Analysis of a Dynamic Reservation Protocol for Interactive Data Services on TDMA-Based Wireless Networks

Tak-Shing Peter Yum, *Senior Member, IEEE,* and Hongbing Zhang

***Abstract*— This paper presents the dynamic reservation protocol for supporting variable-rate data services on time-division multiple-access based wireless networks. It allows a large number of data terminals to access data applications by sharing a reserved data-carrier. Through dynamic reservation data terminals can get their needed radio channels for uplink transmission without contention. The protocol performance is evaluated by queuing analysis and verified by computer simulation.**

***Index Terms*—Mobile communication, multiple priority, reservation protocol, TDMA, wireless data.**

I. Introduction

THERE is an ever-growing demand for mobile data services. To satisfy the wide range of data applications, such as database access, file transfer, credit card verification, wireless e-mail, and web browsing, research and standardization efforts for data services on the existing systems, such as DAMPS and GSM, have been going on in earnest. Bianchi *et al.* [1] have evaluated the performance of a few proposals for the general packet radio service, including the dynamic channel stealing method. Hamalainen *et al.* [2] have proposed a variable-rate reservation-access algorithm for packet data services on time-division multiple-access (TDMA) based cellular networks.

A number of new protocols have also been proposed for third-generation wireless networks. Cleary and Paterakis have proposed and evaluated a reservation random-access scheme for integrating voice and data traffic in outdoor microcellular environments [3]. Khan and Zeghlache have proposed a priority-based multiple-access scheme [4]. These and other proposals [5]–[7] are all based on integrating voice and data services in one physical channel.

In this paper, we present the dynamic reservation protocol for variable-rate data services on TDMA-based wireless networks. This protocol has the following features. First, data terminals can get their needed channels for uplink transmission without contention. Second, the reservation scheme can adapt to traffic variations to achieve superior delay/throughput performance by dynamically changing the transmission cycle lengths. Third, it can accommodate applications with different service priorities, different data rates, and different latency constraints through adaptive transmission scheduling. Fourth, although this protocol is designed for third-generation wireless networks, it can easily be adopted to second-generation mobile systems without modification to the physical layer.

The dynamic reservation protocol is introduced in Section II. Section III gives the bandwidth assignment and transmission-scheduling algorithm. Section IV presents the queuing analysis, and Section V is on numerical results and discussions.

II. Dynamic Reservation Protocol

The dynamic reservation protocol is used for data services on a dedicated data carrier in a wireless network. In time-division duplex systems, each carrier is used for both uplink and downlink directions, while in frequency division duplex systems, a pair of carriers is used. Efficient resource utilization is achieved by dynamically sharing these carriers among all active data terminals. A terminal requiring data service can set up a virtual circuit connection with the base station through a data call setup procedure. A data terminal successfully connected is called a logged terminal, and a large number of logged terminals can share one or more data carriers.

A. Virtual Circuit Connection

When a mobile terminal wants to access certain data service, it sends a connection setup request signal to the base station on the random access channel (could be the same channel for voice connection setup). In this signal, the terminal specifies the service type and its requirements, such as priority and the latency constraint. If the addition of this new connection does not violate the service quality of the existing terminals, the setup request is accepted, and a virtual circuit is assigned to the requesting terminal after a successful authentication procedure by the upper layer protocol. The assignment is announced to the terminal by a setup confirm signal. This signal consists of three parts: 1) the assigned data carrier (or carrier pair) for the terminal to use; 2) a virtual circuit number; and 3) the position of the slot for the terminal to make transmission reservations. The actual transmission of data is in cycles. At the beginning of a cycle, logged terminals can make reservations on their assigned uplink slots. The base station immediately computes a transmission schedule based on these reservations and broadcasts it to all logged terminals. Terminals then transmit in their assigned slots.

Three kinds of control signals are defined for virtual circuit connections. They are cycle_start signals, request signals, and schedule signals. The cycle_start signal is sent by the base station to indicate the beginning of a transmission cycle. Upon

Paper approved by P. E. Rynes, the Editor for Switching Systems of the IEEE Communications Society. Manuscript received January 22, 1998; revised June 1, 1999 and June 28, 1999.

The authors are with the Department of Information Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong.

Publisher Item Identifier S 0090-6778(99)09764-0.

receiving the cycle_start signal, all terminals respond with a request signal in their assigned positions. The request signal shall indicate whether a terminal wants to make a transmission reservation in the current cycle, to change its service priority, to maintain the virtual circuit, or to terminate its virtual circuit connection.

B. Uplink Access

Messages generated by terminals are segmented into data units for fitting in TDMA slots. The uplink transmission is divided into cycles. Each transmission cycle begins with a cycle_start signal sent by the base station. It initiates a sequence of request signals, one from each logged terminal. A requested slot number (RSN) field is used on the request signal to specify the number of slots requested for the current cycle. A logged terminal with no request shall send its request signals with RSN $= 0$, and a terminal who fails to respond in a few consecutive cycles will have its virtual connection terminated. For use in the current wireless networks, such as GSM and DECT (digital enhanced cordless telephone), the request signal need be placed in one TDMA slot. However, if the physical layer allows, minislot can be used for transmitting the request signals to achieve higher efficiency.

After receiving the set of request signals from the terminals, the base station assigns an appropriate number of uplink slots for each logged terminal and specifies its starting slot position. The assignments are then placed in the schedule signal and broadcast to all terminals. The starting position of the schedule signals in a transmission cycle is indicated in the schedule signal starting position field of the cycle_start signal with enough time allocated for the very simple schedule computation to be described in the next section. All terminals transmit according to the schedule and resume the monitoring of a new cycle_start signal afterwards. The base station initiates a new transmission cycle at the end of the scheduled transmission.

C. Priority Assignment

Multipriority data transmission can be achieved by assigning different access rights to different priority classes. Consider a two-class case where class 1 has higher priority than class 2. Let class 2 terminals be further divided into n equal groups, where n is a design parameter. In each transmission cycle, all class 1 terminals and one class 2 groups are allowed to make a reservation. With that, a class 2 terminal is allowed to make reservations only once every n cycles. The delay performance of the two classes can therefore be traded off by changing n. This algorithm can be extended to support three or more terminal classes in a straightforward way.

III. Assignment Algorithm

The base station performs the assignment procedure for each transmission cycle based on the information received in the request signals. The assignment results are the number of assigned slots and the starting slot location of individual logged terminals. Let N_1 be the number of class 1 terminals, N_2 be the number of class 2 terminals, and let $N = N_1 + N_2$. As pointed out in Section II-C, $N_1 + (N_2/n)$ slots are used as reservation slots in each cycle. If L is the maximum cycle length, a maximum of $m_0 = L - N_1 - (N_2/n)$ slots can be used in each cycle.

Let the number of requested slots from terminal i be denoted as r_i and let $\boldsymbol{R} = [r_1, r_2, \cdots, r_N]$. Only $1/n$ of class 2 terminals can send their reservation request each time, so there are at most $N_1 + (N_2/n)$ elements of vector $\boldsymbol{R}$ having nonzero values. Let the actual number of slots assigned to terminal i be denoted as a_i. The assignment algorithm aims at finding an appropriate assignment vector $\boldsymbol{A} = [a_1, a_2, \cdots, a_N]$ and deciding the transmission order of the terminals. If the total number of requested slots is no larger than m_0, the requests of all logged terminals can be granted and we can set $\boldsymbol{A} = \boldsymbol{R}$. Otherwise, some terminals will not get all their requested slots in the current transmission cycle. In this case, $\boldsymbol{A}$ is initialized to $\mathbf{0}$, and the total number of remaining slots m is initialized to m_0. The assignment is performed in cycles. In each assignment cycle, the smallest nonzero r_i, denoted as $r_{\min}$, is found first and a_i is updated as $a_i = a_i + \min(r_{\min}, r_i, m)$, for $0 \leq i \leq N$. At the same time, r_i and m are decremented accordingly. Repeat this cyclic process until $m = 0$.

The transmission order follows the service priority. Class 1 terminals transmit first following by class 2 terminals, etc. Among those belonging to the same class, terminals with shorter messages transmit first.

IV. Performance Evaluation

Consider a wireless network with N_1 class 1 and N_2 class 2 logged terminals sharing one TDMA carrier for their uplink data transmission.[1] Let the generation of messages by the two classes of terminals be according to Poisson processes, with rates λ_1 and λ_2, respectively, and let the message lengths for the two classes of terminals, denoted as U and V, in number of slots, be geometrically distributed, with mean values $\overline{U}$ and $\overline{V}$, respectively. For simplicity, we assume there is no limit on the transmission cycle length and N_2/n is an integer.

A. Message Delay for Class 1 Terminals

We define the message delay as the total delay from the generation of the message at a terminal to the time the message is successfully transmitted. For class 1 terminals, it consists of four parts. The first part D_1 is the reservation delay, i.e., from the generation of a message to the beginning of the transmission cycle in which the reservation for this message is made. The second part D_2 is the reservation period $a = N_1 + N_2/n$. The third part D_3 is the queuing delay for transmission. The last part D_4 is the message transmitting time. Since D_2 is a constant and $E[D_4] = U$, we only need to derive $E[D_1]$ and $E[D_3]$.

B. Reservation Delay

A random arrival of a class 1 message will intercept a particular transmission cycle of length Y, and D_1 is the

[1] Here we assume the use of frequency-division duplex channels. For time-division duplex channels, the analysis is similar.

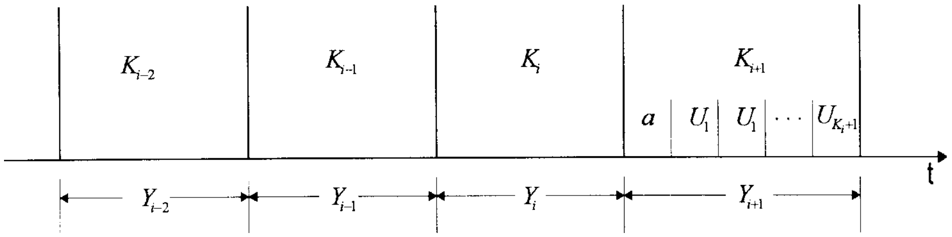


Fig. 1. The transmission cycles.

waiting time until the end of that transmission cycle. From renewal theory [8, pp. 169–174], we obtain

$$E[D_1] = \frac{E[Y^2]}{2E[Y]} = \frac{\overline{Y}^2}{2\overline{Y}}. \tag{1}$$

Let Y_i be the length of cycle i, and let K_i be the number of messages transmitted in it.

As there are two classes of messages, we further let $K_i^{(1)}$ denote the number of class 1 messages and $K_i^{(2)}$ be the number of class 2 messages. The class 1 messages are generated in cycle $i-1$. But the $K_i^{(2)}$ class 2 messages are generated in cycle $i-1$ of number $K_i^{(2)(1)}$ and in cycle $i-2$ of number $K_i^{(2)(2)}$. Combining, we have

$$K_i = K_i^{(1)} + K_i^{(2)(1)} + K_i^{(2)(2)}. \tag{2}$$

Given K_{i-1} and K_{i-2}, the statistical behavior of K_i and K_{i+1} can be completely determined. Therefore, we can choose (K_i, K_{i+1}) as a state vector and solve the two-dimensional Markov chain. To begin, we can express the transition probability as in (3), shown at the bottom of the page, by noting that K_{i+1} depends on K_{i-2} only through K_i and K_{i-1}.

The duration of transmission cycle i consists of a reservation period of length a and the message transmission period (Fig. 1). Therefore, given that $K_i = m$

$$Y_i|_{K_i=m} = a + U_1 + U_2 + \cdots + U_m. \tag{4}$$

Let $Y_i^*(z)$ and $U^*(z)$ be the z-transform of Y_i and U, respectively. Then

$$Y_i^*(z)|_{K_i=m} = z^a[U^*(z)]^m.$$

Taking the inverse z-transform, we obtain the conditional distribution of Y_i as

$$f(n|m) \equiv P[Y_i = n|K_i = m] = Z^{-1}[z^a(U^*(z))^m] \tag{5}$$

where $Z^{-1}[\cdot]$ denotes the inverse z-transform operation.

The $K_{i+1}^{(1)}$ class 1 messages in cycle $i+1$ are generated in cycle i. Therefore, by conditioning on $Y_i = n$, we have

$$\begin{aligned} P\Big[K_{i+1}^{(1)} = l|K_i = m\Big] \\ = \sum_{n=a}^{\infty} P\Big[K_{i+1}^{(1)} = l|Y_i = n, K_i = m\Big] P[Y_i = n|K_i = m]. \end{aligned} \tag{6}$$

But given $Y_i = n, K_{i+1}^{(1)}$ has a Poisson distribution with rate λ_1. Therefore

$$P\Big[K_{i+1}^{(1)} = l|K_i = m\Big] = \sum_{n=a}^{\infty} \frac{e^{-n\lambda_1}(n\lambda_1)^l}{l!} f(n|m). \tag{7}$$

Next, consider the class 2. We can argue similarly that the number of arrivals in cycle $i-1$ depends only on the duration of cycle $i-1$. Therefore

$$P\Big[K_{i+1}^{(2)(1)} = l|K_i = m, K_{i-1} = i\Big] = P\Big[K_{i+1}^{(2)(1)} = l|K_i = m\Big]. \tag{8}$$

Similarly, we can use this condition on $Y_i = n$ and obtain

$$P\Big[K_{i+1}^{(2)(1)} = l|K_i = m\Big] = \sum_{n=a}^{\infty} \frac{e^{-n\lambda_2/2}(n\lambda_2/2)^l}{l!} f(n|m). \tag{9}$$

The conditional distribution of those that arrive during cycle $i-2$, $P[K_{i+1}^{(2)(2)} = l|K_{i-1} = j]$, is given by a similar expression. Therefore, from (2), we can obtain

$$\begin{aligned} P[K_{i+1} = l|K_i = m, K_{i-1} = j] \\ = P\Big[K_{i+1}^{(1)} = l|K_i = m\Big] * P\Big[K_{i+1}^{(2)(1)} = l|K_i = m\Big] \\ * P\Big[K_{i+1}^{(2)(2)} = l|K_{i-1} = j\Big] \end{aligned} \tag{10}$$

where $*$ denotes the convolution operation. This is the first term in (3). The second term is similar with index m replaced with index j.

$$P[K_{i+1} = l, K_i = m|K_{i-1} = j, K_{i-2} = k] = P[K_{i+1} = l\,|K_i = m, K_{i-1} = j]P\,[K_i = m|K_{i-1} = j, K_{i-2} = k] \tag{3}$$

Having derived the transition probabilities, we can use any mathematical package to solve numerically the set of steady-state probabilities $\{P[K_{i+1} = l, K_i = m]\}$ and from it the distribution of Y as

$$p[Y = n] = \sum_{m=0}^{\infty} f(n|m)P[K = m]. \tag{11}$$

With that, $\overline{Y}$ and $\overline{Y}^2$ required in (1) can be obtained.

Recall that the $K^{(2)}$ class 2 messages belonging to a specific group are generated in the previous cycles of length $Y_s = Y_i + Y_{i+1}$. Its distribution is given by

$$\begin{aligned} P[Y_s &= Y_i + Y_{i+1} = n] \\ &= \sum_{l=0}^{\infty} \sum_{m=0}^{\infty} P[Y_i + Y_{i+1} = n | K_{i+1} = l, K_i = m] \\ &\quad \cdot P[K_{i+1} = l, K_i = m] \\ &= \sum_{i=0}^{\infty} \sum_{m=}^{\infty} Z^{-1}\left[z^{2a}(U^*(z))^{l+m}\right] P[K_{i+1} = l, K_i = m]. \end{aligned} \tag{12}$$

The distribution of $K^{(1)}$ and $K^{(2)}$ are needed in the next section. They can be obtained from (7) and (9) shown as (13) and (14), at the bottom of the page.

C. Queuing Delay

Consider the transmission of $K^{(1)} = k$ class 1 messages. The total queuing delay of these $K^{(1)}$ messages is $w_{(1)} + w_{(2)} + \cdots + w_{(K^{(1)})}$, where $w_{(i)}$ is the waiting time of the ith transmitted message. The average waiting time of these k messages is

$$E[D_3] = E\left[\frac{w_{(1)} + w_{(2)} + \cdots w_{(k)}}{k}\right]. \tag{15}$$

Let $\{u_1, u_2, \cdots, u_k\}$ be the set of ordered class 1 message lengths, i.e., $u_{(1)} \leq u_{(2)} \leq \cdots \leq u_{(k)}$. As stated in Section III, shortest messages are transmitted first. Therefore, $w_{(1)} = 0$ and

$$w_{(j)} = u_{(1)} + u_{(2)} + \cdots + u_{(j-1)}, \qquad j = 2, 3, \cdots, k. \tag{16}$$

Substituting into (15) and removing the conditioning on K_1, we obtain

$$E[D_3] = \sum_{k=2}^{\infty} \frac{1}{k} \left\{ \sum_{i=1}^{k-1} (k-i)\overline{u}_{(i)} \right\} P\left[K^{(1)} = k\right] \tag{17}$$

where the distribution of the ordered message length $u_{(j)}$ from a total of k messages obtained from [9] is

$$\begin{aligned} P[u_{(j)} = m] = &\frac{k! P[U = m]}{(k-j)!(j-1)!} \{P[U \leq m]\}^{j-1}, \\ &\cdot \{p[U > m]\}^{k-j}, \qquad j = 1, 2, \cdots, k \end{aligned}$$

and for geometrically distributed messages

$$\begin{aligned} \overline{u_{(j)}} = &\frac{k!}{(k-j)!(j-1)!} \sum_{m=0}^{\infty} \frac{m}{\overline{U}-1} \left[1 - \left(1 - \frac{1}{\overline{U}}\right)^m\right]^{j-1} \\ &\cdot \left(1 - \frac{1}{\overline{U}}\right)^{m(k-j+1)}, \qquad j = 1, 2, \cdots, k. \end{aligned} \tag{18}$$

D. Message Delay for Class 2 Terminals

The message delay D_1' for class 2 terminals can be obtained in a similar way, except that D_1' is given by (1) with Y_S replacing Y and the evaluation of D_3' should consider the lower transmission priority of the class 2 messages. In other words

$$E[D_3'] = \overline{K}_1\overline{U} + \sum_{k=2}^{\infty} \frac{1}{k} \left\{ \sum_{j=1}^{k-1} (k-j)\overline{v}_{(j+1)} \right\} P\left[K^{(2)} = k\right] \tag{19}$$

where $\overline{v}_{(j+1)}$ is given by (18) with $\overline{U}$ replaced by $\overline{V}$.

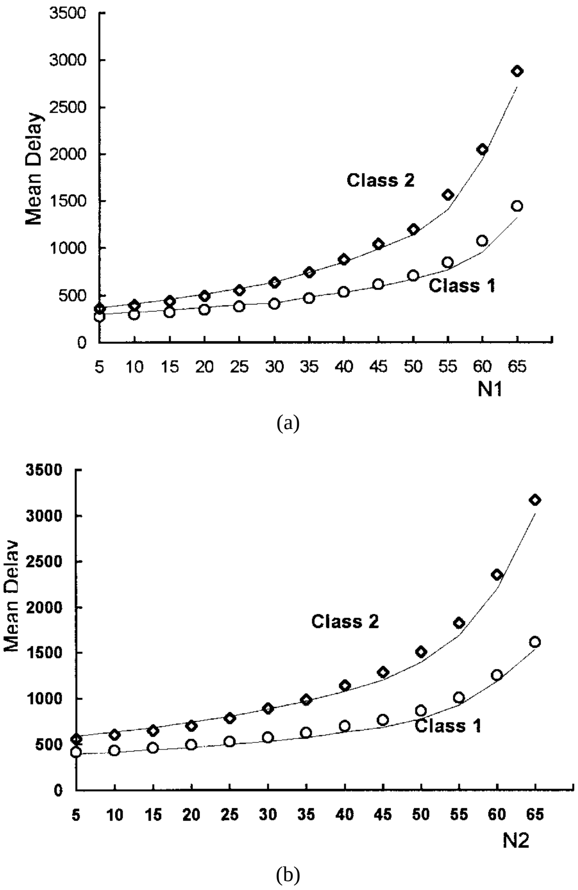
V. Numerical Examples

Let the message length be geometrically distributed, with a mean value of 128 slots, and let each logged terminal generate messages according to a Poisson process with rate $\lambda_0 = 1/20\,480$. This gives a per terminal channel utilization factor $\rho_0 = 1/160$, so that a theoretical maximum of 160 terminals can be supported in the reserved carrier. On a single wireless carrier, we do not expect N to be very large.

Fig. 2(a) shows the average message delay for class 1 and class 2 terminals in slots versus N_1 with $N_2 = 80$. Two class 2 groups are considered. Simulation results are shown

$$P\left[K_{i+1}^{(1)} = l\right] = \sum_{m=0}^{\infty} P\left[K_{i+1}^{(1)} = l | K_i = m\right] P[K_i = m] = \sum_{m=0}^{\infty} \left[\sum_{n=a}^{\infty} \frac{e^{-n\lambda_1}(n\lambda_1)^l}{l!} f(n|m) \right] P[K_i = m] \tag{13}$$

$$\begin{aligned} P\left[K_{i+1}^{(2)} = l\right] &= P\left[K_{i+1}^{(2)(1)} + K_{i+1}^{(2)(2)} = l\right] = \sum_{m=0}^{\infty} \sum_{j=0}^{\infty} \left[\sum_{x=0}^{l} P\left[K_{i+1}^{(2)(1)} = l - x | K_i = m\right] P\left[K_{i+1}^{(2)(2)} = x | K_{i-1} = j\right] \right] \\ &\quad \cdot P[K_i = m, K_{i-1} = j] \end{aligned} \tag{14}$$



(a)

(b)

Fig. 2. (a) Message delay versus the number of (a) class 1 terminals and (b) class 2 terminals.

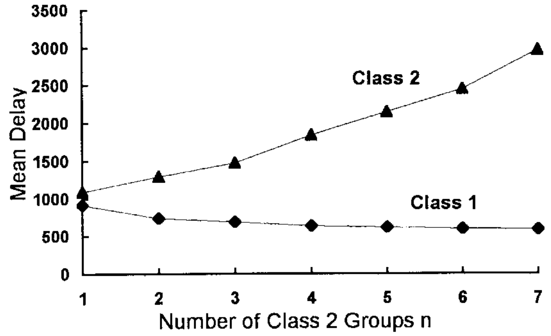


Fig. 3. Message delay versus number of class 2 groups.

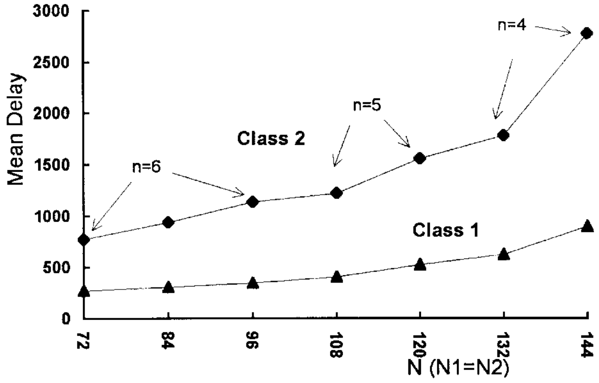


Fig. 4. Message delay tradeoff between two classes.

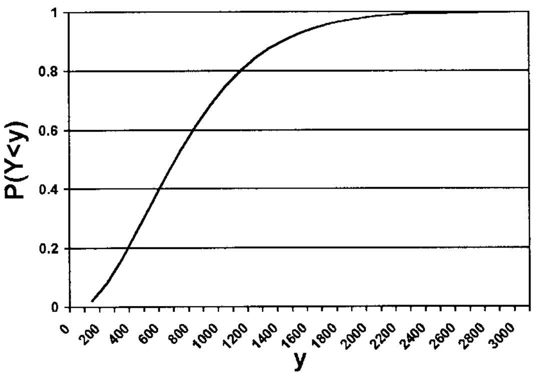


Fig. 5. A typical distribution function of the cycle length.

by markers with the 95% confidence intervals all smaller than the marker size. We find that when $N_1 = 60$ and $N_2 = 80$, which represents a channel utilization of 0.875, the message delays for class 1 and class 2 terminals are 828.9 and 1547.1 slots, respectively. For GSM physical layer with slot size 4.615 ms, the corresponding message delays are 478.2 and 892.5 ms, respectively. We also find that the analytical and the simulation results match very well. Fig. 2(b) shows the message delay versus N_2 with $N_1 = 80$. The results are similar, except that the message delays are slightly larger. This is due to the increase of reservation overhead when N_1 is larger.

Fig. 3 gives the message delay versus n, the number of class 2 groups for $N_1 = 60, N_2 = 75$. It shows that the message delays of class 1 and class 2 terminals can be traded off through changing n. These are simulation results as analysis is limited to $n = 2$.

Fig. 4 shows the simulation result of message delays versus the number of total logged terminals N when a delay requirement $\overline{D_1} = \overline{D_2}/3$ is imposed. The number of class 2 groups n required to achieve this delay difference is also shown. This figure shows that the message delays of class 1 and class 2 terminals can be easily traded off by changing n.

Fig. 5 shows a typical distribution function of the transmission cycle length Y under the same condition as that for Fig. 2(a) with $N_1 = 60$ and $N_2 = 80$. We see that Y starts at 100 slots $(N_1 + N_2/2)$ and can be as large as 2800 slots (or about 21 messages in a cycle). By simple scaling, smaller message size leads to higher reservation overhead and lower throughput. In the computation of the distributions of Y, K_1 and K_2, we need to truncate infinite series. For Y, we extend its range until the relative difference between $\overline{Y}$ obtained by the transform method and that from the distribution is less than 0.2%. For the case in Fig. 5, Y is truncated to 5000. For K_1 and K_2, we use a relative error of 10^{-7} as the stopping rule. For the same case, $P[K_1 = 20] = 0.000\,000\,2$ and $P[K_2 = 20]$ is even smaller, and so they are both truncated to 20.

VI. Conclusion

The protocol studied has promising features, such as contention-free, traffic variation tracking to achieve efficient channel utilization, and supporting of multipriority services. It can be adapted to TDMA-based systems, such as GSM and DECT, without modifying the physical layer. A framing structure can also be introduced for ease of synchronization. In that case, transmission cycles are in units of frames, and the first slot of every frame can be reserved for the cycle_start signal. This allows very easy detection of cycle boundaries by newly activated terminals. Also, by setting a limit to the transmission cycle length, high-priority terminals can have bounded delay on their messages.

References

[1] G. Bianchi *et al.*, "Packet data service over GSM networks with dynamic stealing of voice channels," in *Proc. IEEE GLOBECOM*, 1995, pp. 1152–1156.

[2] J. Hamalainen *et al.*, "Multi-slot packet radio air interface to TDMA systems–variable rate reservation access (VRRA)," in *Proc. IEEE PIMRC*, Sept. 1995, pp. 366–371.

[3] A. C. Cleary and M. Paterakis, "Design and performance evaluation of a scheme for voice-data channel access in third generation microcellular wireless networks," in *Proc. ICUPC*, Cambridge, MA, 1996, pp. 1–6.

[4] F. Khan and D. Zeghlache, "Priority-based multiple access (PBMA) for statistical multiplexing of multiple services in wireless PCS," in *Proc. ICUPC*, Cambridge, MA, 1996, pp. 17–21.

[5] D. Turina *et al.*, "A protocol for multi-slot MAC layer operation for packet data channel in GSM," in *Proc. ICUPC*, Cambridge, MA, 1996, pp. 572–576.

[6] S. Noerskov, U. Gliese, and K. Stubkjaer, "Adaptive packet reservation multiple access (A-PRMA) for broadband wireless ATM," presented at the MoMuC-3 Workshop, Princeton, NJ, USA, Sept. 1996.

[7] F. Watanabe *et al.*, "Performance evaluation of reserved idle signal multiple access with collision resolution," presented at the MoMuC-3 Workshop, Princeton, NJ, USA, Sept. 1996.

[8] L. Kleinrock, *Queueing Systems—Volume I: Theory*. New York: Wiley, 1975.

[9] S. Ross, *A First Course in Probability*, 4th ed. Englewood Cliffs, NJ: Prentice-Hall, 1994.