which it arrived. Secondly, the mean time from the instant the server arrives at Q_i, till the moment the tagged message starts to receive service. This component consists of all the service times from the messages that arrived after the start of the previous visit period, but before the tagged message arrived. Hence it is given by $\lambda_i E\mathbf{rc}_i\beta_i$ $(=\rho_i E\mathbf{rc}_i)$.

2) *Exhaustive service strategy at Q_i.* For this case we shall prove that

$$E\mathbf{W}_i = (1-\rho_i)E\tilde{\mathbf{rc}}_i, \tag{3.5}$$

where $E\tilde{\mathbf{rc}}_i$ is the mean residual cycle time at Q_i that now corresponds to a type-i cycle, $\tilde{\mathbf{C}}_i$, being defined as the time between two successive *departures* of S from Q_i. After we have proven (3.5) we shall just ignore the (small) difference between $E\mathbf{rc}_i$ and $E\tilde{\mathbf{rc}}_i$, hence arriving at relation (3.2).
With this different definition of a cycle, denote the LST of the intervisit time, visit time and cycle time for Q_i by $\tilde{I}_i(\cdot)$, $\tilde{V}_i(\cdot)$, and $\tilde{C}_i(\cdot)$, respectively. As in Bux & Truong [5], it may be proven that

$$\tilde{C}_i(s) = \tilde{I}_i(s+\lambda_i-\lambda_i\gamma_i(s)), \tag{3.6}$$

where $\gamma_i(\cdot)$ is the LST of the length of the busy period at Q_i starting with one message present.

Hence we obtain the following result for the second moment of the intervisit time for Q_i:

$$E\tilde{\mathbf{I}}_i^2 = (1-\rho_i)^2 E\tilde{\mathbf{C}}_i^2 - \frac{\lambda_i \beta_i^{(2)}}{1-\rho} s. \tag{3.7}$$

Furthermore we have,

$$E\mathbf{W}_i = \frac{E\tilde{\mathbf{I}}_i^2}{2E\tilde{\mathbf{I}}_i} + \frac{\lambda_i \beta_i^{(2)}}{2(1-\rho_i)}, \tag{3.8}$$

cf. for instance Doshi [6]. Substituting (3.7) into (3.8) and using $E\tilde{\mathbf{rc}}_i = E\tilde{\mathbf{C}}_i^2 / 2E\tilde{\mathbf{C}}_i$, now yields (3.5). Note that the derivation above is completely independent of the service strategies at the other queues.

3) *1-Limited service strategy at Q_i.* Denote by $\mathbf{X}_i$ the number of waiting messages at Q_i just before the arrival of a type-i message, and by $\mathbf{C}_{b,i}$ 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. Similar reasoning as in [4] leads to the approximation

$$E\mathbf{W}_i \approx E\mathbf{rc}_i + E\mathbf{X}_i E\mathbf{C}_{b,i}. \tag{3.9}$$

This is an approximation, because the use of Wald's lemma, leading to the last term in the right-hand side of (3.9), is not justified here. Using the fact that Poisson arrivals see time averages, and Little's formula, we may write $E\mathbf{X}_i = \lambda_i E\mathbf{W}_i$. Substitution of this in (3.9) leads to (3.3).

For the approximation of $E\mathbf{C}_{b,i}$, note that it consists of a type-i service, and, possibly, services of messages at the other queues, plus the total switch-over time. It is now assumed that the number of type-j messages *arriving* during $E\mathbf{C}_{b,i}$, equals the number of type-j messages *departing* during $E\mathbf{C}_{b,i}$ (balance of flow). Hence we approximate the average number of type-j services during $E\mathbf{C}_{b,i}$ by $\lambda_j E\mathbf{C}_{b,i}$. The idea of this assumption is due to Kuehn [11]. However, for the case that Q_j has a 1-limited service strategy, for obvious reasons we bound $\lambda_j E\mathbf{C}_{b,i}$ by 1. So our approximation for $E\mathbf{C}_{b,i}$ will be calculated from the equation

$$E\mathbf{C}_{b,i} = \beta_i + s + \sum_{j \in e,g} \lambda_j E\mathbf{C}_{b,i} \beta_j + \sum_{j \in 1l} \min(1, \lambda_j E\mathbf{C}_{b,i}) \beta_j. \tag{3.10}$$

If $\lambda_j E\mathbf{C}_{b,i} \leqslant 1$ for all queues Q_j with 1-limited service, then (3.10) simplifies to

$$E\mathbf{C}_{b,i} = \frac{\beta_i + s}{1-\rho+\rho_i}, \tag{3.11}$$

and (3.3) in that case reduces to:

$$E\mathbf{W}_i \approx \frac{1-\rho+\rho_i}{1-\rho-\lambda_i s} E\mathbf{rc}_i. \tag{3.12}$$

In general, the assumption that $E\mathbf{rc}_i = E\mathbf{rc}$ for all i is fairly accurate. However cf. Everitt [7], for a discussion of some factors which may influence the accuracy.

For the two-queue case from Section 2, it was noted that (cf. (2.14)), if the switch-over time from Q_1 to Q_2 is a constant, $E\mathbf{W}_1 = (1-\rho-\lambda_2 s) E\mathbf{W}_2$. It may be easily seen that the approximation for this case is exact. Furthermore the approximation is exact in the cases where $N=1$, and also in the cases where we have a symmetrical system, i.e. all N queues having the same characteristics. In the case that all queues have exhaustive, or all queues have gated service, the approximation reduces to an approximation by Everitt [7]. In the case that all queues have a 1-limited service strategy, it closely resembles the approximation of Boxma & Meister [4].

RESULTS

We shall now present some results of the approximation. The results are compared with simulation. In all cases considered, the service-time distributions are taken negative exponential. Furthermore the switch-over time distributions are taken deterministic; this is not a serious limitation, since varying the distribution of the switch-over times has only a marginal effect in most cases. In Tables 1-12, the individual switch-over times are taken equal to 0.1, in Tables 13 and 14 equal to 0.05.

Tables 1-12 are all for 3-queue models with a total traffic intensity of 0.3, 0.5 and 0.8. Tables 1-6 present cases where $\lambda_1 = 0.6$, $\lambda_2 = \lambda_3 = 0.2$; $\beta_1 = \beta_2 = \beta_3$. Tables 7-12 present cases where $\lambda_i = 1/3$, $i = 1,2,3$; $\beta_2 = \beta_3 = (1/3)\beta_1$.

The relative error presented in the tables below, is defined as

$$\frac{\textit{approximation result} - \textit{simulation result}}{\textit{simulation result}} 100\%.$$

Tables 13-14 treat 12-queue models, all with a total traffic intensity of 0.5. In cases 1 to 3 of Table 13 we have asymmetric arrival streams, while all β_i are equal to 0.5. In case 1, $\lambda_1 = 0.56$, $\lambda_2 = ... = \lambda_{12} = 0.04$. In case 2, $\lambda_4 = 0.56$, $\lambda_1 = ... = \lambda_3 = \lambda_5 = ... = \lambda_{12} = 0.04$. In case 3, $\lambda_8 = 0.56$, $\lambda_1 = ... = \lambda_7 = \lambda_9 = ... = \lambda_{12} = 0.04$ (note that $\rho + \lambda_8 s = 0.836$, so that Q_8 is quite heavily "loaded"). In cases 1 to 3 of Table 14 one queue has a larger mean service time than the other queues; $\lambda_1 = ... = \lambda_{12} = 1/12$. In case 1, $\beta_1 = 3.36$, $\beta_2 = ... = \beta_{12} = 0.24$, in case 2, $\beta_4 = 3.36$, $\beta_1 = ... = \beta_3 = \beta_5 = ... = \beta_{12} = 0.24$, and in case 3, $\beta_8 = 3.36$, $\beta_1 = ... = \beta_7 = \beta_9 = ... = \beta_{12} = 0.24$.

REMARK Some improvements for the case that there is a heavily loaded 1-limited queue in the system may be made by a modifying procedure as described in [4], leading to a reduced system. Note that we can improve somewhat upon this procedure by substituting the mean waiting times for the reduced system in the pseudo-conservation law for the original system, hence obtaining a better approximation for the mean waiting time of the (omitted) heavy traffic queue. The modifying procedure may lead to considerable improvement: for instance for Table 3 the modification applied to the case $\rho = 0.8$ yielded errors of 2.6, 4.0 and 0.7 percent for the mean waiting times at Q_1, Q_2 and Q_3 respectively, instead of the old values -0.6, 20.3 and -17.2 percent (the fact that all three errors are positive indicates an inaccuracy in the simulation); for Table 4 we obtain errors of -3.1, 6.1 and -1.2 percent instead of -6.1, 31.3 and 22.6.

CONCLUSIONS:

A very simple and straightforward mean waiting-time approximation for cyclic-service systems with mixed service strategies has been derived and investigated. The approximation unifies and generalizes some previous approximations. It is constructed in such a way that it fulfills the pseudo-conservation law. Despite its simplicity, the approximation provides considerable insight into both the qualitative and quantitative behavior of the mean waiting times. For systems where some of the queues have a 1-limited service strategy, the approximation clearly is a low- or medium-traffic approximation. The approximation is better for the case that we have a strong asymmetry with respect to the arrival process, than for the case that we have a strong asymmetry with respect to the service times. When only mixtures of exhaustive and gated queues are involved, the approximation is good over the whole range of admissible traffic intensities.

4. Refinement of the approximation

As may be seen from the results presented in the previous section, the approximation has difficulties handling the case of heavy, asymmetric traffic in combination with the presence of 1-limited queues. This is especially due to approximation (3.3). Comparison with exact results from [3] reveals that (3.3) is quite close in most cases, but there are a few exceptions, with errors up to 20 percent. In [10] an approximation will be presented which adds a correction term A_i to formula (3.3), so that we get

$$E\mathbf{W}_i = \frac{E\mathbf{rc}_i}{1-\lambda_i E\mathbf{C}_{b,i}} + A_i \cdot \tag{4.1}$$

The correction term is derived using similar arguments as in [13] (which only considers the 1-limited strategy). It appears that this approximation is much less sensitive to the traffic intensity and can better cope with strong asymmetry than the one described in this paper. It is however more complicated and much less transparent.

References

[1] Boxma, O.J., Groenendijk, W.P. (1986) Pseudo-conservation laws in cyclic-service systems. Report OS-R8606, Centre for Mathematics and Computer Science, Amsterdam. To appear in: *J. Appl. Prob.* Vol. **24**, 1987.

[2] Boxma, O.J., Groenendijk, W.P. (1987) Waiting times in discrete-time cyclic-service systems, Report OS-R8707, Centre for Mathematics and Computer Science, Amsterdam. To appear in: *IEEE Trans. on Communications,* January 1988.

[3] Boxma, O.J., Groenendijk, W.P. (1987) Two queues with alternating service and switching times, Report OS-R8712, Centre for Mathematics and Computer Science, Amsterdam.

[4] Boxma, O.J., Meister, B. (1986) Waiting-time approximations for cyclic-service systems with switch-over times. *Performance Evaluation Review* **14**, 254-262.

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

[6] Doshi, B.T. (1986) Queueing systems with vacations - a survey, *Queueing Systems* **1**, 29-66.

[7] Everitt, D. (1986) Simple approximations for token rings. *IEEE Trans. on Communications,* Vol. COM-**34**, 719-721.

[8] Ferguson, M.J., Aminetzah, Y.J. (1985) Exact results for nonsymmetric token ring systems. *IEEE Trans. on Communications,* Vol. COM-**33**, 223-231.

[9] Fuhrmann, S.W., Wang, Y.T. (1987) Mean waiting time approximations of cyclic service systems with limited service. To appear in the proceedings of Performance '87, Brussels.

[10] Groenendijk, W.P. (1987) To appear.

[11] Kuehn, P.J. (1979) Multiqueue systems with nonexhaustive cyclic service, *The Bell System Techn. Journ.* **58**, 671-698.

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

[13] Srinivasan, M.M. (1986) An approximation for mean waiting times in cyclic server systems with non-exhaustive service. Technical report 86-36, The University of Michigan, Ann Arbor.

[14] Takagi, H. (1987) Analysis of polling models with a mix of exhaustive and gated service disciplines. Report Tokyo Research Laboratory, IBM Japan.

[15] Watson, K.S. (1985) Performance evaluation of cyclic service strategies - a survey. In: *Performance '84,* ed. E. Gelenbe, North-Holland Publ. Cy., Amsterdam, 521-533.

Table 1

Q_i	EW_i	$\rho=.3$	$\rho=.5$	$\rho=.8$
e	simul.	0.286	0.58	1.64
	approx.	0.288	0.59	1.91
	error	0.6	2.3	15.9
1/	simul.	0.419	1.17	9.96
	approx.	0.417	1.15	9.42
	error	-0.6	-1.9	-5.4
1/	simul.	0.418	1.17	9.96
	approx.	0.417	1.15	9.42
	error	-0.3	-1.7	-5.4

$\lambda_1=0.6$, $\lambda_2=\lambda_3=0.2$, $\beta_1=\beta_2=\beta_3$

Table 2

Q_i	EW_i	$\rho=.3$	$\rho=.5$	$\rho=.8$
g	simul.	0.381	0.86	2.99
	approx.	0.383	0.89	3.88
	error	0.5	4.6	29.9
1/	simul.	0.393	0.99	8.78
	approx.	0.386	0.94	6.74
	error	-1.9	-5.8	-23.3
1/	simul.	0.392	1.00	8.78
	approx.	0.386	0.94	6.74
	error	-1.7	-6.2	-23.2

$\lambda_1=0.6$, $\lambda_2=\lambda_3=0.2$, $\beta_1=\beta_2=\beta_3$

Table 3

Q_i	EW_i	$\rho=.3$	$\rho=.5$	$\rho=.8$
1/	simul.	0.535	1.57	57.83
	approx.	0.534	1.55	57.49
	error	-0.1	-1.5	-0.6
e	simul.	0.294	0.54	1.18
	approx.	0.297	0.56	1.42
	error	0.8	4.2	20.3
1/	simul.	0.373	0.81	2.56
	approx.	0.375	0.84	3.09
	error	0.4	4.9	-17.2

$\lambda_1=0.6$, $\lambda_2=\lambda_3=0.2$, $\beta_1=\beta_2=\beta_3$

Table 4

Q_i	EW_i	$\rho=.3$	$\rho=.5$	$\rho=.8$
1/	simul.	0.532	1.56	60.38
	approx.	0.531	1.53	56.68
	error	-0.2	-2.4	-6.1
g	simul.	0.328	0.63	1.47
	approx.	0.332	0.67	1.93
	error	1.3	6.9	31.3
1/	simul.	0.370	0.79	2.48
	approx.	0.372	0.83	3.04
	error	0.6	4.8	22.6

$\lambda_1=0.6$, $\lambda_2=\lambda_3=0.2$, $\beta_1=\beta_2=\beta_3$

Table 5

Q_i	EW_i	$\rho=.3$	$\rho=.5$	$\rho=.8$
e	simul.	0.291	0.61	2.51
	approx.	0.291	0.62	2.53
	error	0.1	0.5	1.1
g	simul.	0.379	0.97	5.60
	approx.	0.376	0.97	5.65
	error	-0.9	0.5	1.0
g	simul.	0.378	0.98	5.81
	approx.	0.376	0.97	5.65
	error	-0.6	-1.1	-2.7

$\lambda_1=0.6$, $\lambda_2=\lambda_3=0.2$, $\beta_1=\beta_2=\beta_3$

Table 6

Q_i	EW_i	$\rho=.3$	$\rho=.5$	$\rho=.8$
g	simul.	0.392	0.95	4.94
	approx.	0.393	0.96	4.98
	error	0.3	1.1	0.7
e	simul.	0.314	0.67	2.88
	approx.	0.313	0.66	2.83
	error	-0.5	-0.6	-2.0
e	simul.	0.313	0.66	2.80
	approx.	0.313	0.66	2.83
	error	0.1	0.8	0.9

$\lambda_1=0.6$, $\lambda_2=\lambda_3=0.2$, $\beta_1=\beta_2=\beta_3$

Table 7

Q_i	EW_i	$\rho=.3$	$\rho=.5$	$\rho=.8$
e	simul.	0.321	0.68	1.99
	approx.	0.324	0.71	2.42
	error	1.0	4.6	21.5
1/	simul.	0.503	1.58	18.51
	approx.	0.501	1.53	16.77
	error	-0.4	-2.9	-9.4
1/	simul.	0.510	1.60	18.38
	approx.	0.501	1.53	16.77
	error	-1.8	-4.4	-8.8

$\lambda_1=\lambda_2=\lambda_3=1/3,\ \beta_2=\beta_3=(1/3)\beta_1$

Table 8

Q_i	EW_i	$\rho=.3$	$\rho=.5$	$\rho=.8$
g	simul.	0.420	0.97	3.06
	approx.	0.426	1.06	4.82
	error	1.4	8.9	57.6
1/	simul.	0.467	1.38	16.95
	approx.	0.457	1.22	11.73
	error	-1.8	-11.2	-30.8
1/	simul.	0.469	1.39	17.09
	approx.	0.457	1.22	11.73
	error	-2.6	-11.9	-31.4

$\lambda_1=\lambda_2=\lambda_3=1/3,\ \beta_2=\beta_3=(1/3)\beta_1$

Table 9

Q_i	EW_i	$\rho=.3$	$\rho=.5$	$\rho=.8$
1/	simul.	0.515	1.44	12.99
	approx.	0.515	1.47	13.74
	error	0.1	1.9	5.8
e	simul.	0.323	0.63	1.49
	approx.	0.330	0.66	1.70
	error	2.2	5.5	14.0
1/	simul.	0.452	1.12	10.33
	approx.	0.445	1.10	7.28
	error	-1.6	-8.1	-29.6

$\lambda_1=\lambda_2=\lambda_3=1/3,\ \beta_2=\beta_3=(1/3)\beta_1$

Table 10

Q_i	EW_i	$\rho=.3$	$\rho=.5$	$\rho=.8$
1/	simul.	0.516	1.43	12.53
	approx.	0.511	1.45	13.53
	error	-0.9	1.2	8.0
g	simul.	0.357	0.73	1.85
	approx.	0.369	0.80	2.31
	error	3.4	9.6	24.7
1/	simul.	0.453	1.20	10.04
	approx.	0.441	1.08	7.16
	error	-2.6	-9.4	-28.6

$\lambda_1=\lambda_2=\lambda_3=1/3,\ \beta_2=\beta_3=(1/3)\beta_1$

Table 11

Q_i	EW_i	$\rho=.3$	$\rho=.5$	$\rho=.8$
e	simul.	0.326	0.73	3.14
	approx.	0.328	0.75	3.22
	error	0.6	2.1	2.6
g	simul.	0.419	1.16	7.01
	approx.	0.424	1.17	7.18
	error	1.1	1.6	2.4
g	simul.	0.433	1.24	7.60
	approx.	0.424	1.17	7.18
	error	-2.3	-5.2	-5.5

$\lambda_1=\lambda_2=\lambda_3=1/3,\ \beta_2=\beta_3=(1/3)\beta_1$

Table 12

Q_i	EW_i	$\rho=.3$	$\rho=.5$	$\rho=.8$
g	simul.	0.439	1.16	6.23
	approx.	0.438	1.14	6.22
	error	-0.3	-1.3	-0.1
e	simul.	0.342	0.77	3.46
	approx.	0.349	0.79	3.53
	error	1.9	2.4	2.0
e	simul.	0.350	0.80	3.53
	approx.	0.349	0.79	3.53
	error	-0.4	-1.4	-0.1

$\lambda_1=\lambda_2=\lambda_3=1/3,\ \beta_2=\beta_3=(1/3)\beta_1$

Table 13

	case 1			case 2			case 3		
Q_i	simul.	approx./error		simul.	approx./error		simul.	approx./error	
e	0.84	0.85	(1.4)	1.00	1.01	(0.8)	0.85	0.89	(4.4)
e	1.15	1.16	(0.9)	1.00	1.01	(0.6)	0.85	0.89	(4.3)
e	1.14	1.16	(1.4)	1.00	1.01	(0.4)	0.85	0.89	(4.4)
g	1.19	1.21	(1.5)	1.31	1.32	(0.8)	0.89	0.93	(4.2)
g	1.20	1.21	(1.0)	1.05	1.05	(-0.6)	0.89	0.93	(3.9)
g	1.21	1.21	(-0.1)	1.06	1.05	(-0.9)	0.89	0.93	(3.5)
g	1.20	1.21	(0.9)	1.05	1.05	(0.0)	0.89	0.93	(3.6)
ll	1.32	1.29	(-1.7)	1.15	1.12	(-2.5)	4.40	4.31	(-2.0)
ll	1.33	1.29	(-2.6)	1.15	1.12	(-2.5)	0.93	0.99	(5.9)
ll	1.32	1.29	(-2.4)	1.15	1.12	(-2.5)	0.93	0.99	(5.6)
ll	1.33	1.29	(-2.9)	1.15	1.12	(-2.6)	0.94	0.99	(5.3)
ll	1.33	1.29	(-2.6)	1.15	1.12	(-2.3)	0.94	0.99	(5.0)

Table 14

	case 1			case 2			case 3		
Q_i	simul.	approx./error		simul.	approx./error		simul.	approx./error	
e	1.91	2.11	(10.5)	2.09	2.26	(7.9)	1.78	1.99	(11.9)
e	2.65	2.87	(8.2)	2.13	2.26	(5.9)	1.81	1.99	(9.9)
e	2.71	2.87	(5.9)	2.17	2.26	(4.0)	1.83	1.99	(8.6)
g	2.83	2.99	(5.6)	2.81	2.95	(4.9)	1.91	2.07	(8.4)
g	2.86	2.99	(4.3)	1.99	2.35	(17.9)	1.94	2.07	(6.9)
g	2.96	2.99	(1.1)	2.02	2.35	(16.5)	1.97	2.07	(5.4)
g	3.03	2.99	(-1.5)	2.05	2.35	(14.7)	2.01	2.07	(3.2)
ll	4.03	3.38	(-15.9)	3.15	2.66	(-15.6)	3.50	3.52	(0.6)
ll	4.04	3.38	(-16.2)	3.20	2.66	(-16.9)	2.59	2.35	(-9.2)
ll	4.10	3.38	(-17.4)	3.23	2.66	(-17.7)	2.60	2.35	(-9.6)
ll	4.08	3.38	(-17.0)	3.21	2.66	(-17.0)	2.60	2.35	(-9.8)
ll	4.15	3.38	(-18.4)	3.25	2.66	(-18.2)	2.60	2.35	(-9.7)