UNE IMP

EHE CWI

QMIPS

INR UER

ZAR UTO

Quantitative Modeling in Parallel Systems

Waiting Time Distribution in a Class of Wireless Multi-Channel Local Area Networks

Peter G. Harrison   Sérgio Coury
Department of Computing
Imperial College
180 Queen's Gate, London SW7 2BZ

D W2.T1.11

22 September 1994

Distribution Level: All partners

Approved by: Peter G. Harrison

Abstract The waiting time distribution in a class of wireless multi-channel local area networks is derived. The networks comprise base stations controlling several channels operating under a time-slotted sharing policy. This system maps naturally onto a model in which the inter-visit time has a probability mass function of phase-type. We derive a set of matrix equations with efficient iterative procedures for their solution and controllable accuracy in their numerical evaluation. Furthermore, the methodology can be extended to several other polling strategies.

The QMIPS Project The QMIPS project is a collaborative research project supported by the CEC as ESPRIT-BRA project no 7269. It is being carried out by the following organisations: CWI (Amsterdam), EHEI (University of Paris V), Imperial College (London), INRIA (Sophia-Antipolis), University of Erlangen, University of Newcastle, University of Torino and University of Zaragoza.

1 Introduction

Cellular Networks represent the extension of the packet radio technology into a global environment. They are a result of a natural evolution of packet radio networks, enabling them to become an integrated part of a wider packet exchange network. Their topology extends over a large geographical area formed by restricted regions (called *cells*), each having one central station functioning as part of a high speed network, and several mobile stations communicating with it through radio links.

A common strategy used for the management and access control of the stations transmission channels is polling. There are various analytical models for these systems based on several queues visited by a single server in a cyclic arbitration order. The main performance measure of interest in such systems is the sojourn time of the messages at the transmission queues.

Medium access control systems which have channel activity controlled by time-slotted policies operate at discrete time intervals. Although some important results concerning waiting and sojourn times in discrete systems are available in the literature (e.g. [9], [12], [11]), analytical results are available only in some cases. Some exact results, involving average delays were derived using Kleinrock's *work conservation principle* [1]. Several other methods have been developed for computing the mean delay, the mean queue length for each queue, the amount of work of the server and the cycle time. A survey of these methods is given by Levy and Sidi(see [6]).

This paper tackles the problem of finding an approximate method for evaluating the waiting time probability density function in multi-queue systems with discrete service time. More abstractly, it also considers the sojourn time distribution of systems with exceptional first service times. A direct application for such systems can be found in time-slotted medium access protocols. In these systems several traffic sources compete to use a group of transmission channels. These transmitters are available only at the beginning of fixed length time-slots. At each discrete interval a single server starts visiting queues in cyclic order and assigning packets to the transmitters.

In order to evaluate these distributions, we introduce a new way of looking at the server. Our approach is based on the fact that the server inter-visit time can be regarded as a random variable represented by transitions in a discrete-time Markov chain. Each state of this Markov chain represents the server position at the beginning of a time-slot. We show how to construct the transition probability matrix of this Markov chain and then derive a recursive matrix equation that yields the probability mass function of the inter-visit time. Based on this equation we obtain an equivalent phase-type system that produces the same probability mass function, following the modelling approach presented by Neuts (see [2], [3], [4]). This approach is particularly suitable for our problem, since it produces a compact matrix formula for the Laplace transform of the sojourn time probability density function, from which its moments and the probability density function itself follow. Furthermore, these equations are easily converted into iterative procedures for numerical evaluation. These procedures can be written down straight from the equations, producing results of great accuracy since they are based on the expansion of rapidly convergent series.

The paper is organised as follows: section 2 describes the network architecture and its protocols; section 3 describes the model to be analysed and the assumptions made; section 4 reviews basic properties of the phase-type renewal process, and introduces two new results required for our analysis which extends the results presented in [2] for the residual service time and waiting time probability density functions. Section 5 presents a model of the system based on a discrete Markov chain. This section also shows how this chain can be mapped into a phase-type probability mass function. The solution is presented in algorithmic form in section 6 with numerical results being given in section 7. Finally, in section 8 we give our conclusions.

2 Network Architecture

The network of stations considered here can be considered as an extension of the network described in [8]. It consists of a group of Base Stations (BS) and of Portable Stations (PS). A typical portable station can be either a laptop computer or an electronic point-of-sale. The base stations are devices which are connected to a wired packet exchange network. They can communicate with the portable stations via multi-channel radio links. A typical transaction would be a laptop having access to a remote host computer or a point-of-sale consulting a central data-base.

Base stations are located in a cellular pattern to provide continuous coverage over a wide area. The topology inside each cell is centralised, i.e., the portables communicate only with the base station and vice-versa. The base station is responsible for forwarding their messages to their final destinations. No direct link between portables is available, however their communication is possible using the base station as an intermediate node.

The cellular technique of spatial frequency reuse is employed, which implies that the total bandwidth is divided into groups of frequencies and these groups are assigned to the cells.

The novel medium access protocol described here is a combination of a multi-channel TDMA (time division multiple access) and polling. It is a centrally controlled protocol, since the right to access each channel is controlled by the base station. The allocated group of channels is divided into three sub-groups: control, uplink data transfer and downlink data transfer channels. In all channels, packets are transmitted at discrete time intervals referred to as *time slots*.

The control sub-group contains one broadcast channel used to transmit system packets (SYS) from the base station to portables. The SYS packet contains a Header (H), the Data transfer Assignment Table (DAT) and a Frame Checking Sequence (FCS).

The Header contains a synchronisation preamble and the base station identification. The synchronisation preamble allows the portable and the base station to operate with the same timing. The base station identification informs the portable location, and gives the information regarding the number of available channels in the cell and it corresponding frequencies.

The Data transfer Assignment Table is divided into two parts. The first part informs which portables are addressed in the downlink transmission during the next time-slot. This allows the selected portables to tune in the corresponding frequencies and receive the packet. The second part informs which stations can access the uplink channels in the next time-slot (it is the equivalent of a poll command). If the portable has a packet to transmit it tunes to the channel selected and transmits its message, otherwise a null reply is sent.

The base station reserves a *contention time-slot* in certain periods. This is used when a new portable has moved into a cell and it wants to inform of its existence. Since more than one station can transmit in the same contention slot, the sending station waits for an acknowledgement. After receiving a correct acknowledgement, the portable can wait for a time-slot dedicated to it in the downlink channels.

Figure 1 gives an example of a cell containing 3 channels for uplink, 3 channels for downlink data transfer and 6 portable stations.

We can see that the contents of the first DAT shown in fig. 1 indicate that the mobile stations number 1,4 and 6 are addressed during the next time slot in the downlink channels number 1,2 and 3 respectively. The DAT also indicates that the mobile stations number 2,3 and 5 are allowed to transmit a packet during the next time slot. The selected mobile stations (i.e. the stations number 2,3, and 5) start their transmission in the uplink channels number 1,2 and 3 respectively.

If there is no packet to be transmitted to a mobile station the DAT contain an X. In the same way, if no mobile station is selected to transmit a packet the DAT will contain the code C, which indicates a contention time slot.

This protocol has been designed to provide a good channel utilisation, since the base station

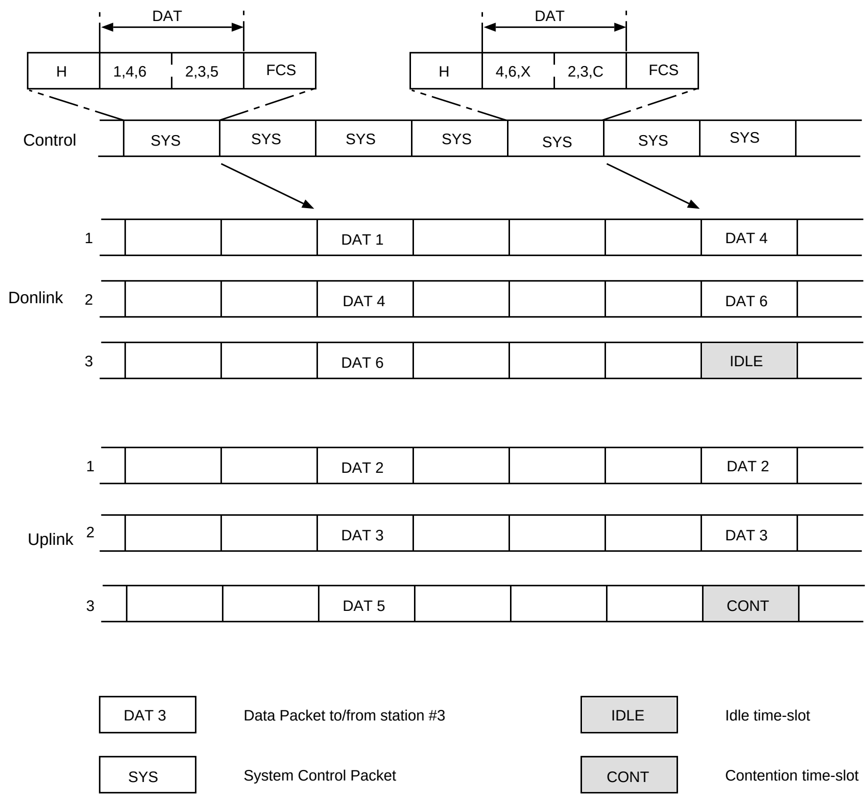

Figure 1: MCN medium access protocol.

can minimise the number of idle time slots by allocating different packets to the channels. The base station can also control how frequently the stations are polled, providing in this way a dynamic traffic control.

3 Model Description

We model the network architecture by a queuing system with N parallel queues $(Q_1, \ldots, Q_N)$ representing the channels. Each queue has infinite buffer capacity to store packets, i.e. its customers (see fig 2). Clearly, if the buffer capacity is big enough in the actual system, i.e. the probability of packet loss is small, this is a good approximation. Although we make some further approximating assumptions, our model structure reflects faithfully the principles of operation of the network described in section 2.

The packets in each queue wait to be placed in a *channel transmit buffer* (which effectively is the head of their queue), administered by a *channel controller* on a first-come-first-served basis. This system can be regarded as N independent external Poisson streams, generating packets to the parallel queues, each one with an arrival rate of λ_i, $(1 \leq i \leq N)$.

The channel controller operates at discrete time intervals τ_n $(n = 0, 1, 2, \ldots)$ called *time-slots*. It is assumed that the interval between time-slots is greater than the transmission time of each packet. At each time τ_n, a group of K transmitters are available for simultaneous operation. The channel controller starts visiting the queues in cyclic order. If the channel transmit buffer of the visited queue is empty, then the channel controller immediately visits the next queue, otherwise the packet is removed and allocated to the first transmitter available.

The channel controller repeats this operation until either all the transmitters have been filled or all queues have been visited. At time τ_{n+1} the first queue to be visited is the next in sequence.

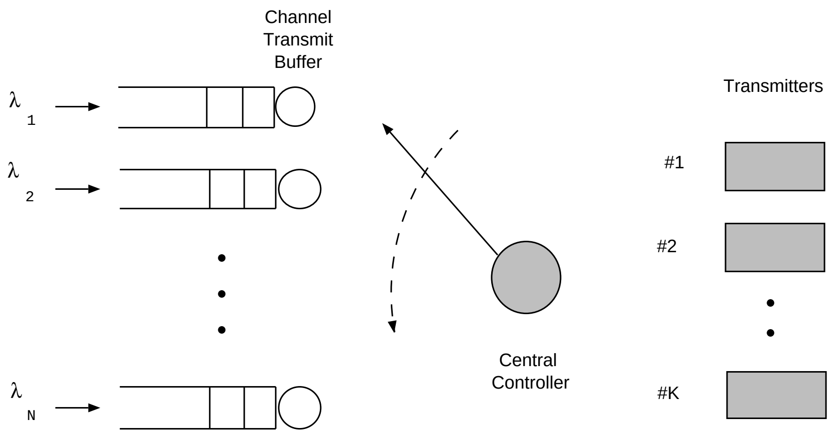

Figure 2: Base station queuing system.

We suppose that the time spent during the queue scanning operation is negligible compared with the time-slot duration.

This allocation policy prevents two packets from the same queue being transmitted at the same time, as required in our model. Packets from the same queue generally go to the same destination (the mobile station). This destination is usually unable to receive more than one packet per time-slot. However, the model is easy to adapt to other protocols, for example where the transmitters are allocated until they or all waiting packets are exhausted.

We consider *sojourn time* as being the time spent by a packet in the queue (including the time spent in the channel transmit buffer), together with its *service time*, i.e. the time spent in the channel transmit buffer.

4 Probability distributions of phase-type

The class of probability distributions of phase-type and some of its properties are summarised in this section. An extensive and comprehensive treatment of this subject can be found in [2]. We also introduce two new theorems, that will be used in section 5.

Consider a discrete Markov chain with $(m+1)$ states (called *phases*) and transition probability matrix given by:

$$P = \begin{pmatrix} S & \mathbf{S}^\circ \\ \mathbf{0}' & 1 \end{pmatrix} \tag{1}$$

where S is a $(m \times m)$ sub-stochastic matrix, such that $I - S$ is non-singular, $\mathbf{S}^\circ$ is a column vector such that P is stochastic and $\mathbf{0}'$ is the null vector. We denote column vectors in bold type and row vectors also in bold type with a prime operator. Obviously the $(m+1)$ state is absorbing. We suppose that once an absorption occurs, the process is immediately re-started in state i $(1 \le i \le m+1)$ with probability β_i. Thus, the initial probability vector is (β', β_{m+1}), where $\beta' = (\beta_1 \ldots \beta_m)$.

The probability mass function $\{p_k\}$ (called *phase-type* or *PH*) of the number of steps $\mathcal{N}$ until absorption has *representation* (β', S) and is given by:

$$p_k = \begin{cases} \beta_{m+1} & k = 0 \\ \beta' S^{k-1} \mathbf{S}^\circ & k = 1, 2, \ldots \end{cases} \tag{2}$$

Note that $\mathbf{S}^\circ$ is uniquely determined by S. In the analysis presented here we will assume $\beta_{m+1} = 0$. For this case, the probability generating function, Laplace Transform (i.e. the expectation $E[e^{-s\mathcal{N}}]$) and the factorial moments of the probability mass function $\{p_k\}$ with representation (β', S) are given by:

$$\mathcal{P}(z) = z\beta'(I - zS)^{-1}\mathbf{S}^\circ \tag{3}$$

$$f(s) = \mathcal{P}(e^{-s}) = e^{-s}\beta'(I - e^{-s}S)^{-1}\mathbf{S}^\circ \tag{4}$$

$$\mathcal{P}^{(k)}(1) = k!\beta' S^{k-1}(I - S)^{-k}\mathbf{e} \tag{5}$$

where $\mathbf{e}' = (1, 1, \ldots, 1)$

We further consider an equivalent version of the Markov chain in which the absorbing state $m+1$ is removed; it is replaced by an "invisible" instantaneous state. We obtain a Markov chain with m states and transition probability matrix given by

$$S^* = S + S^\circ B^\circ \tag{6}$$

where S° is a $(m \times m)$ matrix with identical columns $\mathbf{S}^\circ$ and $B^\circ = diag(\beta_1, \ldots, \beta_m)$. The representation (β', S) is called *irreducible* if and only if the matrix S^* is non-singular. Clearly, matrix S^* is stochastic and the stationary solution of the equations $\theta' S^* = \theta'$ and $\theta'\mathbf{e} = 1$ is given by:

$$\theta' = \mu\, \beta'(I - S)^{-1} \tag{7}$$

where $\overline{m} = \mu^{-1}$ is the mean of the probability mass function with representation (β', S), which can be obtained using equation 5. The following two theorems are the discrete case versions of the theorems presented in [2].

Theorem 1 *If p_k is a discrete distribution of phase-type with (irreducible) representation (β', S) and generating function $\mathcal{P}(z)$, then*

$$\mathcal{Q}(z) = \mu z\, \frac{1 - \mathcal{P}(z)}{1 - z} \tag{8}$$

is the generating function of the discrete PH distribution with representation (θ', S), where θ' is the stationary solution of the equations $\theta'\, S^* = \theta'$ and $\theta'\mathbf{e} = 1$.

Theorem 2 *If p_k is a discrete PH distribution with (irreducible) representation (β', S), then the convolution of p_k with itself is also a discrete PH-distribution with representation (γ', L) given by:*

$$\gamma' = (\beta', \mathbf{0}') \tag{9}$$

$$L = \begin{pmatrix} S & S^\circ B^\circ \\ 0 & S \end{pmatrix} \tag{10}$$

Proof

This theorem is a particular case of the theorem presented in [2].

5 Analysis of the model

We now concentrate on the analysis of one queue in our packet-switching system, let us say Queue 1 (called the *reference queue*). Associated with this queue, we consider a N-state Markov chain, where each state represents the queue to which the channel controller is pointing at times τ_n, after the transmitter allocation. This Markov chain has a $(N \times N)$ transition probability matrix $T = [t_{ij}]$ where t_{ij} is the probability of the server pointing to queue j at τ_{n+1} given that it was pointing to queue i at τ_n.

Assuming the system to be in a steady state, let σ_i be the probability of the channel controller finding the channel transmit buffer of queue i non-empty ($1 \leq i \leq N$). We make an approximating assumption that the probabilities σ_i are independent. The validity of this approximation is considered in section 7. At each discrete time interval the channel controller visits queues $i+1$,$i+2$,...,$i+j$ (for simplicity no special notation will be used to denote arithmetic $mod(N)$). The probabilities t_{ij} can be evaluated in three different cases:

(a) When $i+1 \leq j \leq i+K-1$, we have $t_{ij} = 0$, since at least K queues are visited every time-slot.

(b) When $i+K \leq j \leq i-1$ we have a situation in which the K transmitters were filled, therefore queue $j-1$ was non-empty. Thus we have:

$$t_{ij} = \sigma_{j-1}\, P[K-1 \text{ non-empty queues between } i \text{ and } j-2] \tag{11}$$

Let ψ be a set of $K-1$ queues between i and $j-2$, and Ψ the set of all possible ψ. Then, under the independence assumption above, the probability t_{ij} is given by:

$$t_{ij} = \sigma_{j-1} \sum_{\psi \in \Psi} \prod_{m \notin \psi} (1-\sigma_m) \prod_{n \in \psi} \sigma_n \tag{12}$$

(c)When $i = j$ we have a situation in which not necessarily all transmitters were filled. Let ψ_x be a set of any x queues and Ψ_x the set of all possible ψ_x. The probability t_{ii} is given by:

$$\begin{aligned} t_{ii} \;\; = \;\; & \sum_{x=0}^{K-1} \sum_{\psi_x \in \Psi x} \prod_{m \notin \psi_x} (1-\sigma_m) \prod_{n \in \psi_x} \sigma_n \\ & + \sigma_{i-1} \sum_{\psi_{K-1} \in \Psi_{K-1}} \prod_{m \notin \psi_{K-1}} (1-\sigma_m) \prod_{n \in \psi_{K-1}} \sigma_n \end{aligned} \tag{13}$$

The first term of equation 13 evaluates t_{ii} when there are less than K non-empty channel transmit buffers. The second term evaluates t_{ii} when there are exactly K non-empty channel transmit buffers. This occurs only when the channel controller stops after queue $i-1$ (which is not empty).

5.1 The Inter-Visit Time Distribution

Consider now a discrete random variable Y which describes the number of time-slots between consecutive visits of the server to the reference queue. Let $r_k = P[Y = k]$. We can easily show that there is a steady state solution for the system of equations $\pi' T = \pi'$ and $\pi' \mathbf{e} = 1$ (the Markov chain defined by T is clearly irreducible, aperiodic and has finite state space). We first evaluate r_1 and r_2 and then generalise by giving an expression for r_k, which is easily verified by induction.

We know that the reference queue is visited whenever there is a jump in the corresponding Markov chain from a state i to a state $j \leq i$. Thus, we have:

$$r_1 = \pi_1(t_{11}) + \pi_2(t_{21} + t_{22}) + \ldots + \pi_N \sum_{j=1}^{N} t_{Nj} = \sum_{i=1}^{N} \sum_{j \leq i} \pi_i t_{ij} \tag{14}$$

This uses the fact that an equilibrium state exists just after a polling cycle with a visit to queue 1, since this queue was chosen arbitrarily. Using matrix notation for equation 14 yields:

$$r_1 = \pi' R \mathbf{e} \tag{15}$$

where R is the lower triangular part of the matrix T, defined by $r_{ij} = t_{ij}$ if $i \geq j$; $r_{ij} = 0$ otherwise. The probability r_2 is calculated using the fact that there are two jumps in the corresponding Markov chain: the first goes from a state i to a state $j > i$ and the second goes from a state j to a state $k \leq j$. Thus we have:

$$r_2 = \sum_{i=1}^{N} \sum_{j>i} \sum_{k \leq j} \pi_i t_{ij} t_{jk} \tag{16}$$

Using the matrix form yields:

$$r_2 = \pi' (T - R) R \mathbf{e} \tag{17}$$

The probability r_k is calculated in the same way, i.e., by considering that there are $k-1$ intermediate jumps in the corresponding Markov chain to states numbered successively higher than the initial state, and one jump to a lower numbered state. Thus the expression for r_k can be written:

$$r_k = \pi'(T - R)^{k-1} R \mathbf{e} \tag{18}$$

We observe that this probability mass function is equivalent to a phase-type with representation $(\pi', (T - R))$, since it produces the same probabilities $\{r_k\}$. This representation is reducible, since $(T - R)^*$ is singular (all elements of its two last rows are identical). An irreducible representation of these probabilities can be obtained by defining a reduced matrix S and vector β and solving a system of linear equations (this method is discussed in more detail in the Appendix). Henceforth, we assume (β', S) to be an irreducible representation of the probabilities $\{r_k\}$.

The system so formed can be regarded as a $M/G/1$ queue visited by a single server at discrete times. The visit times can be considered as a renewal process of phase-type with representation (β', S).

5.2 The Reference Queue Operation

A typical reference queue operation is depicted in Fig. 3. We can see that instants t_i correspond to the arrival of a customer i (denoted A_i). The instants T_i correspond to a visit of the channel controller and a departure of a customer (if any).

In the example shown we observe that customer 1 found the queue empty, therefore its sojourn time is given by the time until the next visit of the server. We call this time the *residual time* of the inter-visit renewal process. The probability mass function $\{r_k\}$ of the renewal period of this process is a discrete phase-type with (irreducible) representation (β', S).

From the same example we also observe that the sojourn time of customer 3 is the residual time plus the inter-visit time, since it found one customer already in the queue when it arrived. This shows us that the reference queue operates as a queue with *exceptional first service times* (see [14] and [15]), where the first service is the time until the next renewal, and ordinary service is the inter-visit time.

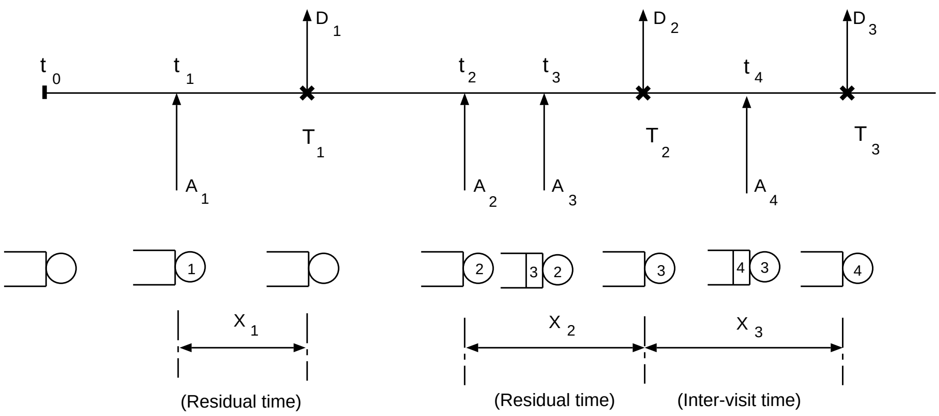

Figure 3: Reference queue operation.

We proceed by considering the sojourn time of the reference queue. Let θ' be the stationary solution of the system of equations $\theta' S^* = \theta'$ and $\theta' \mathbf{e} = 1$. Let $\{p_k\}$ be the probability mass function of the PH-distribution with representation (θ', S) and the vector $\mathbf{q} = \{q_k = p_{k+1} - p_k | k = 0, 1, \ldots\}$.

Let $\overline{m}$ be the mean inter-visit time and $\rho = \lambda\overline{m}$. The following definitions will be used in sequence:

Definition 1 *Let the matrix Q contain the convolutions of the vector $\mathbf{q}$ with itself, i.e. $Q = [q_{ij}]$ and q_{ij} is the i-th element of the j-th convolution of $\mathbf{q}$ with itself $(i, j = 0, 1, \ldots)$. We also denote $\mathbf{q}_i^{*'}$ as the i-th row of Q. We note that the 0-th column of Q is the vector $\mathbf{q}$ itself (i.e. $q_{k0} = q_k$).*

Definition 2 *Let the matrix $E(t)$ contain the elements of the serial expansion of the functions $exp(\rho(t-i))$ $(i = 0, 1, \ldots)$, i.e. $E = [e_{ij}(t)]$ and $e_{ij}(t) = \frac{\rho(t-j)^i}{i!}$. We also denote $\mathbf{e}_j(t)$ as the j-th column of E.*

Definition 3 *Let the operation d of a square matrix A (denoted to as A^d) produce a vector which contains the elements of the diagonal of A.*

The following two theorems in sequence evaluate the residual time and the waiting time probability density function of a queue with phase-type service time of representation (θ', S). Their proofs are given in [13]. They are useful for evaluating the sojourn time probability density function of the reference queue (whose first and ordinary service time have distributions $\hat{r}(t)$ and $\{r_k\}$ respectively).

Theorem 3 *For a renewal process with discrete renewal period distribution of phase type with representation (β', S), having mass function $\{r_k\}$, the stationary probability density function of the residual life is given by:*

$$\hat{r}(t) = \mathbf{q}' \mathbf{H}(t) \tag{19}$$

where $\mathbf{H}(t) = (H(t), H(t-1), \ldots)$ and $H(t)$ is the unit-step function ($H(t) = 1$ if $t > 0$, $H(t) = 0$, otherwise).

Remarks

1. Notice that time is normalised with respect to the time-slot duration $\tau_{n+1} - \tau_n$.

2.The first and second moments of the residual time probability density function are given by:

$$\overline{m}_r = -\frac{d}{ds}\hat{r}^*(s)\mid_{s=0} = \theta'(I-S)^{-1}\mathbf{e} - \frac{1}{2} \qquad (20)$$

$$\overline{\overline{m}}_r = \frac{d^2}{ds^2}\hat{r}^*(s)\mid_{s=0} = \frac{1}{3} + 2\,\theta' S(I-S)^{-2}\mathbf{e} \qquad (21)$$

Theorem 4 *In the stable $M/G/1$ queue with service time probability mass function $\{r_k\}$ of phase-type with representation (β', S), the stationary probability density function of the waiting (queuing) time is given by:*

$$w(t) = (1-\rho)\,[\delta(t) + \rho\,(QE(t))^d\,\mathbf{H}(t)] \qquad (22)$$

where $\rho = \lambda\overline{m}$ and $\overline{m}$ is the mean service time.

Remarks

1. Notice again that time is normalised with respect to the time-slot duration.

2.For $0 \le t \le 1$ w(t) can be explicitly derived as:

$$w(t) = (1-\rho)\,[\delta(t) + \rho\,exp(\theta'\,\mathbf{S}^\circ\,\rho\,t)] \qquad (23)$$

3.The first and second moments of the waiting time distribution are given by:

$$\overline{m}_w = -\frac{d}{d\,s}w^*(s)\mid_{s=0} = \frac{\rho\overline{m}_r}{(1-\rho)} \qquad (24)$$

$$\overline{\overline{m}}_w = \frac{d^2}{ds^2}w^*(s)\mid_{s=0} = \frac{\rho\,[2\rho(\overline{m}_r)^2 + \overline{\overline{m}}_r(1-\rho)]}{(\rho-1)^2} \qquad (25)$$

The following theorem (its proof is given in [13]) considers the sojourn time of a packet in the reference queue. It operates with exceptional first service time, where the distribution of the ordinary service time is discrete phase-type of representation (β', S) and the distribution of the first service time is $\hat{r}(t)$.

Let the matrices $\dot{E}(t) = [\dot{e}_{ij}(t)]$ and $\ddot{E}(t) = [\ddot{e}_{ij}(t)]$ be defined as follows:

$$\begin{aligned} \dot{e}_{ij}(t) &= e_{i+1\,j}(t) = \frac{[\rho(t-j)]^{i+1}}{(i+1)!} \\ \ddot{e}_{ij}(t) &= (i+1)e_{i+2\,j}(t) = (i+1)\frac{[\rho(t-j)]^{i+2}}{(i+2)!} \quad (i,j = 0,1,\ldots) \end{aligned} \qquad (26)$$

As in definition 2, the j-th columns of $\dot{E}(t)$ and $\ddot{E}(t)$ are referred to as $\dot{\mathbf{e}}_j(t)$ and $\ddot{\mathbf{e}}_j(t)$ respectively.

Let the matrix $\Gamma(t) = [\gamma_{ij}(t)]$ be defined as follows:

$$\gamma_{ij}(t) = q_i[\varphi - (t-i-j)\,\mathbf{q}_j^{*'}\dot{\mathbf{e}}_{i+j}(t) + \frac{1}{\rho}\,\ddot{\mathbf{e}}_{i+j}(t)]H(t-i-j) \qquad (27)$$

where $\rho = \lambda\overline{m}$, $\overline{m}$ is the mean ordinary service time and $\varphi = \frac{1}{\lambda} - \overline{m}$.

Let the vector $\mathbf{u}(t) = \{u_i(t) | i = 0,1,\ldots\}$ be defined as follows:

$$u_i(t) = [\varphi - (t-i)]H(t-i) \qquad (28)$$

Theorem 5 *In a stable $M/G/1$ queue in which the probability density function of the first service time is $\hat{r}(t)$ and the distribution of the ordinary service time is of phase-type with representation (β', S), the sojourn time has density function $W(t)$ given by:*

$$W(t) = \pi_0 \lambda[(1 + \mathbf{q}'\mathbf{u}(t))H(t) + (Q\,\dot{E}(t))^d \mathbf{H}(t) + \mathbf{e}\Gamma(t)\mathbf{e}'] \tag{29}$$

where π_0 is the steady state probability of having 0 customers in the reference queue (observed at a random instant), i.e. :

$$\pi_0 = \frac{1 - \lambda \overline{m}}{1 + \lambda(\overline{m}_r - \overline{m})} \tag{30}$$

Remarks

1. *Time is normalised with the time-slot duration.*
2. *The first moment of the sojourn time distribution is given by:*

$$\overline{m}_s = \frac{\pi_0 \lambda}{(1-\rho)^2}[(1-\rho)(\frac{\overline{\overline{m}}_r}{2} + \varphi \overline{m}_r) + \rho \overline{m}_r(\varphi + \overline{m}_r)] \tag{31}$$

6 The algorithmic solution

The algorithm presented here is based on successive convolutions of the sequence formed by the vector $\mathbf{q}$. It is assumed that the probability mass function $\{r_k\}$ of the inter-visit time probability mass function with representation (β', S) is given. The objective of the algorithm is to produce the value of $W(t)$ for a given t.

The algorithm starts reading (β', S), σ and t. From this, it constructs the probability mass function $\{p_k\}$ with representation (θ', S). The vector $\mathbf{q}$ is calculated using the iterative formula $q_k = p_{k+1} - p_k$, $(k = 0, 1, \ldots)$. The successive convolutions $q_k^{(m)}$ can be evaluated either directly from the definition or in two steps: (1) Calculating $p_k^{(m)}$ using theorem 2 and (2) Calculating the convolution of $p_k^{(m)}$ with $d_k^{(m)}$, where $d_k^{(m)}$ is the m-th convolution of the anti-transform of $z/1-z$ with itself.

Knowing that $W(t)$ is a sum of convergent series, a precision parameter ε can also be given and tested every iteration.

The algorithm can be summarised in the following steps:

Algorithm 1

1. *Read (β', S), λ and t.*
2. *Calculate the mean of the probability mass function $\{r_k\}$ using the expression*

$$\overline{m} = \frac{1}{\mu} = \beta'(I - S)^{-1}\mathbf{e} \tag{32}$$

(The parameters $\rho = \lambda\overline{m}$ and $\varphi = \lambda^{-1} + \overline{m}$ immediately follow).

3. *Calculate θ' using the expression*

$$\theta' = \mu\,\beta'(I - S)^{-1} \tag{33}$$

4. *Calculate probabilities $\{p_k\}$ of representation (θ', S) using equation 2.*
5. *Calculate the sequence $\mathbf{q}$ using the expression*

$$q_k = p_{k+1} - p_k \tag{34}$$

6. *Calculate the matrix Q by evaluating the convolutions of $\mathbf{q}$ with itself.*

7. *Create the functions $\dot{e}_{ij}(t)$ and $\ddot{e}_{ij}(t)$ given by:*

$$\begin{aligned} \dot{e}_{ij}(t) &= \frac{[\rho(t-j)]^{(i+1)}}{(i+1)!} \\ \ddot{e}_{ij}(t) &= (i+1)\, \dot{e}_{i+1\,j}(t) \end{aligned} \tag{35}$$

8. *Create the function $\gamma_{ij}(t)$ given by:*

$$\gamma_{ij}(t) = q_i[\sum_{k=0}^{max} q_{jk}\dot{e}_{k\,i+j}(t) + \rho^{-1}\sum_{k=0}^{max} q_{jk}\ddot{e}_{k\,i+j}(t)] \tag{36}$$

where $Q = [q_{ij}]$ and $\mathbf{q} = \{q_i\}\,(i, j = 0, 1, \ldots)$.

9. *Evaluate $s_1(t)$, $s_2(t)$ and $s_3(t)$ given by:*

$$\begin{aligned} s_1(t) &= 1 + \sum_{i=0}^{\lfloor t \rfloor} q_i(\varphi - t - i) \\ s_2(t) &= \sum_{i=0}^{\lfloor t \rfloor}\sum_{j=0}^{max} q_{ij}\dot{e}_{ji}(t) \\ s_3(t) &= \sum_{i=0}^{\lfloor t \rfloor}\sum_{j=0}^{\lfloor t \rfloor - i} \gamma_{ij}(t) \end{aligned} \tag{37}$$

10. *Evaluate $W(t)$ using the expression*

$$W(t) = \pi_o\lambda(s_1(t) + s_2(t) + s_3(t)) \tag{38}$$

where π_0 is calculated from equation 30.

Remarks

1. Observe that since equations 36 and 37 contain convergent series, a parameter max can be chosen so that $W(t) < \varepsilon$, where ε is the pre-defined precision parameter.

2. The matrix S of representation (β', S) is given in terms of the probabilities σ_i. These probabilities can be evaluated from the arrival rates λ_i (since $1 - \sigma_i$ is the probability of the channel controller finding queue i empty). This happens whenever queue i is empty at a random instant (this probability is given by π_{oi}) and there is no arrival until the next visit of the channel controller. Therefore, σ_i is calculated by finding the solution of the following system of equations:

$$1 - \sigma_i = \pi_{0i}\int_0^\infty e^{-\lambda x}\hat{r}(x)dx = \pi_{0i}\hat{r}^*(\lambda_i) \quad (i = 0, 1, \ldots, N) \tag{39}$$

7 Numerical results

We examine the case where $\sigma_i = \sigma$, $(i = 1, 2, \ldots, N)$. Firstly we consider the case where $N = 4$ and $K = 2$. The matrix T can be constructed using equations 12 and 13. Thus we have:

$$T = \begin{pmatrix} 1-3\sigma^2+2\sigma^3 & 0 & \sigma^2 & 2\sigma^2(1-\sigma) \\ 2\sigma^2(1-\sigma) & 1-3\sigma^2+2\sigma^3 & 0 & \sigma^2 \\ \sigma^2 & 2\sigma^2(1-\sigma) & 1-3\sigma^2+2\sigma^3 & 0 \\ 0 & \sigma^2 & 2\sigma^2(1-\sigma) & 1-3\sigma^2+2\sigma^3 \end{pmatrix} \tag{40}$$

Obviously, the steady state solution for the discrete Markov chain with state transition matrix T is $\pi' = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. In section 5.1 we defined r_k as being the probability of having k time-slots between consecutive visits to the reference queue. Using equation 18 we have:

$$r_k = \begin{cases} 0 & k = 0 \\ 1 - \sigma^2(1 - \frac{\sigma}{2}) & k = 1 \\ \frac{\sigma^2}{2}(2 - \sigma) & k = 2 \\ 0 & k > 2 \end{cases} \tag{41}$$

We observe that an irreducible representation of the probabilities $\{r_k\}$ is (β', S) given by:

$$\beta' = (\frac{1}{2}, \frac{1}{2}) \tag{42}$$

$$S = \begin{pmatrix} 0 & \sigma^2(2-\sigma) \\ 0 & 0 \end{pmatrix} \tag{43}$$

Figure 4 shows $W(t)$ for $\lambda = 0.15$, $\lambda = 0.25$ and $\lambda = 0.45$. Notice that explicit equations for its first moment are given by equation 31.

Expressions for the first and second moments of the residual time can be calculated using equations 20 and 21. Thus we have:

$$\overline{m}_r = \frac{2(\sigma^3 - 2\sigma^2 - 1)}{\sigma^3 - 2\sigma^2 - 2} - \frac{1}{2} \tag{44}$$

$$\overline{\overline{m}}_r = \frac{-7\sigma^3 + 14\sigma^2 + 2}{-6\sigma^3 + 12\sigma^2 + 12} \tag{45}$$

From the equations above the first moment $\overline{m}_s$ immediately follows.

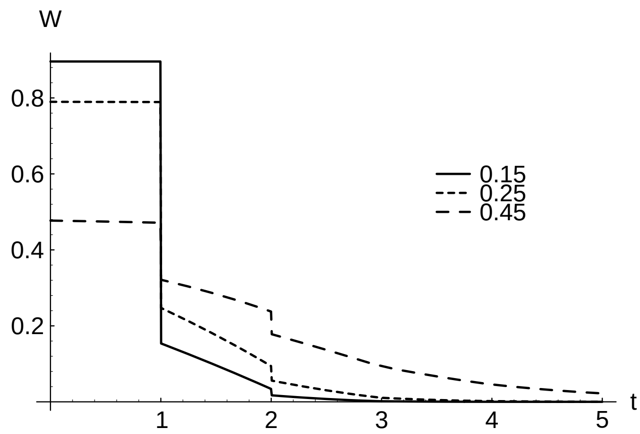

Figure 4: Sojourn time probability density function.

The sojourn time probability density function has points of discontinuity when $t = 1$ and $t = 2$. These points correspond to the points of discontinuity of $\hat{r}(t)$.

7.1 Model validation

In this section we validate our analytical model by comparing its results with the results obtained by simulation. We carried out a discrete-event simulation of the system described in section 3. A Sun/Sparc workstation was used and the program was written in the C programming language.

λ	Anal. Model	Simulation	0.95 Conf.Interval	Relative Error
0.10	0.558	0.560	(0.540,0.580)	-0.36
0.20	0.659	0.660	(0.643,0.677)	-0.15
0.30	0.834	0.849	(0.821,0.877)	-1.80
0.40	1.160	1.204	(1.152,1.256)	-3,65
0.48	1.684	2.101	(2.079,2.123)	-19.8

Table 1: Mean Sojourn Time Results.

Initially we estimated the mean sojourn time for balanced systems comprising four queues and two transmitters. The results are shown in Fig. 5 (observe that the dashed line represents the simulation results). We also present the same results in table 1. The confidence intervals were estimated using the batch means approach. Each sample ran 300,000 units of time (time-slots) and 6 samples were taken.

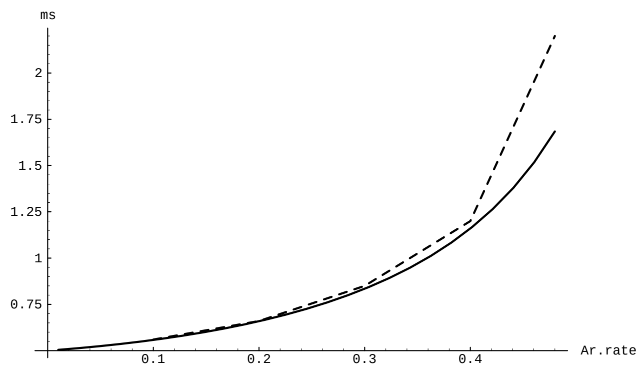

Figure 5: Mean sojourn time.

As can be seen from the table below, the analytical model is very accurate when the arrival rate is less than or equal to 0.4 packets/time unit. The relative errors are more noticeable at higher arrival rates where our analytical model underestimates the mean sojourn time. We therefore examine our initial assumption of independence of the probabilities σ_i.

Let Z_i be a random variable which is equal to 0 when the channel controller finds queue i empty, otherwise it is equal to 1 (obviously $\sigma_i = P(Z_i = 1)$). We use our simulation model to estimate the covariance and correlation between Z_i and Z_j ($j \neq i$). We also estimate the autocorrelation function of Z_i regarded as a time series and the results are presented below.

7.1.1 Estimation of the covariance and correlation

We used our simulation model to collect 2000 observations of the random variables Z_i and Z_j. Although we show here the results for $i = 1$ and $j = 2$, similar results were obtained for the other cases. We estimated the means $\overline{z}_1$ and $\overline{z}_2$, the covariance $cv(Z_1, Z_2) = E[(Z_1 - \overline{z}_1)(Z_2 - \overline{z}_2)] = E(Z_1Z_2) - \overline{z}_1\overline{z}_2$, and the correlation $cr(Z_1, Z_2) = cv(Z_1, Z_2)/\sqrt{E[(Z_1 - \overline{z}_1)^2]E[(Z_2 - \overline{z}_2)^2]}$. Again, we adopted the batch means approach, where 6 samples were taken. Each sample run 300,000 units of time giving a total simulation time of 1,800,000 units.

We simulated systems comprising $N = 4$ queues and $K = 2$ transmitters (see table 2). We also simulated systems with $N = 8$ queues and $K = 4$ transmitters (see table 3). The above mentioned parameters were calculated for various total incoming traffic (referred to as

Balanced					
λ_0	$\overline{z}_1$	$\overline{z}_2$	$cv(Z_1, Z_2)$	95% conf. interval	$cr(Z_1, Z_2)$
0.5	0.127	0.130	-0.00133	(-0.00134 – -0.00130)	-0.011
1.0	0.263	0.264	0.00318	(0.00315 – 0.00321)	0.016
1.4	0.428	0.422	0.01130	(0.01116 – 0.01145)	0.046
1.8	0.640	0.652	0.02110	(0.02102 - 0.02117)	0.092
1.95	0.860	0.861	0.00910	(0.00907 – 0.00928)	0.082
Unbalanced					
λ_0	$\overline{z}_1$	$\overline{z}_2$	$cv(Z_1, Z_2)$	95% conf.interval	$cr(Z_1, Z_2)$
0.5	0.352	0.075	-0.00265	(-0.00270 – -0.00260)	-0.015
0.7	0.508	0.116	0.00612	(0.00601 – 0.00623)	0.038
0.9	0.657	0.148	0.00343	(0.00339 – 0.00347)	0.022
1.1	0.785	0.178	0.00209	(0.00200 – 0.00218)	0.015
1.35	0.964	0.212	-0.00021	(-0.00021 – -0.00020)	-0.002

Table 2: Estimation of the covariance and correlation ($N = 4$,$K = 2$).

Balanced					
λ_0	$\overline{z}_1$	$\overline{z}_2$	$cv(Z_1, Z_2)$	95% conf. interval	$cr(Z_1, Z_2)$
0.5	0.054	0.063	0.00198	(0.00190 – 0.00206)	0.019
1.5	0.205	0.180	0.00365	(0.00355 – 0.00375)	0.022
3.0	0.404	0.408	0.01347	(0.01339 – 0.01355)	0.055
3.8	0.662	0.669	0.01861	(0.01855 – 0.01858)	0.084
Unbalanced					
λ_0	$\overline{z}_1$	$\overline{z}_2$	$cv(Z_1, Z_2)$	95% conf. interval	$cr(Z_1, Z_2)$
0.4	0.304	0.064	-0.00101	(-0.00110 – -0.00092)	-0.014
0.8	0.573	0.115	0.00027	(0.00030 – 0.00024)	0.001
1.0	0.721	0.168	-0.00502	(-0.00510 – -0.00490)	-0.033
1.3	0.915	0.193	-0.00018	(-0.00020 – -0.00017)	0.001

Table 3: Estimation of the covariance and correlation ($N = 8$,$K = 4$).

$\lambda_0 = \sum \lambda_i$). We considered both balanced systems (where $\lambda_i = \lambda_0/N$) and unbalanced systems (where we assumed $\lambda_1 = 0.70\lambda_0$, $\lambda_2 = 0.15\lambda_0$, $\lambda_3 = 0.10\lambda_0$ and $\lambda_i = 0.05\frac{\lambda_0}{(N-3)}$ $(i = 4, \ldots, N)$). By observing tables 2 and 3 we can say that there is no evidence that variables Z_1 and Z_2 are correlated. We observe, for example, in table 2 that when $\lambda_0 = 1.95$ (note that the maximum total arrival rate supported is $\lambda_0 = 2.0$), we have $cv(Z_1, Z_2) = 0.009$. This is a quite small number compared with $\overline{z}_1\overline{z}_2 = 0.78$. We therefore conclude that it is not unreasonable to assume that the variables are independent.

7.1.2 Estimation of the autocorrelation function

From our simulation model, we collected 2000 observations of the variable Z_i and estimated its autocorrelation function. This function gives the correlation between Z_i observed at time t with Z_i observed at time $t + k$. The interval k between these observations is called *lag*. Clearly, successive observations of Z_i form a stationary stochastic process. We therefore used *Bartlett's formula* (see [16]) to estimate the smallest lag q beyond which the autocorrelation function is zero.

We show in table 4 the auto-correlation function for both balanced and unbalanced systems comprising four queues and two transmitters. We estimated the autocorrelation function for queue 1 (i.e. we collected observations of the variable Z_1) and we show the results for several

Balanced Systems							
λ_0	*lag*						q
	1	2	3	4	5	6	
0.5	0.07	0.01	-0.01	0.02	0.01	-0.00	0
1.0	0.24	0.11	0.04	0.03	0.05	0.04	1
1.4	0.25	0.14	0.07	0.05	0.04	-0.04	2
1.8	0.42	0.30	0.23	0.20	0.17	0.04	5
Unbalanced Systems							
λ_0	*lag*						q
	1	2	3	4	5	6	
0.5	0.16	0.09	0.06	0.04	0.03	0.00	1
0.7	0.20	0.10	0.05	0.06	0.03	0.01	1
0.9	0.28	0.18	0.13	0.07	0.07	0.05	3
1.1	0.33	0.22	0.15	0.08	0.06	0.04	3

Table 4: Estimation of the autocorrelation function ($N = 4$,$K = 2$).

values for the total incoming traffic λ_0.

The simulation results indicate that successive values of the value Z_1 are autocorrelated, but only slightly at low-medium traffic, and increasing at higher traffic. As a consequence of this, our model underestimates the probability σ_i, specially at high traffic. This is so because it does not fully capture the effect of the interdependence of successive values of Z_i. This is the "price we pay" by using a simplified Markov chain to represent the operation of the channel controller.

A more accurate model at higher traffic can be obtained by considering a phase-type process which includes the state of every queue in the multi-queue system. In this way, the probabilities t_{ij} (see section 5) will reflect the state of the system, rather than only the position of the channel controller. This would lead to greater complexity in our algorithm but still it would be more efficient than using explicit Markov chain. We leave it as a suggestion for further work.

Based on the numerical results obtained in this section we can say that our model is very accurate for $\lambda \leq 0.4$ (for systems with $N = 4$ and $K = 2$). Considering that the maximum arrival rate supported in such a system is $\lambda = 0.5$, we assume that the analytical model can be used for multi-queue systems in which the arrival rate does not exceed 80% of its total capacity.

8 Conclusion

A solution for the sojourn time distribution in multi-queue systems served in cyclic order in discrete time intervals has been presented. The analytical model is based on the inter-visit time renewal process, which we considered as being of phase-type. We have seen that its operation can be represented by a Markov chain, the state of which represents the position of the server. The advantage of this method is that it produces compact matrix equations and efficient recursive algorithms for numerical evaluation. These algorithms have been used to compare quantitatively a number of alternative cellular network designs (see [13]). Moreover, this methodology can be extended to several other polling strategies, which is a suggestion for future research.

References

[1] L. Kleinrock, *Queueing Systems, Volumes I & II*, John Wiley & Sons, 1975 and 1976.

[2] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models - an algorithmic approach*, Johns Hopkins, 1981.

[3] M. F. Neuts, "Algorithms for the waiting time distributions under various queue disciplines in the $M/G/1$ queue with service time distributions of phase-type", *Algorithmic Methods in Probability*, North-Holland Publishing Company, 1977.

[4] M. F. Neuts, "A Versatile Markovian Point Process", *J. Appl. Prob.*, **16**, pp. 764–779, 1979.

[5] O. J. Boxma and W. P. Groenendijk, "Waiting Times in Discrete-Time Cyclic-Service Systems", *IEEE Trans. Comm.*, **36**, no 2, pp. 164–170, Feb. 1988.

[6] H. Levy and M. Sidi, "Polling Systems: Applications, Modeling, and Optimization", *IEEE Trans. Comm.*, **38**, no 10, pp. 1750–1760, Oct. 1990.

[7] R. B. Cooper and G. Murray, "Queues Served in Cyclic Order", *B.S.T.J*, pp. 675–689, Mar. 1969.

[8] M.C.Davie and J.B.Smith, "A cellular packet radio data network", *Elect. and Comm. Eng. Journal*, pp. 137–143, Jun. 1991.

[9] H. Takagi, *Analysis of Polling Systems*, M.I.T. Press, 1986.

[10] F. A. Tobagi, "Multiaccess Protocols In Packet Communication Systems", *IEEE Trans. Comm.*, **28**, no 4, pp. 468 – 488.

[11] H. Kobayashi and A. G. Konheim, "Queueing Models for Computer Communications System Analysis", *IEEE Trans. Comm.*, **25**, no 1, pp. 2–28, Jan. 1977.

[12] I. Rubin and J. E. Baker, "Media Access Control for High-Speed Local Area and Metropolitan Area Communications Networks", *Proceedings of the IEEE*, **78**, no 1, pp. 168–203, Jan. 1990.

[13] S. Coury, "Quantitative Models for the Design of Cellular Networks," PhD thesis, Dept. of Computing, Imperial College, University of London, in preparation.

[14] P.G. Harrison and N.M.Patel, *Performance Modelling of Communication Networks and Computer Architectures*, Addison-Wesley Publishing Company, 1992.

[15] P.J.B. King, *Computer and Communication Systems Performance Modelling*, Prentice-Hall, 1990.

[16] G. E. P. Box and G. M. Jenkins, *Time Series Analysis*, Holden-day, 1971.