

An Approximation for Mean Waiting Times in Cyclic Server Systems with Nonexhaustive Service

Mandyam M. Srinivasan

Department of Industrial and Operations Engineering, The University of Michigan, 1205 Beal Avenue, Ann Arbor, MI 48109-2117, U.S.A.

Received 13 July 1987

Revised 22 December 1987 and 29 February 1988

The cyclic server system has been the subject of considerable research over the last few years. Interest in analyzing such systems has gained momentum due to their application in the performance analysis of token ring networks. In this paper we consider cyclic server systems with nonexhaustive service discipline. The performance measures of interest here are the mean waiting times at the nodes in the system. Exact analysis of such systems for these performance measures is very difficult in general, and a number of approximation schemes have been proposed in the past to evaluate these quantities. This paper presents a new approximation technique that gives accurate estimates of these mean waiting times, based on extensive validation with simulations.

Keywords: Cyclic Server Model, Token Ring Network, Nonexhaustive Service, M/G/1 Server Vacation Model, Queueing Analysis, Computer Communication Network, Local Area Network.



Mandyam M. Srinivasan received the Master of Technology degree from the Indian Institute of Technology, Madras, India, and a Ph.D. in Industrial Engineering and Management Sciences from Northwestern University, Evanston, IL, in 1973 and 1985, respectively. He is an Assistant Professor in the Industrial and Operations Engineering Department at The University of Michigan, Ann Arbor, MI. His current research interests are in performance evaluation of computer communication networks and flexible manufacturing systems.

Dr. Srinivasan is a member of ACM and TIMS. His work has been published in various journals including *IEEE Transactions on Computers*, *IEEE Network*, *Computers and Operations Research* and the *Large Scale Systems Journal*.

1. Introduction

A cyclic server system is a system in which a single server attends, in a cyclic manner, to a number of centers (nodes) at which requests arrive, and queue up for service. The number of requests serviced at a node, during a visit there by the server, depends on the service discipline that the server adopts. The service disciplines that are typically modeled are the exhaustive, the gated, and the nonexhaustive service disciplines [22]. When the server departs from a node, he can take a finite amount of time to switch to the next node, and this is termed the switch-over time.

Interest in the performance analysis of such systems has gained considerable momentum recently, especially owing to their direct application in the modeling and analysis of token ring networks. In modeling such networks, the token is modeled as the single server, and the packets that are generated by the nodes, for transmission to other nodes, are the requests for service from the system. When the bandwidth is constant, the size of the packet determines the service time required to transmit it. The overhead involved in buffering data and switching control from one node to the next and the propagation delay, together, constitute the switch-over time between nodes. Some typical performance measures of interest here are the mean waiting time for a request and the distribution of the cycle time (the time required to make one complete scan of the system).

In this paper we consider systems with the nonexhaustive service discipline where at most one request is attended to by the server during a visit to a node. This discipline has been widely adopted in the implementation of token ring networks, due to its perceived fairness. Requests are assumed here to arrive at the nodes according to independent Poisson processes and it is assumed that there is unlimited waiting room at each node to hold these requests. It is also assumed that in

North-Holland

Performance Evaluation 9 (1988) 17-33

each cycle the server visits each node exactly once. If the server finds no requests at a node when he visits it, it is assumed that he immediately begins to switch over to the next node.

The analysis of such systems presents considerable difficulties; in general, the exact analysis for even the mean waiting times in systems with more than two nodes is unknown at present, and a number of approximate analytical schemes have been proposed in the past for obtaining these mean waiting times. In this paper, a new approximation technique, termed Myopic Analysis of Cyclic Non-Exhaustive Service Systems (MAC-NESS) is presented. This approximation technique appears to be very effective in obtaining estimates of the mean waiting times, in comparison with techniques reported previously.

2. Previous work on cyclic server systems

The seminal work on cyclic server systems may be attributed to the analyses by Cooper and Murray [9] and by Cooper [8], who consider systems with exhaustive and gated service disciplines without switch-over times. Eisenberg [10] obtains the Laplace–Stieltjes transforms (LSTs) for the waiting times and the intervisit time distributions at each node in a system with exhaustive service and nonzero switchover times. The analysis of both exhaustive and gated service systems having nonzero switchover times is also presented by Hashida [14] and Ferguson and Aminetzah [12]. Bux [5] analyzes the system in which all nodes have identical arrival patterns, service time distributions, and switch-over times (the symmetric system). The above analyses are all exact. Simple, approximate analytical models for nonsymmetric systems are proposed by Bux and Truong [6] and Carsten et al. [7]. An excellent overview on the state-of-the-art in the analysis of polling systems is presented by Takagi [21].

2.1. Previous work on cyclic server systems with nonexhaustive service

The analysis of systems using the nonexhaustive service discipline presents considerable difficulties. A complete analysis of the system with two nodes, without switch-over times is presented by Eisenberg [11]; the system with two nodes with

switch-over times, having identical characteristics, is analyzed by Boxma [2]. These require a complex analysis of Riemann–Hilbert boundary value problems and, even for the mean waiting times, no simple expression results. For the symmetric case, a simple closed form expression for the mean waiting times has been obtained (see [22,20,13]). In addition, a conservation law exists for these systems [22], which presents one equation for a weighted sum of the mean waiting times in terms of known data parameters. In general, an exact analysis of such systems appears extremely difficult. A number of approximation techniques have been proposed for obtaining these mean waiting times. These approximations are usually validated through extensive simulations.

A notable contribution towards approximate analysis of cyclic server systems with nonexhaustive service is the work of Kuehn [17] who considered systems with batch Poisson input. The analysis obtains the generating function for the stationary state probabilities, the LSTs of the delay distributions, and the mean waiting time at each node in the system. To obtain these estimates, two conditional cycle times are considered: a cycle time which includes a service at node i , and a cycle time which has no service at node i . The variance of each of these cycle times is then approximated assuming that, in either of these cycles, the sojourn time at each node is independent of the sojourn times at the other nodes (the independence assumption). An imbedded Markov chain analysis now obtains the desired estimates. In addition, a stability criterion is derived for general GI/G/1 systems with cyclic service. Following the work of Kuehn, a number of approximate analytical results have been reported for such systems [1,3,15,19].

The paper by Berry and Chandy [1] requires identical distributions for the service times at all nodes. Arrivals at individual nodes are assumed to be Poisson. The switch-over time is assumed to be a small constant, and is the same for each pair of nodes. With these assumptions, the approximation technique then views the entire system as a single M/G/1 queue with an arrival rate set equal to the sum of the arrival rates over all nodes. It calculates the overall mean queue length in this M/G/1 system, and then allocates this quantity among the nodes using an iterative heuristic that is developed in the paper. A simple application of Little's rule

[18] then provides the mean waiting times at these nodes.

Kimura and Takahashi [15] present a diffusion approximation to analyze systems in which each node can be subject to batch arrivals having arbitrary distributions. The analysis considers conditional cycle times, and uses the independence assumption on these conditional cycle times in a manner similar to Kuehn's analysis. For the special case where the arrivals are Poisson, the mean waiting times obtained by this analysis are very close to the values obtained by the method of Kuehn.

Boxma and Meister [3] consider a nonsymmetric system with Poisson arrivals at each node. The approximation makes use of the conservation law presented by Watson [22], and obtains a closed form expression for the mean waiting times in systems with switch-over times. This approximation appears to provide the most accurate estimates for the mean waiting times among the techniques reported in the past. A similar result for systems without switch-over times is presented by Boxma and Meister in [4].

3. The analysis

Let N denote the number of nodes in the cyclic server system. Requests for service arrive, at node n in this system, according to an independent Poisson process with rate λ_n . The service time demands made by requests at node n are assumed to be independent, identically distributed (i.i.d.) random variables with mean b_n and second moment $b_n^{(2)}$. The switch-over times between node n and node $(n \bmod N) + 1$ are i.i.d. random variables with mean s_n and second moment $s_n^{(2)}$. The utilization of the server at node n , ρ_n , is defined as $\rho_n = \lambda_n b_n$.

In the ensuing discussion, unless specified otherwise, the index for any summation is assumed to be over the range 1 through N . Let

$$s = \sum_n s_n \quad \text{and} \quad \rho = \sum_n \rho_n.$$

The expected cycle time, c , is obtained from [17]

$$c = \sum_n s_n + \sum_n \rho_n c, \quad (1)$$

from which

$$c = s / (1 - \rho). \quad (2)$$

It can be shown [17] that the following conditions are necessary and sufficient for the stability of this system:

$$\rho < 1 \quad (3a)$$

and

$$\max_n \lambda_n c < 1. \quad (3b)$$

It is assumed that the system being analyzed satisfies the above stability conditions.

3.1. The expression for the mean waiting times

Consider the mean waiting time, w_n , experienced by a tagged request (customer) arriving at node n , $1 \leq n \leq N$, in the system. Let $p_n(i)$ denote the probability that the arriving customer sees i customers already present at this node. Owing to the fact that Poisson arrivals see time averages [24], this tagged customer sees the equilibrium distribution of customers present at the node. Let t_n denote the mean system time (mean waiting time + mean service time) for this tagged customer and let $t_n(i)$ represent the mean time in the system for this customer, conditioned on the fact that it sees i customers at node n on arrival. Thus,

$$w_n + b_n = t_n = \sum_{i \geq 0} p_n(i) t_n(i). \quad (4)$$

If we can estimate the $t_n(i)$ values, we can thus determine w_n from equation (4). To estimate $t_n(i)$ we consider two cases: $i = 0$ and $i > 0$.

Case $i = 0$. If the tagged customer finds $i = 0$ customers at node n upon arrival then it always sees the server on vacation from that node and interrupts this vacation. Let \tilde{v}_n denote the expected residual life of this interrupted vacation. Thus,

$$t_n(0) = \tilde{v}_n + b_n. \quad (5)$$

Case $i > 0$. Suppose, on the other hand, that the tagged customer finds $i > 0$ customers at node n upon arrival. In this case we choose to identify the customer at the head of this queue as the Head-Of-Line (HOL) customer. The expected time in the system, for the tagged customer, is then the sum of two quantities: (i) the expected time, r_n , from the time of its arrival till the HOL customer

departs the system, and (ii) the expected time for the server to complete service on the $i - 1$ remaining customers, followed by a service on the tagged customer, namely to complete i successive cycles, each of which includes a service at node n . To determine r_n we consider two possible situations:

(a) The customer found the server on vacation: the fraction of time this occurs is, on the average, x_n , where x_n denotes the fraction of the time that the server is away on vacation from node n , given at least one customer is present at the node. In this case, we assume that the customer interrupts a special vacation which has an expected residual life \tilde{v}_n . (This is clearly an approximation since, in general, the length of a vacation is dependent on the number of customers present at the node.) This implies that the expected time till the departure of the HOL customer is, approximately, $\tilde{v}_n + b_n$.

(b) The arriving customer found the server servicing the HOL customer. The fraction of time this occurs is, on the average, $1 - x_n$. The expected time till the departure of the HOL customer here is just the expected residual life of this special service time which is equal to $b_n^{(2)}/2b_n$. The term r_n is thus given as

$$r_n = x_n(\tilde{v}_n + b_n) + (1 - x_n)\left(\frac{b_n^{(2)}}{2b_n}\right). \quad (6)$$

The term x_n in equation (6) is evaluated as follows: let f_n denote the fraction of time that the server is away from the node, on a vacation. Since the utilization of the server at node n is given by ρ_n , we can write

$$f_n = 1 - \rho_n.$$

In this system, the server immediately begins a new vacation if he finds no customer present at a node when he visits it. Note that the server is away on vacation at least $p_n(0)$ of the time, on the average. It is clear that f_n can then also be written as

$$f_n = p_n(0) + (1 - p_n(0))x_n.$$

Equating the two expressions for f_n , we have

$$x_n = \frac{1 - \rho_n - p_n(0)}{1 - p_n(0)}. \quad (7)$$

We still need to determine the expected time for the tagged customer in the system, from the moment that the HOL customer departs, till the time of service completion on the tagged customer.

To this end, let v_n denote the expected length of a vacation which begins after a normal service at node n (i.e., after a service of expected length b_n), ending when the server returns to node n . Also, let

$$C_n = v_n + b_n \quad (8)$$

denote the expected length of a cycle which begins with a normal service at node n , ending with the next arrival instant of the server at node n . This cycle includes possible services at the other nodes, plus the sum of all the switch-over times. To determine C_n , let α_{mn} denote the probability that this cycle contains a service at another node m . Then, since at most one customer is served at node m , we can write (refer also to [17] for a similar derivation)

$$\alpha_{mn} \approx \lambda_m C_n, \quad m \neq n.$$

Note that it is possible that when the arrival rate to some node, m , is high, then α_{mn} may exceed 1, in which case it can no longer be interpreted as a probability. In such cases, it will be necessary to restrict this quantity to be at most 1. Hence, the expected length of this cycle is given by

$$C_n \approx s + b_n + \sum_{m \neq n} \min\{\lambda_m C_n, 1\} b_m. \quad (9)$$

If $\alpha_{mn} \leq 1$ for all m , then the above expression simplifies to

$$C_n = \frac{s + b_n}{1 - \rho + \rho_n}, \quad (9a)$$

otherwise, computing the C_n values will involve some iteration.

Each of these vacations following the departure of the HOL customer is preceded by a service at node n . We could now assume that the expected length of each of these vacation equals v_n and proceed to develop the expression for w_n . However, this assumption would be incorrect, at least in the case of the vacation immediately following the departure of the HOL customer. To illustrate this, consider the situation where the tagged customer found the server at node n , attending to the HOL customer. The arrival then interrupts a special service which has expected duration $b_n^{(2)}/b_n$. Adopting a similar reasoning as was used earlier to obtain C_n , the cycle which includes this special service has expected length $C_n(b)$, where

$$C_n(b) \approx s + b_n + \sum_{m \neq n} \min\{\lambda_m C_n(b), 1\} b_m. \quad (10)$$

As before, it may be necessary in some cases to restrict the term $\lambda_m C_n(b)$ to be at most 1. If this term is less than 1 for all m , the above expression simplifies to

$$C_n(b) \approx \frac{s + b_n^{(2)}/b_n}{1 - \rho + \rho_n}, \quad (10a)$$

otherwise obtaining $C_n(b)$ will generally involve some iteration. Let

$$v_n(b) = C_n(b) - b_n^{(2)}/b_n$$

denote the expected length of the special vacation following this interrupted service. Clearly, this is not equal to v_n . Similarly, if the tagged customer arrives at node n when the server is on vacation, then he interrupts a special vacation which, in turn, influences the vacation following the subsequent departure of the HOL customer at node n .

Let the expected length of the vacation, following the departure of the HOL customer, be denoted by \hat{v}_n . Using arguments similar to that presented above, and assuming that the remaining $i - 1$ vacations all have expected length v_n , we can write

$$t_n(i) = x_n(\tilde{v}_n + b_n) + (1 - x_n) \left(\frac{b_n^{(2)}}{2b_n} \right) + (\hat{v}_n + b_n) + (i - 1)(v_n + b_n), \quad i > 0. \quad (11)$$

From equations (4) through (7) and equation (11), after some algebra we get

$$w_n(1 - \lambda_n v_n - \rho_n) = \tilde{v}_n(1 - \rho_n) + \rho_n \frac{b_n^{(2)}}{2b_n} + \rho_n v_n + (\hat{v}_n - v_n)(1 - p_n(0)). \quad (12)$$

In the above expression, to determine w_n , we still need to evaluate the terms \hat{v}_n , \tilde{v}_n , and $p_n(0)$. The expressions for these terms are developed in Sections 3.2 and 3.3.

3.2. Determining the probability $p_n(0)$

For this system, determining $p_n(0)$ exactly can be very complex. Instead, we choose to estimate $p_n(0)$ by considering this cyclic server system from a different viewpoint. Consider a single node, single server system with Poisson arrivals in which, for each customer, the server requires a set-up

time that is independent of the service time. Further, suppose that this set-up time has a different distribution for the customer that arrives at an empty system than for a customer that arrives at a nonempty system. Such a system was also studied by Welch [23] who obtained the distribution of the number of customers present at the node, given the first two moments of the distributions for the two set-up times and for the service time. We shall refer to this system, for convenience, as system W .

Suppose the arrival rate to system W is λ . Let β denote the first moment of the distribution for the service time, and let ψ (respectively $\tilde{\psi}$) denote the first moment of the set-up time required for arrivals to a nonempty system (respectively, arrivals to an empty system). The probability π_0 of finding zero customers at a random point in time in this system requires only these first moments, and is presented below as Lemma 3.1. A proof of Lemma 3.1 may be found in [23].

3.1. Lemma

$$\pi_0 = \frac{1 - \lambda(\psi + \beta)}{1 - \lambda(\psi - \tilde{\psi})}.$$

It can be observed that the behavior of the cyclic server system with nonexhaustive service, as we have modeled it, closely resembles system W . When a customer arrives at an empty system at node n it interrupts a special 'set-up' time, which has a mean residual life equal to \tilde{v}_n at the time of interruption, before the server can begin actual service on this customer. Assuming that the mean 'set-up' time in this case is twice this mean residual life (this is, of course, an approximation), the mean set-up time here is $2\tilde{v}_n$. On the other hand, if the customer arrives at a nonempty system, then the mean 'set-up' time in progress, between services to customers at this node, is given by v_n (with the exception, which we choose to overlook, of the service following the HOL customer, which involves a mean 'set-up' time of \hat{v}_n as per our assumptions). Hence the probability $p_n(0)$ of finding zero customers in this system, is then equated to π_0 , with appropriate substitution of parameters, to obtain the following result.

3.2. Proposition

$$p_n(0) = \frac{1 - \lambda_n(v_n + b_n)}{1 - \lambda_n(v_n - 2\tilde{v}_n)}. \quad (13)$$

3.3. Evaluating the terms \tilde{v}_n and \hat{v}_n

The terms \tilde{v}_n and \hat{v}_n are evaluated by conditioning on the position of the server, as observed by the tagged customer on arrival. Let γ_m denote the event that the server is at node m at the time of arrival of the tagged customer. Similarly, let σ_m denote the event that the server is switching from node m to node $(m \bmod N) + 1$ at the time of arrival of the tagged customer. Let $q(\gamma_m)$ and $q(\sigma_m)$, respectively, denote the probabilities of these events. From equation (1) it can be seen that

$$\sum_m s_m/c + \sum_m \rho_m = 1.$$

Thus, the term ρ_m can be interpreted as the probability that the server is present at node m at a random point in time and, similarly, the term s_m/c can be interpreted as the probability that the server is switching between nodes m and $(m \bmod N) + 1$ at a random point in time. Noting the fact that a Poisson arrival takes a random look at the system, we must have $q(\gamma_m) = \rho_m$ and $q(\sigma_m) = s_m/c$.

We also need to define the expected length of two cycles: (i) a cycle which includes a special service at node n of expected length $b_n^{(2)}/b_n$, denoted $C_n(b)$, and (ii) a cycle which includes a special switch-over between nodes n and $n + 1$ of expected length $s_n^{(2)}/s_n$, denoted $C_n(s)$. The expression for $C_n(b)$ was given by equation (10). The expression for $C_n(s)$ is obtained using a similar reasoning:

$$C_n(s) = \frac{s + s_n^{(2)}/s_n - s_n}{1 - \rho}. \quad (14)$$

The expression for \tilde{v}_n is then presented below as Proposition 3.3. The derivation for this expression is given in Appendix A at the end of this paper.

3.3. Proposition

$$\begin{aligned} \tilde{v}_n = & \sum_{m, m \neq n} (q(\gamma_m)/(1 - \rho_n)) \tilde{v}(n | \gamma_m) \\ & + \sum_m (q(\sigma_m)/(1 - \rho_n)) \tilde{v}(n | \sigma_m), \end{aligned} \quad (15)$$

where

$$\begin{aligned} \tilde{v}(n | \gamma_m) = & \frac{b_m^{(2)}}{2b_m} + \sum_{k=m}^n s_k \\ & + \sum_{k=m+1}^n \min\{\lambda_k C_m(b), 1\} b_k \end{aligned} \quad (15a)$$

and

$$\begin{aligned} \tilde{v}(n | \sigma_m) = & \frac{s_m^{(2)}}{2s_m} + \sum_{k=m+1}^n s_k \\ & + \sum_{k=m+1}^n \min\{\lambda_k C_m(s), 1\} b_k. \end{aligned} \quad (15b)$$

In equations (15a) and (15b) it is assumed that $1 \leq m \leq n \leq N$. This avoids the use of the mod function. Note that this does not lead to any loss of generality. Proposition 3.4 now develops the expression for the term \hat{v}_n . The derivation of this expression is also presented in Appendix A.

3.4. Proposition

$$\hat{v}_n = (1 - x_n)v_n(b) + x_n v_n(1), \quad (16)$$

where

$$\begin{aligned} v_n(1) = & \sum_{m, m \neq n} (q(\gamma_m)/(1 - \rho_n)) v_n(1 | \gamma_m) \\ & + \sum_m (q(\sigma_m)/(1 - \rho_n)) v_n(1 | \sigma_m), \end{aligned} \quad (17)$$

with

$$\begin{aligned} v_n(1 | \gamma_m) = & s + \sum_{k, k \neq n} \min\{\lambda_k \max[C_m(b), C_n], 1\} b_k \end{aligned} \quad (17a)$$

and

$$\begin{aligned} v_n(1 | \sigma_m) = & s + \sum_{k, k \neq n} \min\{\lambda_k \max[C_m(s), C_n], 1\} b_k. \end{aligned} \quad (17b)$$

It is easy to show that the expression for the mean waiting times, given by equation (12), is exact for some limited cases. This is stated as Proposition 3.5. The proof of this proposition is straightforward, and is therefore omitted.

3.5. Proposition. *The expression for the mean waiting times given by equation (12) is exact for the single node vacation system, and for the symmetric system having deterministic service times, and deterministic switch-over times.*

4. The accuracy of the approximation

The mean waiting time estimates obtained by MACNESS were validated for their accuracy

through extensive simulation on a substantial number of test cases. For relatively low traffic ($\rho \leq 0.5$), the estimates obtained by MACNESS were very close to the simulation estimates. Under conditions of heavy traffic, when the system was quite asymmetric, some differences were observed between the two estimates. In this section, using the conservation law of Watson [22], a means of improving the accuracy of the estimates obtained by MACNESS is presented.

4.1. The conservation law

The conservation law [22], which provides one equation for the mean waiting times in terms of known data parameters, is presented below:

$$\begin{aligned} & \sum_n \rho_n (1 - \lambda_n c) w_n \\ &= \frac{\rho}{2(1 - \rho)} \sum_n \lambda_n b_n^{(2)} + \frac{\rho}{2s} \sum_n (s_n^{(2)} - s_n^2) \\ & \quad + \frac{1}{2} c \sum_n \rho_n (1 + \rho_n). \end{aligned} \quad (18)$$

This law can indicate how effective the approximation is. Suppose we substitute the mean waiting times obtained from equation (12) in place of the w_n values in equation (18) and evaluate the resulting expression on the left-hand side. Let the factor ϕ denote the ratio of the expression on the right-hand side of this equation to the quantity evaluated on the left-hand side. Obviously, the closer this ratio is to 1, the more confidence one would have in the above approximation.

A substantial number of experiments were conducted to determine the accuracy of the approximation. At low traffic intensities ($\rho \leq 0.5$), the factor ϕ was between 0.96 and 1.0 in all of these test cases. In addition, the estimates of w_n were very close to the simulation estimates here. At heavy traffic intensities, this factor ranged, in general, between 0.90 to 1.10, with a few exceptions outside this range occurring as the system was nearing the limits of stability. This is to be expected since we have made several assumptions in arriving at the expression for the mean waiting times. It is conjectured that the approximation used to obtain the expression for \tilde{v}_n is a major contributor in causing the factor to be different from 1 for the heavy traffic case. We now make use of the conservation law to obtain an improved estimate for \tilde{v}_n in the following section.

Remark. It is to be noted that Boxma and Meister [3] use the conservation law directly, to obtain their estimates of mean waiting times. This technique (henceforth referred to as the B&M technique) proceeds as follows. At the time of arrival of the tagged customer at node n , the server is at some place in the system. The arrival sees the equilibrium distribution of customers at the various nodes and, hence, on the average, sees $q_n = \lambda_n w_n$ customers waiting at node n . So the mean waiting time for this customer consists of two quantities: (i) the expected residual cycle time, rc_n , for the server to reach node n , from the place he is currently at, and (ii) the time for the server to complete $\lambda_n w_n$ cycles, each of length C_n , to serve q_n customers. The mean waiting time is then obtained as $w_n = rc_n / (1 - \lambda_n C_n)$. The B&M technique now assumes that the residual cycle times are the same for each node. It then uses the conservation law to obtain a closed form expression for the w_n 's. This technique is simple and the approximation assumptions used are very easily understood. It is remarked, though, that the resulting expression does not provide an intuitive understanding for the behavior of the system.

4.2. Improved estimates of \tilde{v}_n using the conservation law

Equation (12) is first rewritten as

$$w_n (1 - \lambda_n v_n - \rho_n) = \kappa_n + v_n, \quad (19)$$

where

$$\kappa_n = \tilde{v}_n (1 - \rho_n) + \rho_n \frac{b_n^{(2)}}{2b_n} + \rho_n v_n \quad (19a)$$

and

$$v_n = (\hat{v}_n - v_n)(1 - p_n(0)). \quad (19b)$$

In order to improve on the estimates of \tilde{v}_n , it is assumed that the term κ_n is the same for all nodes. Note the similarity between this assumption and the assumption made by the B&M technique. In fact, if we had not accounted for the 'special' vacation following the departure of the HOL customer, then this approximation would just reduce to the B&M technique and κ_n would represent the 'residual cycle time' here. (We have, of course, presented an intuitive approach for obtaining this 'residual cycle time'.) Applying the conservation law to the above expression for w_n ,

Table 1

$N = 3$; $\lambda_1 = 0.6$, $\lambda_2 = \lambda_3 = 0.2$. All service time distributions are exponential with $b_1 = b_2 = b_3$. All switch-over times are equal to zero

Node	Method	Utilization (ρ)		
		0.3	0.5	0.8
1	Simulation	0.135	0.553	4.144
	MACNESS	0.137 (1.5)	0.557 (0.7)	4.251 (2.6)
	Boxma & Meister	0.136 (0.7)	0.556 (0.5)	3.942 (-4.9)
2 to 3	Simulation	0.115	0.393	1.477
	MACNESS	0.116 (0.9)	0.414 (5.3)	1.623 (9.9)
	Boxma & Meister	0.118 (2.6)	0.417 (6.1)	2.087 (41.3)

and setting $\kappa_n = \kappa$, an expression for κ can be obtained in a straightforward manner as

$$\kappa = \left[X - \sum_n v_n \theta_n \right] / \left[\sum_n \theta_n \right], \quad (20)$$

where X represents the expression on the right-hand side of the conservation law given in equation (18), and

$$\theta_n = \frac{\rho_n (1 - \lambda_n c)}{1 - \lambda_n v_n - \rho_n}.$$

Thus, a new estimate of \tilde{v}_n is obtained from equations (20) and (19a). This estimate is now used in equation (12) for obtaining the w_n 's. The use of the conservation law in this manner does imply that this is an iterative algorithm for obtaining these values. However, in practice these values converge within a few iterations and, in using this approach, we choose not to iterate more than once.

Finally, note that the use of equation (18) in this manner guarantees that the resulting improved estimates of w_n do satisfy the conservation law. Thus, it can be noted that these values for w_n

are the exact mean waiting times, even for symmetric systems where the service times and the switch-over times can be random variables.

4.3. The special case of systems with zero switch-over times

The MACNESS approach has a direct extension to systems with zero switch-over times. We merely set the mean switch-over time between all nodes to be some arbitrarily small but finite value. Then, all the expressions presented earlier hold. (An alternate view of this approach would be to consider that each of the $q(\sigma_n)$ terms are uniformly replaced by a factor equal to $(1 - \rho)/N$.) The resulting mean waiting time estimates appear to be as accurate as in the case with nonzero switch-over times.

Boxma and Meister [4] also present an approximation technique for such systems. This is also based on the conservation law and is very similar to their technique for systems with nonzero switch-over times.

Table 2

$N = 3$; $\lambda_1 = 0.6$, $\lambda_2 = \lambda_3 = 0.2$. All service time distributions are exponential with $b_1 = b_2 = b_3$. All switch-over times are exponential and equal to 0.05

Node	Method	Utilization (ρ)		
		0.3	0.5	0.8
1	Simulation	0.333	0.976	9.090
	MACNESS	0.334 (0.3)	0.970 (-0.6)	9.162 (0.8)
	Boxma & Meister	0.334 (0.3)	0.959 (-1.7)	8.360 (-8.0)
2 to 3	Simulation	0.261	0.599	1.920
	MACNESS	0.261 (0.0)	0.614 (2.5)	2.083 (8.5)
	Boxma & Meister	0.262 (0.4)	0.628 (4.8)	1.480 (-22.9)

5. Experimental results

The estimates of the mean waiting times obtained by MACNESS for some test cases are presented in Tables 1 through 16.¹ Of these, Tables 1 through 12 are taken from [3,4], and these cover all the examples they present therein. The simulation results presented in these tables are also taken from their papers. It must be noted that the simulation estimates could have some statistical error. However, the simulation results that are presented in these tables usually satisfy the conservation law fairly closely. The MACNESS results presented are obtained using the improved method proposed in Section 4.2. For comparison, the estimates obtained using the B&M technique are also presented in these tables. (As noted earlier, the B&M technique gave the best results among all the techniques previously re-

ported. For example, the methods of Kuehn, and Kimura and Takahashi, when applied to these test cases, often gave large errors, sometimes in excess of 50%.²) The errors, indicated in parentheses in these tables are calculated relative to the values obtained by simulation.

5.1. Discussion of results

For ease of presentation, these tables present mean waiting times that are averaged over groups of queues which have identical characteristics, and for which the mean waiting times obtained were quite close. It is to be noted that the B&M technique obtains the same values for these groups of queues. In general, however, there will be some (possibly small) difference in mean waiting times between two nodes, depending on their position, even though they may have identical characteristics with regard to their service time demands, arrival rates, and switch-over times. We have not, however, been able to draw any conclusive in-

¹ Tables 1 through 16 compare the mean waiting times obtained by MACNESS with the values obtained by simulation and the technique of Boxma and Meister. In all these tables, the mean waiting times reported for MACNESS and for simulation have been averaged over the corresponding group of queues. Errors are indicated in parentheses.

² Note, however, that the analyses of Kuehn, and Kimura and Takahashi obtain more than just the mean waiting times.

Table 3

$N = 3; \lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$. All service time distributions are exponential with $b_2 = b_3 = \frac{1}{3}b_1$. All switch-over times are equal to zero

Node	Method	Utilization (ρ)		
		0.3	0.5	0.8
1	Simulation	0.175	0.677	4.473
	MACNESS	0.182 (4.0)	0.731 (8.0)	4.905 (9.7)
	Boxma & Meister	0.180 (2.9)	0.733 (8.3)	5.203 (16.3)
2 to 3	Simulation	0.156	0.569	3.570
	MACNESS	0.151 (-3.2)	0.553 (-2.8)	3.203 (-10.3)
	Boxma & Meister	0.155 (-2.5)	0.550 (-4.8)	2.755 (-23.6)

Table 4

$N = 3; \lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$. All service time distributions are exponential with $b_2 = b_3 = \frac{1}{3}b_1$. All switch-over times are exponential and equal to 0.10

Node	Method	Utilization (ρ)		
		0.3	0.5	0.8
1	Simulation	0.570	1.384	11.260
	MACNESS	0.570 (0.0)	1.470 (6.2)	12.591 (11.8)
	Boxma & Meister	0.570 (0.0)	1.494 (7.2)	13.020 (15.6)
2 to 3	Simulation	0.502	1.196	8.600
	MACNESS	0.493 (-1.8)	1.157 (-3.3)	7.554 (-12.2)
	Boxma & Meister	0.493 (-1.8)	1.121 (-6.3)	6.890 (-19.9)

Table 5

$N = 16$; $\lambda_1 = \dots = \lambda_{16} = \frac{1}{16}$. All service time distributions are exponential with $b_1 = b_7$, $b_2 = \dots = b_6 = b_8 = \dots = b_{16} = \frac{1}{3}b_1$. All switch-over times are equal to zero

Node	Method	Utilization (ρ)		
		0.3	0.5	0.8
1	Simulation	0.175	0.679	4.490
	MACNESS	0.176 (0.5)	0.711 (4.7)	4.718 (5.1)
	Boxma & Meister	0.170 (-2.9)	0.681 (0.3)	4.965 (10.6)
2 to 6	Simulation	0.163	0.602	3.891
	MACNESS	0.160 (-1.8)	0.610 (1.3)	3.832 (-1.5)
	Boxma & Meister	0.161 (0.0)	0.622 (3.3)	3.724 (-4.3)
7	Simulation	0.175	0.675	4.468
	MACNESS	0.176 (0.6)	0.710 (5.2)	4.708 (5.4)
	Boxma & Meister	0.170 (-2.9)	0.681 (0.9)	4.965 (11.1)
8 to 16	Simulation	0.161	0.620	3.869
	MACNESS	0.160 (-0.6)	0.610 (-1.6)	3.831 (-1.0)
	Boxma & Meister	0.163 (1.2)	0.622 (0.3)	3.724 (-3.7)

Table 6

$N = 16$; $\lambda_1 = 0.6$; $\lambda_2 = \dots = \lambda_{16} = \frac{2}{75}$. All service time distributions are exponential with identical means. All switch-over times are equal to zero

Node	Method	Utilization (ρ)		
		0.3	0.5	0.8
1	Simulation	0.140	0.595	4.538
	MACNESS	0.142 (1.4)	0.604 (1.5)	4.601 (1.4)
	Boxma & Meister	0.140 (0.0)	0.584 (-1.8)	4.383 (-3.4)
2 to 16	Simulation	0.110	0.355	1.149
	MACNESS	0.108 (-1.8)	0.345 (-2.8)	1.098 (-4.4)
	Boxma & Meister	0.113 (2.8)	0.375 (5.6)	1.427 (24.2)

Table 7

$N = 16$; $\lambda_1 = \dots = \lambda_{16} = \frac{1}{16}$. All service time distributions are exponential with $b_1 = b_7$, $b_2 = \dots = b_6 = b_8 = \dots = b_{16} = \frac{1}{3}b_1$. All switch-over times are equal to 0.05

Node	Method	Utilization (ρ)		
		0.3	0.5	0.8
1	Simulation	0.823	1.697	8.780
	MACNESS	0.835 (1.5)	1.752 (3.2)	9.349 (6.5)
	Boxma & Meister	0.831 (1.0)	1.742 (2.7)	10.060 (14.6)
2 to 6	Simulation	0.793	1.591	7.980
	MACNESS	0.796 (0.4)	1.586 (-0.3)	7.900 (-1.0)
	Boxma & Meister	0.797 (0.5)	1.590 (-0.1)	7.540 (-5.5)
7	Simulation	0.833	1.720	8.900
	MACNESS	0.835 (0.2)	1.752 (1.9)	9.340 (4.9)
	Boxma & Meister	0.831 (-0.2)	1.742 (1.3)	10.060 (11.8)
8 to 16	Simulation	0.793	1.591	7.910
	MACNESS	0.796 (0.3)	1.586 (-0.3)	7.850 (-0.8)
	Boxma & Meister	0.797 (0.5)	1.590 (-0.1)	7.540 (-4.6)

Table 8

$N = 16$; $\lambda_1 = 0.6$; $\lambda_2 = \dots = \lambda_{16} = \frac{2}{75}$. All service time distributions are exponential with identical means. All switch-over times are equal to 0.01

Node	Method	Utilization (ρ)		
		0.3	0.5	0.8
1	Simulation	0.330	1.015	9.710
	MACNESS	0.325 (-1.5)	1.026 (1.1)	10.284 (5.9)
	Boxma & Meister	0.321 (-2.7)	0.996 (-1.9)	9.790 (0.9)
2 to 16	Simulation	0.222	0.495	1.350
	MACNESS	0.219 (-1.4)	0.484 (-2.2)	1.303 (-3.5)
	Boxma & Meister	0.224 (0.9)	0.521 (5.3)	1.240 (-8.1)

Table 9

$N = 16$, $\lambda_1 = \lambda_7 = 0.15$; $\lambda_2 = \dots = \lambda_6 = \lambda_8 = \dots = \lambda_{16} = 0.05$. Service time distributions at nodes 2, ..., 6 and 8, ..., 16 are exponential with identical means; service at node 1 Erlang-4 with $b_1 = 6b_2$; service at node 7 hyperexponential with coefficient of variation = 2, and $b_7 = 6b_2$. All switch-over times are equal to zero

Node	Method	Utilization (ρ)		
		0.3	0.5	0.8
1	Simulation	0.377	1.479	10.748
	MACNESS	0.349 (-7.4)	1.403 (-5.1)	10.427 (-3.0)
	Boxma & Meister	0.375 (-0.5)	1.512 (2.2)	10.662 (-0.8)
2 to 6	Simulation	0.332	1.107	4.128
	MACNESS	0.300 (-9.6)	1.030 (-7.0)	4.605 (11.6)
	Boxma & Meister	0.328 (-1.2)	1.134 (2.4)	4.719 (14.3)
7	Simulation	0.385	1.547	11.105
	MACNESS	0.422 (9.6)	1.709 (10.5)	11.027 (-0.7)
	Boxma & Meister	0.375 (-2.6)	1.512 (-2.3)	10.662 (-4.0)
8 to 16	Simulation	0.307	1.015	3.888
	MACNESS	0.299 (2.7)	1.015 (0.0)	4.523 (16.3)
	Boxma & Meister	0.328 (6.8)	1.134 (11.7)	4.719 (21.4)

Table 10

$N = 16$, $\lambda_1 = \dots = \lambda_4 = 0.16$; $\lambda_5 = \dots = \lambda_{16} = 0.03$. All service time distributions are exponential with identical means. All switch-over times are equal to zero

Node	Method	Utilization (ρ)		
		0.3	0.5	0.8
1 to 4	Simulation	0.131	0.532	3.905
	MACNESS	0.133 (1.5)	0.535 (0.6)	3.926 (0.5)
	Boxma & Meister	0.131 (0.0)	0.521 (-2.1)	3.612 (-7.5)
5 to 16	Simulation	0.123	0.439	1.896
	MACNESS	0.121 (-1.6)	0.437 (-0.5)	1.910 (0.7)
	Boxma & Meister	0.124 (0.8)	0.463 (5.5)	2.467 (30.1)

Table 11

$N = 16$, $\lambda_1 = \lambda_7 = 0.15$; $\lambda_2 = \dots = \lambda_6 = \lambda_8 = \dots = \lambda_{16} = 0.05$. Service time distributions at nodes 2, ..., 6 and 8, ..., 16 are exponential with identical means; service at node 1 Erlang-4 with $b_1 = 6b_2$; service at node 7 hyperexponential with coefficient of variation = 2, and $b_7 = 6b_2$. All switch-over times are equal to 0.05

Node	Method	Utilization (ρ)		
		0.3	0.5	0.8
1	Simulation	1.198	3.253	41.260
	MACNESS	1.200 (0.2)	3.189 (-2.0)	33.993 (-17.6)
	Boxma & Meister	1.224 (2.2)	3.271 (0.6)	33.840 (-18.0)
2 to 6	Simulation	0.946	2.011	6.270
	MACNESS	0.913 (-3.5)	1.933 (-3.9)	7.059 (12.6)
	Boxma & Meister	0.940 (-0.6)	2.027 (0.8)	4.900 (-21.9)
7	Simulation	1.247	3.335	39.210
	MACNESS	1.273 (2.1)	3.447 (3.3)	34.379 (-12.3)
	Boxma & Meister	1.224 (-1.8)	3.271 (-1.9)	33.840 (-13.7)
8 to 16	Simulation	0.922	1.902	6.170
	MACNESS	0.912 (-1.1)	1.923 (1.1)	7.027 (13.9)
	Boxma & Meister	0.940 (2.0)	2.027 (6.6)	4.900 (-20.6)

Table 12

$N = 16$, $\lambda_1 = \dots = \lambda_4 = 0.16$; $\lambda_5 = \dots = \lambda_{16} = 0.03$. All service time distributions are exponential with identical means. All switch-over times are equal to 0.05

Node	Method	Utilization (ρ)		
		0.3	0.5	0.8
1 to 4	Simulation	0.898	1.929	17.660
	MACNESS	0.901 (0.3)	1.922 (-0.4)	17.901 (1.3)
	Boxma & Meister	0.897 (-0.1)	1.884 (-2.3)	16.870 (-4.2)
5 to 16	Simulation	0.717	1.267	3.570
	MACNESS	0.714 (-0.4)	1.255 (-1.0)	3.967 (11.1)
	Boxma & Meister	0.720 (0.4)	1.307 (3.2)	3.14 (-12.0)

Table 13

$N = 8$, $\lambda_1 = \dots = \lambda_3 = 0.3$, $\lambda_4 = \dots = \lambda_8 = 0.02$; $b_1 = \dots = b_8$. Service at nodes 1 through 3 are Erlang-4. Service at nodes 4 through 8 are hyperexponential with coefficient of variation = 2. All switch-over times are equal to 0.05

Node	Method	Utilization (ρ)		
		0.3	0.5	0.8
1 to 3	Simulation	0.505	1.188	9.250
	MACNESS	0.505 (0.0)	1.160 (-2.4)	10.160 (9.8)
	Boxma & Meister	0.504 (-0.2)	1.041 (3.8)	10.056 (8.7)
4 to 8	Simulation	0.379	0.685	1.597
	MACNESS	0.376 (-0.8)	0.681 (-0.6)	1.782 (11.6)
	Boxma & Meister	0.378 (-0.2)	0.701 (2.3)	1.451 (-9.1)

Table 14

$N = 6$, $\lambda_1 = \dots = \lambda_5 = 0.0673$, $\lambda_6 = 0.2558$. All service times have bimodal distribution with coefficient of variation 1.01, and identical means. All switch-over times are equal to zero

Node	Method	Utilization (ρ)		
		0.3	0.6	0.9
1 to 5	Simulation	0.206	1.251	6.103
	MACNESS	0.204 (-1.0)	1.231 (-1.6)	5.682 (-6.9)
	Boxma & Meister	0.208 (1.0)	1.305 (4.3)	8.580 (28.9)
6	Simulation	0.236	1.875	26.747
	MACNESS	0.240 (1.7)	1.935 (3.2)	24.541 (-8.2)
	Boxma & Meister	0.235 (-0.4)	1.837 (-2.2)	20.727 (-22.51)

Table 15

$N = 5$, $\lambda_1 = \lambda_4 = 0.05$, $\lambda_2 = \lambda_5 = \frac{1}{12}$, $\lambda_3 = 0.15$. Service time distributions are exponential with identical means. All switch-over times are equal to 0.2

Node	Method	Utilization (ρ)		
		0.25	0.50	0.75
1, 4	Simulation	0.952	2.251	5.915
	MACNESS	0.931 (2.2)	2.251 (0.0)	6.590 (11.4)
	Boxma & Meister	0.936 (-1.7)	2.331 (3.6)	5.383 (-11.0)
2, 5	Simulation	0.995	2.675	8.999
	MACNESS	1.009 (1.4)	2.679 (-0.1)	9.951 (10.6)
	Boxma & Meister	1.008 (1.3)	2.679 (0.8)	7.094 (-21.1)
3	Simulation	1.145	3.748	28.870
	MACNESS	1.180 (3.1)	3.730 (-0.5)	26.659 (-7.7)
	Boxma & Meister	1.176 (2.7)	3.639 (-2.9)	23.519 (-18.5)

Table 16

$N = 5$, $\lambda_1 = \dots = \lambda_5 = 0.15$. Service time distributions are exponential with $b_3 = b_4 = b_5$, $b_1 = 5b_3$, $b_2 = 2b_3$. All switch-over times are equal to 0.2

Node	Method	Utilization (ρ)		
		0.25	0.50	0.75
1	Simulation	1.146	3.858	21.405
	MACNESS	1.184 (3.3)	3.564 (-7.6)	22.253 (4.0)
	Boxma & Meister	1.183 (3.2)	3.615 (-6.3)	23.214 (8.5)
2	Simulation	1.077	3.064	15.320
	MACNESS	1.079 (-0.2)	2.906 (-5.2)	15.540 (1.4)
	Boxma & Meister	1.082 (-0.5)	2.892 (-5.6)	14.857 (-3.0)
3 to 5	Simulation	1.050	3.113	15.574
	MACNESS	1.047 (-0.3)	2.726 (-12.4)	13.210 (-15.2)
	Boxma & Meister	1.048 (-0.2)	2.651 (-14.8)	12.070 (-22.5)

ference as to how these differences would be expected to behave in general. The errors presented in these tables are based on comparison with simulation estimates.

When the system is quite asymmetric, with one or more nodes approaching saturation (as indicated by equation (3b)), then the B&M technique stipulates a modification to their algorithm if switch-over times are not insignificant. In this modified approach, the conservation law is used to obtain the mean waiting times at the nodes which are nearing saturation. Then, these nodes are removed, and their presence in the system is accounted for by inflating the means and second moments of some of the switch-over times accordingly. The mean waiting times for the nodes in the resulting system (which now has less nodes than in the original system) is now evaluated using the conservation law once again (it is suggested that this procedure be repeated several times if necessary). While this appears to improve on the estimates, there are two potential problems with this approach. First, using this modified procedure clearly implies that the resulting mean waiting times need not now satisfy the conservation law on which the technique is based. Second, it is hard to recognize exactly when, and to what extent, this method is to be applied, namely, whether this really does improve the estimates at all, and if so how many nodes are to be removed in this manner (Boxma and Meister do present some rules of thumb for guiding this choice; however, see the remark regarding Table 11 below). Among the test cases reported, this modified approach to the B&M technique is applied in Tables 2, 8, and 10 through 13 at $\rho = 0.8$, and in Table 15 at $\rho = 0.75$. The application of this modification does not appear to improve on the estimates in Table 11. However, it does significantly improve on the estimates obtained in the other cases.

The errors reported in the tables are based on comparison with simulation estimates. In making any comparisons between the two techniques, however, it is to be noted that the simulation results could be subject to some statistical error of probably up to 10% in estimating the true mean, especially under very heavy traffic. In general, for comparing the accuracy of the two approximation techniques, we choose to ignore cases where the errors are of the order of about 5% or less under heavy traffic. Comparing MACNESS with the B&

M technique, it can be observed from the tables that both produce estimates that are very close to those obtained by simulation when the traffic is relatively low ($\rho \leq 0.5$). When the traffic is heavy and the systems are quite asymmetric, MACNESS does appear to perform significantly better than the B&M technique. This appears to be especially true when the switch-over times are zero (as in Tables 1, 3, 6, 10, and 14, for example), where the B&M techniques gives estimates that are up to about 40% away from the simulation estimates. The relative accuracy in the MACNESS estimates is significant considering that it does not call for a modification in the algorithm under heavy traffic conditions, as required by the B&M technique (for the case of systems where switch-over times are not negligible).

In general, even under conditions of heavy traffic, for the cases presented here, the estimates obtained by MACNESS are usually within 10% of the simulation estimates, with a notable exception being Table 11, where the errors are as high as about 18%. It is important to note that this is one case where the simulation results appear to be quite in error. This observation is based on the fact that the conservation law, when applied on the simulation estimates, is far from being satisfied. (It was very difficult to get better estimates here as the system is close to saturation in this example.) It is expected that when the systems are even more asymmetric and under even heavier traffic, the approximation would give larger errors. Some such cases were tested; however, the simulation results were unreliable here, and these cases are not reported, as meaningful comparisons cannot be made.

6. Summary and conclusion

The analysis of cyclic server systems with non-exhaustive service is complex. In general, even obtaining the exact mean waiting times for such systems presents considerable difficulties. Hence, a number of approximate techniques have been presented in the past.

Here, a new technique, Myopic Analysis of Cyclic Non-Exhaustive Service Systems (MACNESS), has been proposed. This technique appears to perform much better than techniques reported previously, based on extensive valida-

tions through simulations. A notable feature of the technique is that it presents an intuitively plausible explanation for the average waiting time behavior of these complex systems. This expression produces estimates that are usually very close to the mean waiting times obtained by simulations, and appears to be fairly robust even at very high utilizations, particularly when the conservation law is used to recalculate the values of the residual vacation times.

It can easily be verified that the resulting expression for the mean waiting times, given by equation (12), is exact for the completely symmetric case whenever the service times and switch-over times are both deterministic. With the modifications suggested in Section 4.2, the approximation is exact even in the case of symmetric systems where the service times and switch-over times can be random variables. The approximation has a straightforward extension for analyzing systems with zero switch-over times. It has been observed [4] that the accuracy of approximate techniques usually degrades as switch-over times tend to zero. In fact, the B&M technique does perform relatively quite poorly in some of the test cases when the switch-over times are zero. However, no noticeable change in accuracy of estimates can be noticed in the MACNESS approach. In general, based on the empirical evidence, it appears that the MACNESS performs significantly better than the B&M technique at higher utilizations.

Acknowledgment

The author benefitted considerably from several stimulating discussions on this topic with S.M. Pollock.

Appendix A. Derivations of the expressions in Propositions 3.3 and 3.4

A.1. Derivation of equation (15) in Proposition 3.3

In Section 3.3, the fraction of time that the server is present at a node m was obtained as $q(\gamma_m) = \rho_m$, and the fraction of time that the server was switching between node m and node $(m \bmod N) + 1$ was obtained as $q(\sigma_m) = s_m/c$. Extending this line of reasoning a little further,

given that the tagged customer sees the server on vacation at the time of its arrival, the fraction of time it seems the server at node m , $m \neq n$ (respectively, switching between nodes m and $(m \bmod N) + 1$), is just $q(\gamma_m)/(1 - \rho_n)$ (respectively, $q(\sigma_m)/(1 - \rho_n)$).

Now, let $\tilde{v}(n|\gamma_m)$ (respectively, $\tilde{v}(n|\sigma_m)$) denote the conditional residual vacation time, from the points of view of an arrival at node n , given that the server was found busy at node m (respectively, switching between node m and node $(m \bmod N) + 1$), at the point of arrival of the tagged customer.

Suppose that the arrival at node n found the server on vacation, and at some node m , $m \neq n$. (For ease of discussion, it is assumed that $1 \leq m < n \leq N$.) This arrival then interrupts a special service which has an expected duration of $b_m^{(2)}/b_m$. The expected residual life of this interrupted service is, of course, $b_m^{(2)}/2b_m$. Following the completion of this service, the server then needs to complete the rest of this interrupted vacation. This service interruption has, however, induced a cycle of expected length $C_m(b)$, as given by equation (10). Hence, considering a node k on the path from node m to node n , the expected number of customers served at this node would be $\lambda_k C_m(b)$. (Again, at high arrival rates, this quantity might exceed 1 and, in this case, the probability of a service at node k is assumed to be equal to 1.) Thus we obtain equation (15a).

Similarly, consider the case where the tagged customer interrupts the server switching between nodes m and $(m \bmod N) + 1$. The expected length of this interrupted switchover is $s_m^{(2)}/s_m$, and this induces a cycle of length $C_m(s)$, where

$$C_m(s) = \frac{s + s_m^{(2)}/s_m - s_m}{1 - \rho}.$$

In this case, the conditional residual vacation time $\tilde{v}(n|\sigma_m)$ is obtained as equation (15b).

Finally, the residual vacation time \tilde{v}_n is determined, by unconditioning on these two terms, as equation (15).

A.2. Derivation of equation (16) in Proposition 3.4

Implicit in the discussion here is the understanding that the tagged customer arrives at node n and finds one or more customers already present at the node.

Consider the position of the server at the instant of arrival of the tagged customer. With probability $1 - x_n$, the arrivals finds the server at node n . In this case, it interrupts a special service of expected duration $b_n^{(2)}/b_n$. As discussed earlier (refer to equation (10)), this induces a special cycle of expected length $C_n(b)$, with a corresponding vacation of expected length v_n . This accounts for the first term on the right-hand side.

Suppose, on the other hand, that the tagged customer, on arrival at node n , finds the server away from the node. This happens with probability x_n . In this case, the arrival interrupts a special vacation. Suppose that the server was performing a service at node m , $m \neq n$. The fraction of time this occurs is just $q(\gamma_m)/(1 - \rho_n)$, and the expected length of the interrupted service is $b_m^{(2)}/b_m$. This was the argument used to obtain the residual life of the interrupted vacation \tilde{v}_n , wherein it was proposed that the effect of this interruption induces a special cycle, $C_m(b)$, which continues until the server reaches node n , where he now performs a 'normal' service (i.e., a service of mean duration b_n). Here, this argument is extended a little further: it is proposed that the effect of this interrupted service at node m could continue even after the server completes the service at node n , i.e., during the vacation following this service. In effect, it is proposed that this vacation is governed by either the interrupted service at node m or the service at node n , whichever dominates. Thus, in this vacation, the probability of service at a node k , $k \neq n$, is given by $\hat{\alpha}_{km}$, where

$$\hat{\alpha}_{km} = \min\{\lambda_k \max[C_m(b), C_n], 1\}.$$

Hence, with probability $q(\gamma_m)/(1 - \rho_n)$, the expected length of a vacation, $v_n(1|\gamma_m)$, which follows a service at node n on the customer at the head of the queue, is given by

$$v_n(1|\gamma_m) = s + \sum_{k, k \neq n} \hat{\alpha}_{km}.$$

Suppose, on the other hand, that the server was switching between node m and node $(m \bmod N) + 1$, at the time of arrival of the tagged customer. The fraction of time this occurs is $q(\sigma_m)/(1 - \rho_n)$. In this case, adopting an entirely similar argument, the expected length of this vacation following the service on the customer at the head of the queue at node n , $v_n(1|\sigma_m)$, is given by

$$v_n(1|\sigma_m) = s + \sum_{k, k \neq n} \tilde{\alpha}_{km},$$

where

$$\tilde{\alpha}_{km} = \min\{\lambda_k \max[C_m(s), C_n], 1\}.$$

Hence, the expected length of a vacation following the service on the customer at the head of the line, denoted $v_n(1)$, is obtained by the weighted average of these conditional cycle times as given by equation (17).

Given that the arriving customer saw the server away from the node on arrival, the special vacation following the departure of the customer at the head of the line is just $v_n(1)$. Since this occurs with probability x_n , the result follows.

References

- [1] R. Berry and K.M. Chandy, Performance models of token ring local area networks, *Performance Evaluation Review (Special Issue)* (1983) 266–274.
- [2] O.J. Boxma, Two symmetric queues with alternating service and switching times, *Proc. Performance '84* (North-Holland, Amsterdam, 1984) 475–490.
- [3] O.J. Boxma and B.W. Meister, Waiting-time approximations for cyclic-service systems with switchover times, *Performance '86 and ACM SIGMETRICS 1986, Proc. Joint Conf. on Computer Performance Modeling Measurement, and Evaluation*, Raleigh, NC (May 1986) 254–262; also: *Performance Evaluation* 7 (4) (1987) 299–308.
- [4] O.J. Boxma and B.W. Meister, Waiting-time approximations in multi-queue systems with cyclic service, *Performance Evaluation* 7 (1) (1987) 59–70.
- [5] W. Bux, Local area subnetworks: A performance comparison, *IEEE Trans. Commun.* COM-29 (10) (1981) 1465–1473.
- [6] W. Bux and H.L. Truong, Mean-delay approximation for cyclic-serve queueing systems, *Performance Evaluation* 3 (3) (1983) 187–196.
- [7] R.T. Carsten, E.E. Newhall and M.J.M. Posner, A simplified analysis of scan times in an asymmetric Newhall loop with exhaustive service, *IEEE Trans. Commun.* COM-25 (9) (1977) 951–957.
- [8] R.B. Cooper, Queues served in cyclic order: Waiting times, *Bell System Tech. J.* 49 (3) (1970) 399–413.
- [9] R.B. Cooper and G. Murray, Queues served in cyclic order, *Bell System Tech. J.* 48 (3) (1969) 675–689.
- [10] M. Eisenberg, Queues with periodic service and change-over times, *Oper. Res.* 20 (2) (1972) 440–451.
- [11] M. Eisenberg, Two queues with alternating service, *SIAM J. Appl. Math.* 36 (1979) 287–303.
- [12] M.J. Ferguson and Y.J. Aminetzah, Exact results for non-symmetric token ring systems, *IEEE Trans. Commun.* COM-33 (3) (1985) 223–231.
- [13] S.W. Fuhrmann, Symmetric queues served in cyclic order, *Oper. Res. Lett.* 4 (3) (1985) 139–144.
- [14] O. Hashida, Analysis of multiqueue, *Rev. Electr. Commun. Lab.* 20 (3&4) (1972) 189–199.

- [15] G. Kimura and Y. Takahashi, Diffusion approximation for a token ring system with nonexhaustive service, *IEEE J. Selected Areas Commun. SAC-4* (6) (1986) 794–801.
- [16] L. Kleinrock, *Queueing Systems, Vol. I: Theory* (Wiley, New York, 1975).
- [17] P.J. Kuehn, Multiqueue systems with non-exhaustive cyclic service, *Bell System Tech. J.* **58** (3) (1979) 671–698.
- [18] J.D.C. Little, A proof of the queueing formula $L = \lambda W$, *Oper. Res.* **9** (1961) 383–387.
- [19] V. Rego and W. Szpankowski, *Closed-Network Duals of Multiques with Application to Token-Passing Systems*, Tech. Rept. CSD-TR-660, Purdue Univ., February 1987.
- [20] H. Takagi, Mean message waiting times in symmetric multi-queue systems with cyclic service, *Performance Evaluation* **5** (4) (1985) 271–277.
- [21] H. Takagi, *Analysis of Polling Systems* (The MIT Press, Cambridge, MA, 1986).
- [22] K.S. Watson, Performance evaluation of cyclic service strategies—A survey, *Performance '84* (Elsevier Science Publishers/North-Holland, New York, 1984).
- [23] P.D. Welch, On a generalized M/G/1 queueing process in which the first customer in a busy period receives exceptional service, *Oper. Res.* **12** (1964) 736–752.
- [24] R.W. Wolff, Poisson arrivals see time averages, *Oper. Res.* **30** (1982) 223–231.