Delay Distributions of Slotted ALOHA and CSMA

Yang Yang, *Member, IEEE,* and Tak-Shing Peter Yum, *Senior Member, IEEE*

***Abstract*—In this paper, we derive the closed-form delay distributions of slotted ALOHA and nonpersistent carrier sense multiple access (CSMA) protocols under steady state. Three retransmission policies are analyzed. We find that under a *binary exponential backoff* retransmission policy, finite average delay and finite delay variance can be guaranteed for $G < 2S$ and $G < 4S/3$, respectively, where G is the channel traffic and S is the channel throughput. As an example, in slotted ALOHA, $S < \ln 2/2$ and $S < 3(\ln 4 - \ln 3)/4$ are the operating ranges for finite first and second delay moments. In addition, the blocking probability and delay performance as a function of $r_{\max}$ (maximum number of retransmissions allowed) is also derived.**

***Index Terms*—ALOHA, carrier sense multiple access (CSMA), random access protocol.**

I. Introduction

RANDOM ACCESS protocols, such as ALOHA and carrier sense multiple access (CSMA), are widely used in wireless communication systems such as packet satellite communications, wireless LAN, and the random access channel in cellular mobile systems. During the past three decades, ALOHA- and CSMA-type protocols have been extensively studied with stationary throughput and delay characteristics being derived for slotted and unslotted channels, and finite and infinite population models [1]–[4]. Typically, the system average backlog is derived and the expected delay is obtained by using Little's formula.

Analytical results on delay distributions of the slotted ALOHA and CSMA protocols are obtained only for systems with a finite population [5], [6]. Specifically, in [5], Tobagi derived the z transform and moments of both the waiting time and interdeparture time distribution in slotted ALOHA and CSMA with collision detection (/CD) protocols using a discrete-time Markov chain. In [6], the matrix-geometric method is used to derive the delay distribution of CSMA/CD on a continuous-time Markov chain model. Both approaches are analytically complicated and become intractable when the population size is large.

In this paper, a simple closed-form expression of the delay distribution is derived for slotted ALOHA and CSMA protocols without using transform. Three retransmission policies are analyzed and the conditions for achieving finite delay mean and variance are derived under the *binary exponential backoff* policy. The exact way whereby $r_{\max}$ (maximum number of retransmissions allowed) can be used to tradeoff the blocking probability and delay performance is also given. The analytical results of delay performance are verified by computer simulation.

Paper approved by Y. Fang, the Editor for Wireless Networks of the IEEE Communications Society. Manuscript received January 17, 2002; revised October 3, 2002; January 16, 2003; and April 17, 2003. This work was supported in part by the Hong Kong Research Grants Council under Grant CUHK 4223/00E and Grant CUHK 4325/02E.

Y. Yang was with the Department of Information Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong. He is now with the Department of Electronic and Computer Engineering, Brunel University, Uxbridge UB8 3PH, U.K. (e-mail: yang.yang@brunel.ac.uk).

T.-S. P. Yum is with the Department of Information Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong (e-mail: yum@ie.cuhk.edu.hk).

Digital Object Identifier 10.1109/TCOMM.2003.819201

II. System Model

The system model and the notations follow that in [1]. To summarize, we have the following.

1) Packets are of the same size with transmission time T. The maximum end-to-end propagation delay is denoted by τ with normalized value $a = \tau/T$. The maximum round-trip delay is smaller than the packet transmission time, i.e., $2\tau < T$.
2) The combination of new and retransmitted packet arrivals is a Poisson process with rate G (packets/T), which is referred to as *offered traffic* to the slotted channel. Let S be the corresponding *throughput*. Then, $p_s \triangleq S/G$ is the success probability of a transmission.
3) When a new packet is generated, it accesses the channel at the beginning of next slot. This is called *immediate-first transmission* (IFT). The transmission result is broadcast through a separate reliable acknowledgment channel.

A. Retransmission Policy

When a packet transmission fails, a retransmission is scheduled after a random backoff delay, which is determined by a specific retransmission policy. Let W_i be the ith backoff delay in unit of slots. Then, the ith retransmission takes place at the beginning of the W_ith available slot after knowing the last transmission is unsuccessful. The delay performance of a random access system depends strongly on the distribution of W_i. In this paper, we consider three different retransmission policies.

1) Under a *uniform backoff* (UB) policy, all W_i's are uniformly distributed in the same range, say $[1, \omega]$.
2) Under a *binary exponential backoff* (BEB) policy, backoff delay is uniformly distributed in a binary exponentially expanding range. In other words, the range of backoff delay is doubled every time an unsuccessful retransmission occurs. Let ω be the *initial backoff range*. W_i is then uniformly distributed in $[1, 2^{i-1}\omega]$.
3) Under a *geometric backoff* (GB) policy, backoff delay is geometrically distributed with parameter q.

Table I compares some statistics of W_i under the three retransmission policies.

B. Access Delay

Let R be the number of retransmissions needed and D_i be the delay time due to the ith unsuccessful transmission (or the

TABLE I
STATISTICS OF BACKOFF DELAY W_i UNDER DIFFERENT RETRANSMISSION POLICIES

	UB	BEB	GB
$P\{W_i = k\}$	$\frac{1}{\omega}$	$\frac{1}{2^{i-1}\omega}$	$q(1-q)^{k-1}$
$E[W_i]$	$\frac{1+\omega}{2}$	$\frac{1+2^{i-1}\omega}{2}$	$\frac{1}{q}$
$E[W_i^2]$	$\frac{2\omega^2+3\omega+1}{6}$	$\frac{4^i\omega^2+3\cdot 2^i\omega+2}{12}$	$\frac{2-q}{q^2}$
$Var(W_i)$	$\frac{\omega^2-1}{12}$	$\frac{4^{i-1}\omega^2-1}{12}$	$\frac{1-q}{q^2}$

$(i-1)$th retransmission if $i \geq 2$). Naturally, D_0 is the access delay when the initial transmission is successful. It includes the T seconds of transmission delay and an average of 0.5 slot of slot synchronization delay. The access delay D of a packet is the time duration from its generation to the moment it is successfully transmitted or

$$D = \sum_{i=0}^{R} D_i. \tag{1}$$

Under the Poisson arrival assumption and for large backoff range, e.g., $\omega > 20$, R can be accurately approximated by a geometric distribution with transmission success probability p_s as the parameter [2], [7]

$$P\{R = r\} = p_s(1-p_s)^r. \tag{2}$$

Obviously, different random access protocols have different p_s values. The distributions of $D_1, D_2, \ldots$ are jointly determined by the specific random access protocol and retransmission policy.

III. SLOTTED ALOHA

For slotted ALOHA, the length of a slot is equal to the packet transmission time T. Therefore, D_0 is uniformly distributed in $(T, 2T]$. The success probability p_s was derived in [1] as

$$p_s = \frac{S}{G} = e^{-G}. \tag{3}$$

Fig. 1 shows the access procedure of a tagged packet generated at time t_0 and transmitted at the next slot. The sender waits for a round-trip propagation delay of $2aT$ seconds before receiving the first acknowledgment (*unsuccessful*, in this case). Here, we have implicitly assumed that the sum of packet processing time at the receiver and the round-trip propagation delay is smaller than one slot time. When this is not true, only a fixed constant term needs to be added to the final delay equation. Assume a collision occurs at the ith transmission, the backoff delay caused is W_iT. Add to it the slot due to packet transmission, we have

$$D_i = (W_i + 1)T, \quad i = 1, 2, \ldots. \tag{4}$$

Substitute (4) into (1), we get

$$\begin{aligned} D &= D_0 + T\sum_{i=1}^{R} W_i + RT \\ &= T(X_R + Y_R) \end{aligned} \tag{5}$$

where $X_R = \sum_{i=1}^{R} W_i$ and $Y_R = D_0/T + R$.

A. Delay Distribution

Given $R = r$ (where $r \geq 1$), the distribution of X_r is derived in Appendix A for different retransmission policies as

$$P\{X_r = k\} = \begin{cases} \dfrac{a_r(k)}{\omega^r}, & \text{UB} \\ \dfrac{b_r(k)}{2^{r(r-1)/2}\omega^r}, & \text{BEB} \\ c_r(k)q^r(1-q)^{k-r}, & \text{GB} \end{cases} \tag{6}$$

where $k \geq r$ and ω is the initial backoff range. $a_r(n)$, $b_r(n)$, and $c_r(n)$ are three sequences defined in Appendix A.

Next, Y_r is a random variable uniformly distributed in $(r+1, r+2]$. Let $F_{X_r}(x)$ and $F_{Y_r}(x)$ be the cumulative distribution functions of X_r and Y_r, respectively. As X_r and Y_r are independent, the conditional distribution $F_D(x|R=r)$ for $x > (2r+1)T$ can be computed by convolving $F_{X_r}(x)$ and $(d/dx)F_{Y_r}(x)$. In other words

$$\begin{aligned} F_D(x|R=r) &= \int_{-\infty}^{\infty} F_{X_r}\left(\frac{x}{T} - y\right) \cdot \left(\frac{d}{dy}F_{Y_r}(y)\right) dy \\ &= \int_{r+1}^{r+2} P\left\{X_r \leq \frac{x}{T} - y\right\} dy. \end{aligned} \tag{7}$$

Let $x/T = x_i + x_d$, where $x_i = \lfloor x/T \rfloor$ ($\lfloor x \rfloor$ is the floor function) and $x_d = x/T - x_i$ represent the integer and decimal parts of x/T, respectively. Then, (7) can be simplified to

$$\begin{aligned} F_D(x|R=r) &= \int_{r+1}^{r+1+x_d} P\{X_r \leq x_i + x_d - y\}dy \\ &\quad + \int_{r+1+x_d}^{r+2} P\{X_r \leq x_i + x_d - y\}dy \\ &= \int_{r+1}^{r+1+x_d} P\{X_r \leq x_i - r - 1\}dy \\ &\quad + \int_{r+1+x_d}^{r+2} P\{X_r \leq x_i - r - 2\}dy \\ &= x_d \cdot \sum_{j=r}^{x_i-r-1} P\{X_r = j\} + (1 - x_d) \\ &\quad \cdot \sum_{k=r}^{x_i-r-2} P\{X_r = k\} \\ &= x_d P\{X_r = x_i - r - 1\} \\ &\quad + \sum_{k=r}^{x_i-r-2} P\{X_r = k\}, \\ &\quad r \geq 1, \ x > (2r+1)T. \end{aligned} \tag{8}$$

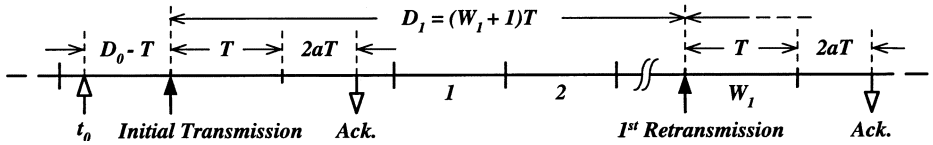

Fig. 1. Access mechanism of slotted ALOHA.

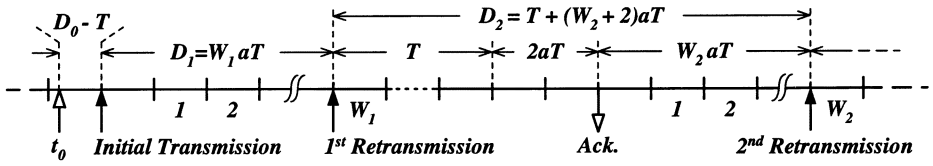

Fig. 2. Access mechanism of slotted nonpersistent CSMA.

Removing the conditioning on R, we have

$$\begin{aligned} F_D(x) &= P\{D_0 \le x\} \cdot P\{R=0\} \\ &\quad + \sum_{r=1}^{\infty} F_D(x|R=r)P\{R=r\} \\ &= P\{D_0 \le x\} \cdot P\{R=0\} \\ &\quad + x_d \sum_{r=1}^{r_1} P\{X_r = x_i - r - 1\}P\{R=r\} \\ &\quad + \sum_{r=1}^{r_1} \sum_{k=r}^{x_i-r-2} P\{X_r = k\}P\{R=r\}, \quad x > T \end{aligned} \tag{9}$$

where $r_1 = \lfloor (x/T - 1)/2 \rfloor$ as each retransmission takes up at least two slots.

B. Moments of Delay

Under UB and GB policies, W_i's are independent and identically distributed (i.i.d.) random variables. Therefore, the mean delay can be derived as

$$\begin{aligned} \overline{D} &= E[D] \\ &= E[D_0] + E[R]E[W_i]T + E[R]T \\ &= \begin{cases} \frac{T}{2}\left[\frac{3+\omega}{p_s} - \omega\right], & \text{UB} \\ \frac{T}{2q}\left[\frac{2+2q}{p_s} + q - 2\right], & \text{GB} \end{cases} \end{aligned} \tag{10}$$

and the delay variance is

$$\begin{aligned} \sigma_D^2 &= \mathrm{Var}(D) \\ &= \mathrm{Var}(D_0) + T^2\mathrm{Var}\left(\sum_{i=1}^{R}(W_i+1)\right) \\ &= \frac{T^2}{12} + T^2\left\{E[R]\cdot\mathrm{Var}(W_i+1) + (E[W_i+1])^2\cdot\mathrm{Var}(R)\right\} \\ &= \begin{cases} \frac{T^2}{12}\left[\frac{3(3+\omega)^2}{p_s^2} - \frac{2(\omega+2)(\omega+7)}{p_s} + 2 - \omega^2\right], & \text{UB} \\ \frac{T^2}{q^2}\left[\frac{(q+1)^2}{p_s^2} - \frac{(q^2+3q)}{p_s} + \frac{q^2}{12} + q - 1\right], & \text{GB.} \end{cases} \end{aligned} \tag{11}$$

For $p_s > 0$, $\overline{D}$ and σ_D^2 are finite.

For BEB policy, W_i's are no longer identically distributed. $\overline{D}$ and σ_D^2 are derived by conditioning on $R = r$. Specifically, the conditional expected delay is given by

$$\begin{aligned} E[D|R=r] &= E[D_0] + T\sum_{i=1}^{r} E[W_i] + rT \\ &= \frac{T}{2}\left(\omega 2^r + 3r + 3 - \omega\right), \quad \text{BEB.} \end{aligned} \tag{12}$$

Removing the conditioning on R, we obtain

$$\begin{aligned} \overline{D} &= \sum_{r=0}^{\infty} E[D|R=r]P\{R=r\} \\ &= \frac{T}{2}\left(\frac{3}{p_s} + \frac{\omega p_s}{1-2(1-p_s)} - \omega\right), \quad p_s > 0.5, \ \text{BEB.} \end{aligned} \tag{13}$$

Note that in (13), the condition for finite expected delay $\overline{D}$ is that $2(1-p_s) < 1$, or $p_s > 0.5$. This is necessary for the infinite summation over r to converge.

The conditional second moment of access delay under BEB is

$$\begin{aligned} E\left[D^2|R=r\right] &= E\left[(D_0 + D_1 + \cdots + D_r)^2\right] \\ &= T^2\left[\frac{5\omega^2}{18}4^r + \frac{3\omega}{2}r2^r - \frac{\omega^2-3\omega}{2}2^r + \frac{9}{4}r^2 \right. \\ &\quad \left. - \left(\frac{3\omega}{2} - \frac{53}{12}\right)r + \frac{2\omega^2}{9} - \frac{3\omega}{2} + \frac{7}{3}\right]. \end{aligned} \tag{14}$$

$E[D^2]$ can then be derived by removing the conditioning on R. But here, the infinite summation over r converges only for $p_s > 0.75$. This is also the condition for finite second moment under BEB policy. Finally, delay variance σ_D^2 is given by

$$\begin{aligned} \sigma_D^2 &= E\left[D^2\right] - (E[D])^2 \\ &= \frac{T^2}{36}\left[\frac{10\omega^2 p_s}{1-4(1-p_s)} + \frac{(54-9\omega p_s)\omega p_s}{[1-2(1-p_s)]^2} \right. \\ &\quad \left. - \frac{54\omega}{1-2(1-p_s)} + \frac{81}{p_s^2} - \frac{84}{p_s} - \omega^2 + 6\right], \\ &\quad p_s > 0.75, \ \text{BEB.} \end{aligned} \tag{15}$$

As a check, (10) is identical to that in [1]. However, (11), (13), and (15) are not found in the literature.

IV. Slotted Nonpersistent CSMA

In slotted nonpersistent CSMA, the length of a slot is defined to be equal to the maximum propagation delay $\tau = aT$. Hence, D_0 is uniformly distributed in $(T, T+aT]$. The success probability p_s was derived in [1] as

$$p_s = \frac{S}{G} = \frac{ae^{-aG}}{1+a-e^{-aG}}. \tag{16}$$

Fig. 2 shows the access procedure of a tagged packet generated at time t_0. At its initial transmission, assume the channel is sensed *busy* so that the transmission is unsuccessful. After the first

backoff delay W_1 (in unit of slots), the tagged packet senses the channel for the second time. If it is *idle*, the packet is transmitted and the sender waits for a round-trip propagation delay of $2aT$ s to learn the transmission result. In case a collision occurs, the tagged packet will try to access the channel again W_2 slots after receiving the acknowledgment. Therefore, the delay time D_1 and D_2 due to the first two unsuccessful transmissions are simply

$$D_1 = W_1 \cdot aT \tag{17}$$

and

$$D_2 = T + (W_2 + 2)aT. \tag{18}$$

Let K $(0 \le K \le R)$ be the number of unsuccessful transmissions due to busy channel. The access delay D in this case is

$$\begin{aligned} D &= D_0 + \sum_{i=1}^{K} W_i aT + \sum_{j=K+1}^{R} [T + (W_j + 2)aT] \\ &= D_0 + aT \sum_{i=1}^{R} W_i + (R-K)(1+2a)T \\ &= aT(X_R + Z_{R,K}) \end{aligned} \tag{19}$$

where $X_R = \sum_{i=1}^{R} W_i$ is the same as in (5) and $Z_{R,K}$ is defined as $Z_{R,K} = ((R-K)(1+2a)T + D_0)/(aT)$.

A. Delay Distribution

Given $R = r(\ge 1)$ and $K = k$, the conditional cumulative distribution function $F_D(x|R = r, K = k)$ for $x > [1 + (1+3a)r - (1+2a)k]T$ can be derived in a similar way as in Section III-A.

$$\begin{aligned} &F_D(x|R=r,K=k) \\ &\quad = x_d P\{X_r = x_i - \eta\} + \sum_{m=r}^{x_i-\eta-1} P\{X_r = m\} \end{aligned} \tag{20}$$

where $x_i = \lfloor x/(aT) \rfloor$ and $x_d = x/(aT) - x_i$ are the integer and decimal parts of $x/(aT)$. η is defined as $\eta = ((r-k)(1+2a)+1)/a$.

The joint distribution of R and K is derived in Appendix B as

$$P\{R = r, K = k\} = \binom{r}{k} p_b^k p_c^{r-k} p_s, \quad 0 \le k \le r \tag{21}$$

where p_b is the probability that the channel is sensed busy and p_c is the probability that a collision occurs. They are given in Appendix B as (B.4) and (B.5), respectively.

Removing the conditioning on R and K, we obtain the distribution of D as

$$\begin{aligned} F_D(x) &= P\{D_0 \le x\} \cdot P\{R=0\} \\ &\quad + \sum_{r=1}^{\infty} \sum_{k=0}^{r} F_D(x|R=r,K=k) P\{R=r,K=k\} \\ &= P\{D_0 \le x\} \cdot P\{R=0\} + x_d \\ &\quad \cdot \sum_{r=1}^{r_2} \sum_{k=0}^{r} P\{X_r = x_i - \eta\} P\{R=r,K=k\} \\ &\quad + \sum_{r=1}^{r_2} \sum_{k=0}^{r} \sum_{m=r}^{x_i-\eta-1} P\{X_r = m\} P\{R=r,K=k\}, \\ &\quad x > T \end{aligned} \tag{22}$$

where $r_2 = \lfloor x/(aT) - 1/a \rfloor$ as x should be larger than $(1 + ar)T$ at $k = r$.

B. Moments of Delay

As in slotted ALOHA, the expected delay $\overline{D}$ under UB and GB policies can easily derived as

$$\begin{aligned} \overline{D} &= E[D_0] + aTE[R]E[W_i] + (1+2a)TE[R-K] \\ &= \frac{(2+a)T}{2} + \frac{(1-p_s)aT}{p_s} \\ &\quad \cdot E[W_i] + \frac{(1-p_s-p_b)(1+2a)T}{p_s} \\ &= \begin{cases} \frac{T}{2}\left[\frac{a\omega+5a+2}{p_s} - \frac{2(1+2a)p_b}{p_s} - a(4+\omega)\right], & \text{UB} \\ \frac{T}{q}\left[\frac{a+q+2aq}{p_s} - \frac{q(1+2a)p_b}{p_s} - \frac{a(3q+2)}{2}\right], & \text{GB.} \end{cases} \end{aligned} \tag{23}$$

The derivation of delay variance σ_D^2 is straightforward, and so is omitted here.

For BEB, the expected delay conditioned on $R = r$ and $K = k$ is

$$\begin{aligned} E[D|R=r,K=k] &= E[D_0] + aT \sum_{i=1}^{r} E[W_i] \\ &\quad + (r-k)(1+2a)T \\ &= \frac{T}{2}\left[a\omega 2^r + (2+5a)r - (2+4a)k \right. \\ &\quad \left. + a + 2 - a\omega\right], \quad \text{BEB.} \end{aligned} \tag{24}$$

The average access delay $\overline{D}$ is, therefore, given by

$$\begin{aligned} \overline{D} &= \sum_{r=0}^{\infty} \sum_{k=0}^{r} E[D|R=r,K=k] P\{R=r,K=k\} \\ &= \frac{T}{2}\left[\frac{a\omega p_s}{1-2(1-p_s)} + \frac{2+5a}{p_s} \right. \\ &\quad \left. - \frac{(2+4a)p_b}{p_s} - a(4+\omega)\right], \quad p_s > 0.5, \ \text{BEB.} \end{aligned} \tag{25}$$

Again, $p_s > 0.5$ is required for finite average delay. Delay variance σ_D^2 can be derived in the same way as that in Section III-B and $p_s > 0.75$ is again found to be the condition for finite variance.

V. Stability Conditions for BEB

In the delay equations for slotted ALOHA and nonpersistent CSMA, (5) and (19), there is a common term $X_R = \sum_{r=1}^{R} W_i$. Under the BEB retransmission policy, W_i's are independent but not identical random variables. As a result, the ith conditional moment of delay $E[D^i|R = r]$ has a sequence of factors of $2^r, 2^{2r}, \ldots, 2^{ir}$. Also, different random access protocols differ only by the parameter $p_s = S/G$, such as (3) for slotted ALOHA and (16) for slotted nonpersistent CSMA, in the delay distribution. Now, removing the conditioning on R (geometrically distributed) requires summation of terms $[2(1 -$

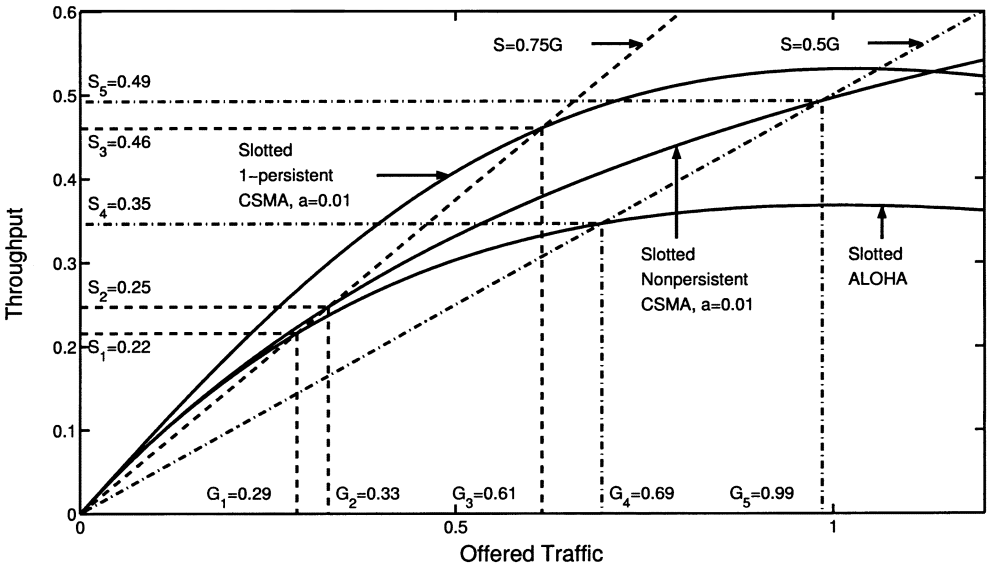

Fig. 3. Conditions for finite expected delay and finite delay variance under the BEB policy.

$p_s)]^r, [4(1-p_s)]^r, \ldots$, and $[2^i(1-p_s)]^r$ over an infinite geometric series. The convergent conditions for all these sums are

$$\begin{cases} 2(1-p_s) < 1 \\ 4(1-p_s) < 1 \\ \vdots \\ 2^i(1-p_s) < 1. \end{cases} \tag{26}$$

Taking the tightest bound, the condition for the finite ith moment of access delay is simply $p_s > 1 - 2^{-i}$. This translates to $p_s > 0.5$ and $p_s > 0.75$ for the first two moments. In Fig. 3, the throughput curves of slotted ALOHA, 1-persistent, and nonpersistent CSMA protocols are plotted against offered traffic G. The two straight lines, $S = 0.5G$ and $S = 0.75G$, represent the lower bounds of p_s for finite first and second moments of access delay. The intersections (G_1, S_1), (G_2, S_2), and (G_3, S_3) give the upper limits of operation for guaranteeing finite first two delay moments in different random access protocols. However, intersections (G_4, S_4) and (G_5, S_5) offers the operating upper bounds for finite expected delay only.

As seen in Fig. 3, slotted nonpersistent CSMA has throughput $(S < S_2 = 0.25)$ not much higher than slotted ALOHA $(S < S_1 = 0.22)$ if finite delay variance needs to be guaranteed. The operating range of slotted 1-persistent CSMA is, however, much larger $(S < S_3 = 0.46)$. We can, therefore, conclude that, under BEB, 1-persistent CSMA is superior to nonpersistent CSMA, although the latter can offer a much higher maximum throughput in theory [2]. Further, when $S < S_3$ (or $G < G_3 = 0.61$), the throughput curves of 1-persistent and nonpersistent CSMA/CD are very close to that of the corresponding CSMA protocols [3], [7]. Hence, the same conclusion applies to CSMA/CD. This confirms the correct choice of 1-persistent CSMA/CD over that of nonpersistent CSMA/CD for the IEEE 802.3 standard a long time ago [8].

VI. When R is Limited to $r_{\max}$

Protocols adopted in applications often block a packet after a certain number of unsuccessful retransmissions. If $r_{\max}$ is the maximum number of retransmissions allowed, the blocking probability P_B is defined as

$$P_B = P\{R > r_{\max}\} = (1-p_s)^{r_{\max}+1}. \tag{27}$$

Figs. 4 and 5 show the relationship between P_B and S with $r_{\max}$ as a parameter for slotted ALOHA and nonpersistent CSMA, respectively. For slotted ALOHA, operating near capacity, say $S = 0.35$, $r_{\max} \geq 9$ is needed to guarantee $P_B < 10^{-3}$ and $r_{\max} \geq 13$ is needed for $P_B < 10^{-4}$. For slotted nonpersistent CSMA, things are quite different. In order to have a reasonable blocking probability, say $P_B < 10^{-3}$ and retransmission delay, say $r_{\max} = 9$, the channel throughput has to be limited to $S < 0.5$, significantly lower than the channel capacity $S_{\max} = 0.86$. This shows that for slotted nonpersistent CSMA, the upper portion of the throughput, i.e., for $0.5 < S < 0.86$, is in fact "unfriendly," or the channel has poor quality of service.

For those packets that are successfully transmitted (i.e., not blocked), their retransmission distribution of R' is just the distribution of R conditioned on $R \leq r_{\max}$. In other words

$$\begin{aligned} P\{R' = r\} &= P\{R = r | R \leq r_{\max}\} \\ &= \frac{p_s(1-p_s)^r}{1-(1-p_s)^{r_{\max}+1}}, \quad r = 0, 1, 2, \ldots, r_{\max}. \end{aligned} \tag{28}$$

Replacing R by R' in the previous results, we obtain the corresponding delay distributions and moment equations under blocking condition. By substituting (28) into (13) and (15), it is easy to see that for $r_{\max}$ finite, both $\overline{D}$ and σ_D are finite. The choice of $r_{\max}$ gives a tradeoff between blocking probability and the latency requirement.

VII. Numerical and Simulation Results

The computer simulation results reported in this section are obtained by the following procedure. New packets are generated according to a Poisson process. Each new packet is time stamped at its birth and the sojourn time (in unit of slots) is measured when the packet is successfully transmitted. The delay sta-

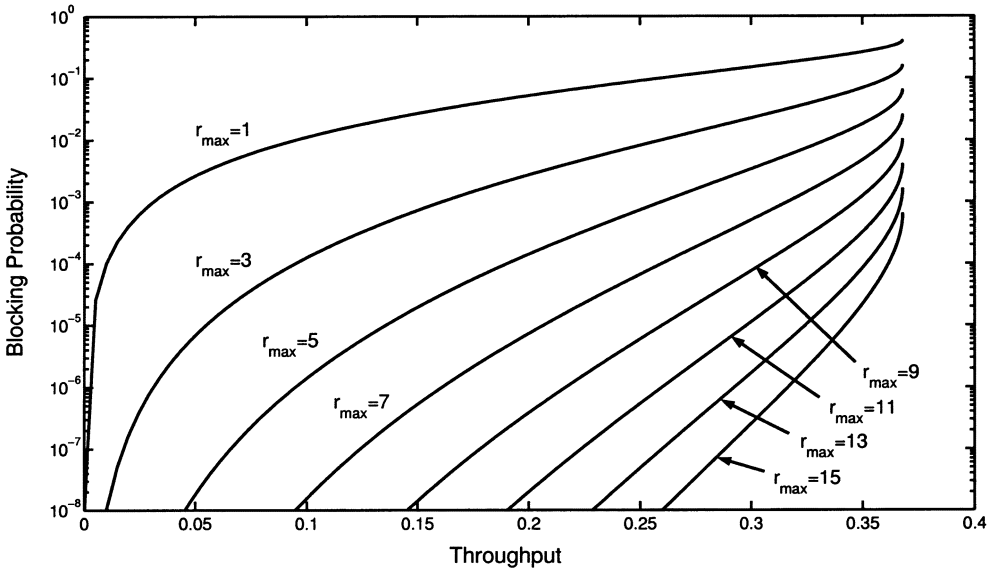

Fig. 4. Blocking probability P_B versus throughput S, slotted ALOHA.

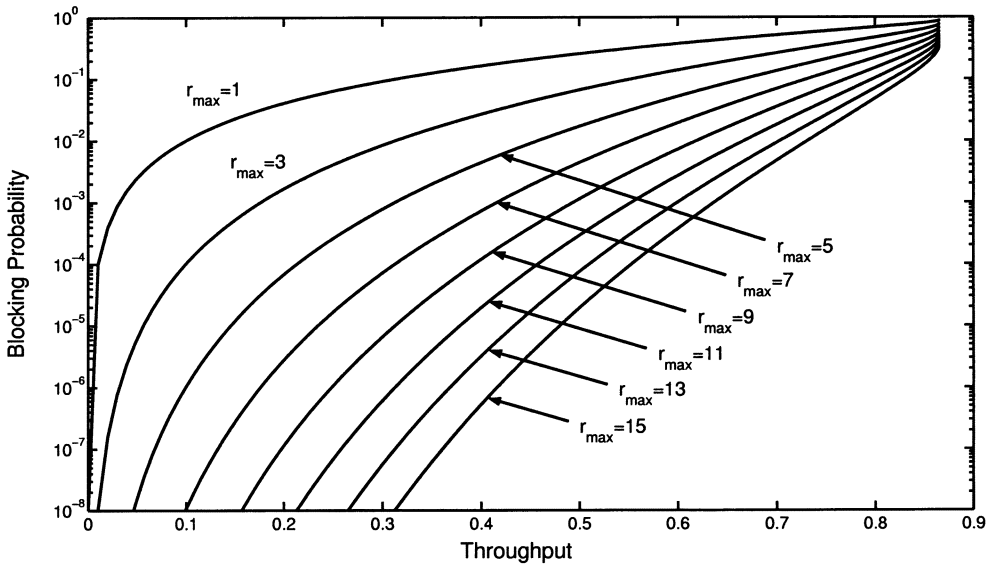

Fig. 5. Blocking probability P_B versus throughput S, slotted nonpersistent CSMA.

tistics are then estimated by processing a large number of such sojourn time values.

For slotted ALOHA with BEB policy, Fig. 6 shows the cumulative distribution function of delay $F_D(x)$ (in unit of slots) for $\omega = 32$, $r_{\max} = 5$, and different p_s values. The analytical results shown in solid lines are obtained by substituting (28) into (9).[1] As seen, they match very well with the simulation results shown in markers. For all those simulation points, the 95% confidence intervals are made to be smaller than the marker size shown.

Since access delay D is no less than D_0 and D_0 is uniformly distributed in $(T, 2T]$, we have $F_D(1) = 0$ and $F_D(2) = P\{R' = 0\} = p_s/(1-(1-p_s)^{r_{\max}+1})$ (the initial transmission is successful), as shown in the figure. If the initial transmission is not successful $(R' \geq 1)$, D is larger than $3T$ as the round-trip propagation delay value(s) should be taken into account (see Fig. 1). Therefore, we have $F_D(3) = F_D(2)$ for all the curves in Fig. 6.

For the delay range (3,35] (note that $35 = 3 + \omega$), the curves appear to be quite straight. This is because the increment of $F_D(x)$ over the above range is dominated (mostly contributed) by the packets with $R' = 1$, especially when the delay value x is close to three. In other words, the probability that a packet with access delay $D \in (3T, 35T]$ is retransmitted two or more times is quite small compared with the probability that it is retransmitted just once. When x is slightly larger than three, the probability $P\{R' \geq 2\}$ is negligible. Hence, the curve slopes are given by

$$h_1 = \frac{1}{\omega} P\{R' = 1\} = \frac{p_s(1-p_s)}{32[1-(1-p_s)^6]}. \tag{29}$$

As x becomes larger and larger, the probability $P\{R' \geq 2\}$ increases so that the curves become, actually, steeper and steeper until the point $x = 35$. In addition, it is seen in the figure that a smaller p_s value gives a steeper curve and a larger curve-slope

[1]Note that the summation over r in (9) is now upper bounded by $\min\{r_{\max}, r_1\}$. Similarly, for slotted nonpersistent CSMA with finite $r_{\max}$ values, the summation over r in (22) is upper bounded by $\min\{r_{\max}, r_2\}$.

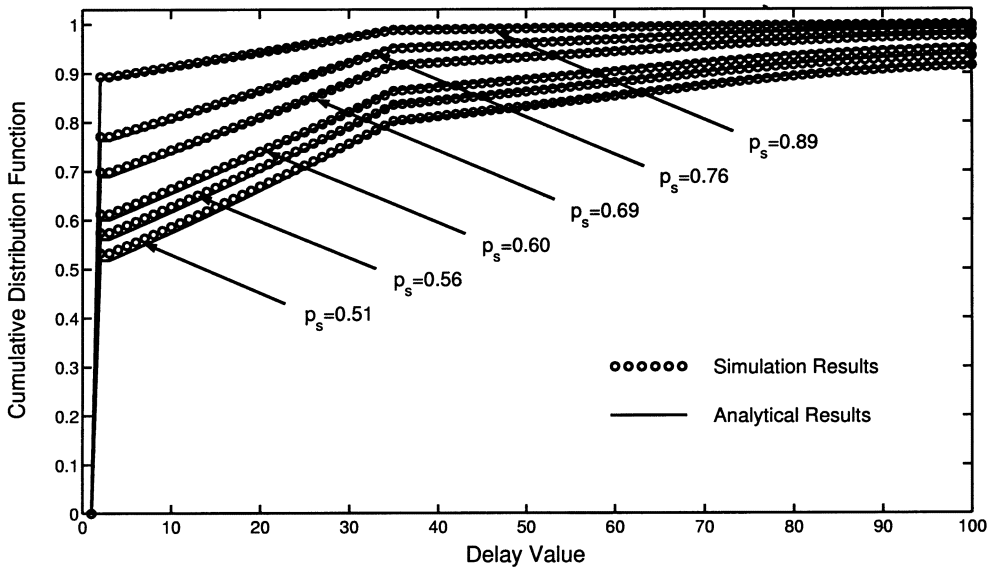

Fig. 6. Cumulative distribution function of delay $F_D(x)$, slotted ALOHA, BEB policy, $\omega = 32$, and $r_{\max} = 5$.

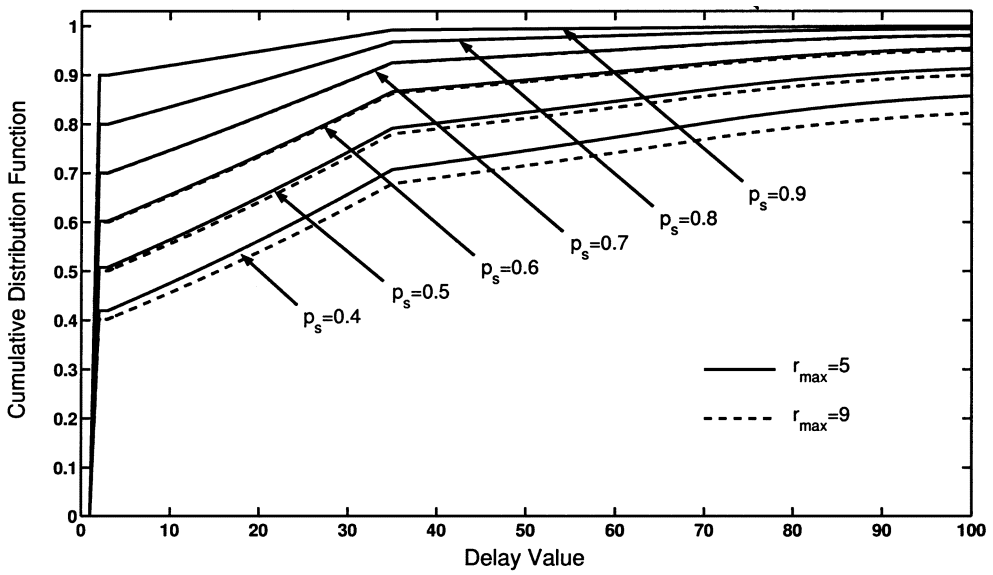

Fig. 7. Cumulative distribution function of delay $F_D(x)$, slotted ALOHA, BEB policy, and $\omega = 32$.

increment (from $x = 3$ to $x = 35$). This is expected, as a smaller p_s implies a larger number of retransmissions, and therefore, higher probability $P\{R' \geq 2\}$.

Consider a packet with access delay larger than $35T$. We know for sure that it has been retransmitted at least twice because the initial backoff range is $\omega = 32$. Hence, the increment of $F_D(x)$ for the delay range $x > 35$ is contributed purely by the packets with $R' \geq 2$. This is quite different from the situation in the range (3,35] and results in a great drop of increment rate (curve slope) of $F_D(x)$. Specifically, when p_s is large and x is close to 35, the slopes can be approximated by[2]

$$h_2 = \frac{1}{2\omega} P\{R' = 2\} = \frac{p_s(1-p_s)^2}{64[1-(1-p_s)^6]}. \qquad (30)$$

[2]This equation accounts for the contributions by the packets with $R' = 2$ only. It underestimates the real curve slopes shown in the figure, especially when p_s is small and x is much larger than 35. For this case, the contributions by the packets with $R' \geq 3$ could not be ignored.

Comparing to the curve slope h_1 given by (29) for the delay range (3,35], this slope is much smaller, leaving $x = 35$ as the common point where all the curve knees are located. To obtain a specific delay value x larger than 35, there are many more possible combinations of R' $tex(R' \geq 2)$ and W_i values. Therefore, no particular R' value could dominate the increment of $F_D(x)$ and no curve knee appears in the range $x > 35$.

The effect of $r_{\max}$ value on the cumulative distribution function of delay is shown in Fig. 7 for slotted ALOHA with BEB policy with p_s as a parameter. As seen, when the success probability p_s is large, the difference between the curves for $r_{\max} = 5$ and $r_{\max} = 9$ is indistinguishable. When p_s is small, say $p_s < 0.6$, the curve for $r_{\max} = 5$ is higher than that for $r_{\max} = 9$ over the entire delay range. This is expected, as the denominator in (28) is a monoincreasing function of $r_{\max}$. The smaller delay for $r_{\max} = 5$ over $r_{\max} = 9$ is at the expense of a larger blocking probability. The results for slotted

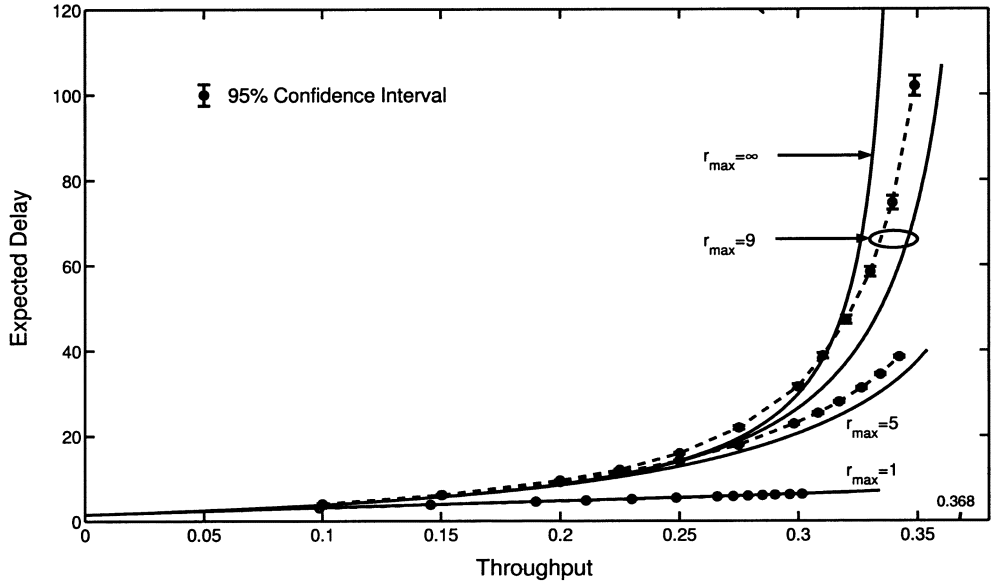

Fig. 8. Expected delay $\overline{D}$ versus throughput S, slotted ALOHA, BEB policy, and $\omega = 32$.

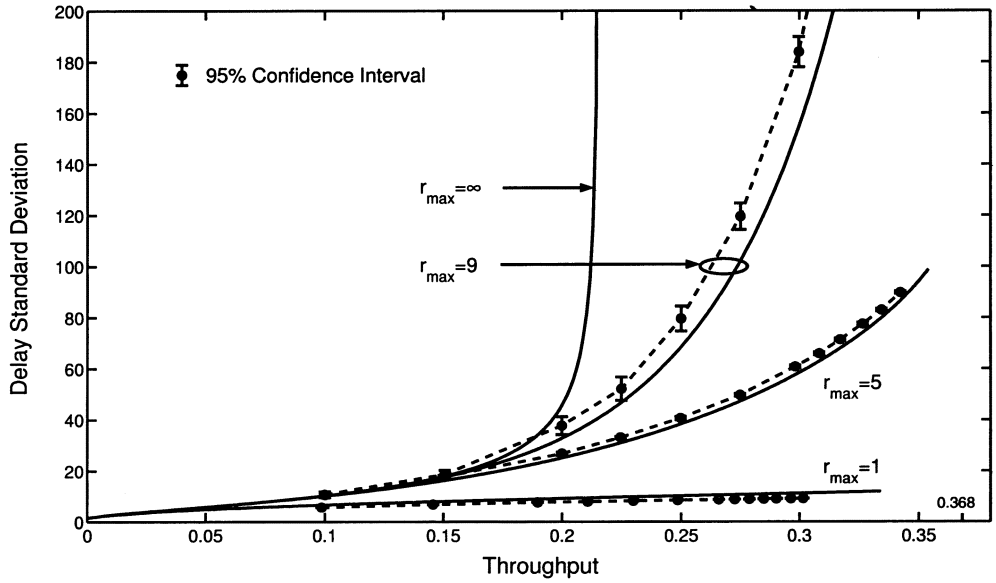

Fig. 9. Delay standard deviation σ_D versus throughput S, slotted ALOHA, BEB policy, and $\omega = 32$.

nonpersistent CSMA and the other two retransmission policies are quite similar (not shown).

Figs. 8 and 9 show the expected delay and the delay standard deviation (both in unit of slots) for slotted ALOHA with BEB policy for various values of $r_{\max}$. The 95% confidence intervals are shown for all the simulation points. We see that for large $r_{\max}$, say $r_{\max} = 9$, the analytical model severely underestimates the delay for systems operating close to the capacity, just like the classical result when $r_{\max} = \infty$. But for $r_{\max} = 5$ (a moderate value), the model gives an accurate prediction of the first two moments of delay. This result is expected, because if the number of retransmissions is smaller, the correlation of packet arrivals is also smaller. Therefore, the combined new and retransmitted traffic is less "bursty," or more Poisson-like, leading to a closer match between the simulation and the analytical results. Similar results and conclusions can be drawn for the other two backoff policies.

For the special case $r_{\max} = 5$, Fig. 10 compares the cumulative distribution functions of delay under three different retransmission policies for slotted ALOHA protocol. For the sake of comparison, we let the fixed backoff range in UB policy equal the initial backoff range in BEB policy, and let the retransmission parameter in GB policy be $q = 2/33$. In doing so, the expected values of first backoff delay $E[W_1]$ under these three policies are the same (see Table I). The curves for UB policy are similar to the corresponding curves for BEB policy, except that they approach unity much faster as x goes large. This is because the backoff range in UB policy is fixed so that the probability that a particular slot is chosen for a packet's next retransmission is larger than that in BEB policy. To illustrate, consider the UB

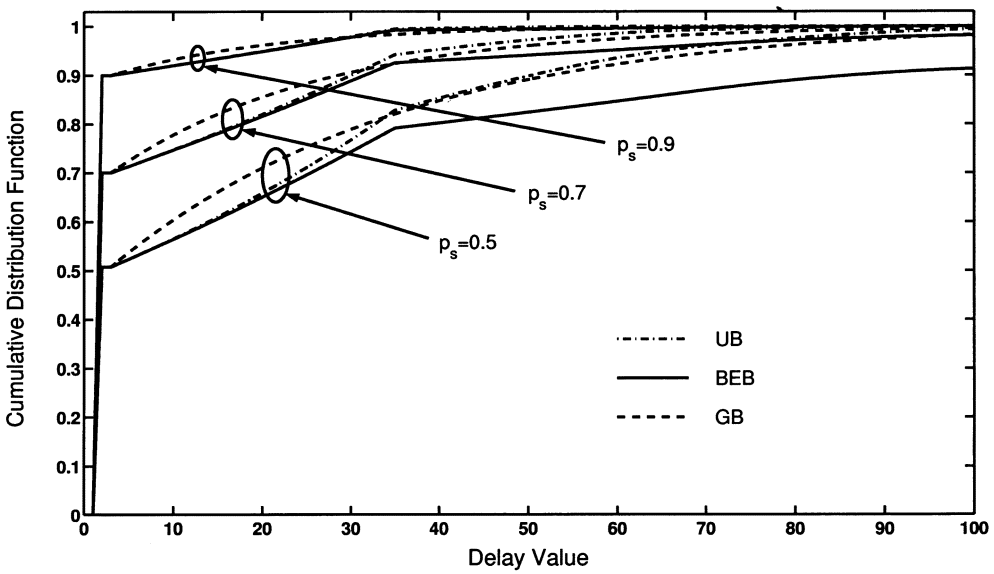

Fig. 10. Cumulative distribution function of delay $F_D(x)$, slotted ALOHA, $r_{\max} = 5, \omega = 32$, and $q = 2/33$.

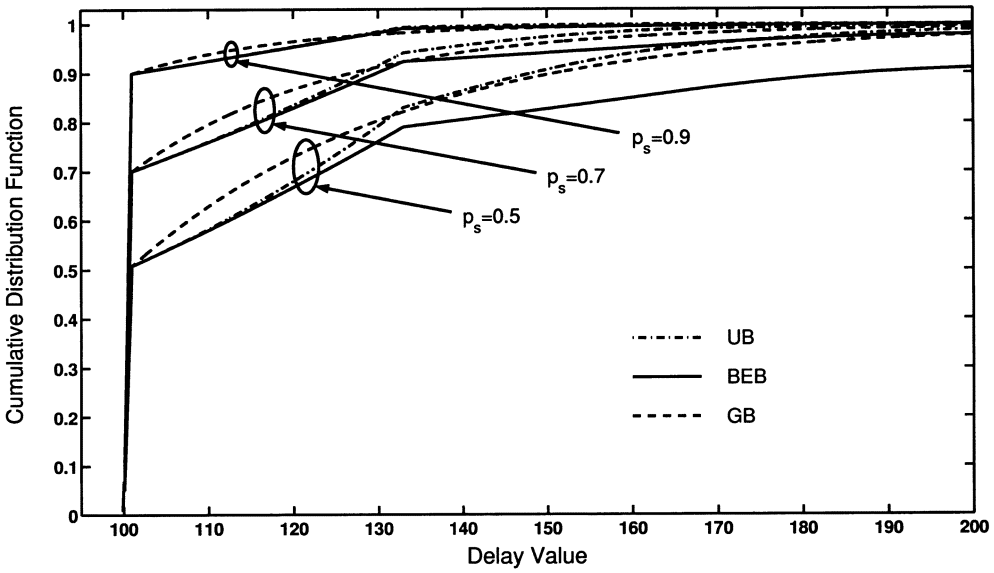

Fig. 11. Cumulative distribution function of delay $F_D(x)$, slotted nonpersistent CSMA, $a = 0.01, r_{\max} = 5, \omega = 32$, and $q = 2/33$.

curves for large p_s values, the curve slopes in the right-hand side range close to 35 and can be approximated by

$$h_3 = \frac{1}{\omega} P\{R' = 2\} = \frac{p_s(1-p_s)^2}{32[1-(1-p_s)^6]} \tag{31}$$

which is about twice that of the BEB curves (approximated by h_2). The retransmissions in GB policy are not limited by any backoff ranges, the resulting GB curves are therefore smooth over the entire delay range.

For slotted nonpersistent CSMA protocol, the delay distributions (in unit of slots) under different retransmission policies are compared in Fig. 11. As seen, these $F_D(x)$ curves are slightly different from those for slotted ALOHA protocol. First, we choose parameter $a = 0.01$ so that the slot size $\tau = aT$ here is one-hundredth of the packet transmission time T. Second, since D_0 is now uniformly distributed in $(T, T + aT]$, we have $F_D(100) = 0$ and $F_D(101) = P\{R' = 0\}$. Finally, for CSMA protocol, when the channel is sensed *busy*, the packet can attempt at the next slot. So, the access delay x is continuous starting from the point $x = 100$. This is different from the case in slotted ALOHA protocol, where the interval (2,3] is a gap in the entire delay range, and hence, we have $F_D(3) = F_D(2)$. As a result, the curve knees of both UB and BEB curves are now located at the point $x = 101 + \omega = 133$.

VIII. Conclusions

We have derived the delay distributions of slotted ALOHA and nonpersistent CSMA under three retransmission policies. For BEB policy, the conditions for finite average delay and finite delay variance are also derived. In addition, we have studied the effect of finite $r_{\max}$ on the blocking and delay performance.

Extending the results to unslotted channel model and other random access protocols is straightforward. Further generalization to variable packet size case [3], to other retransmission policies [9], and to *delayed-first-transmission* (DFT) scheme [5] should also be possible.

APPENDIX A
DERIVATION OF (6)

Since X_r is the sum of r independent random variables W_i's, the probability mass function of X_r is given by

$$P\{X_r = k\} = P\{W_1 = k\} * P\{W_2 = k\} * \cdots * P\{W_r = k\} \quad \text{(A.1)}$$

where "$*$" represents convolution operation.

Consider first the UB policy. Let sequence $a_1(n)$ equal to one for $1 \le n \le \omega$ and zero otherwise. Then

$$P\{W_i = k\} = \frac{a_1(k)}{\omega}. \quad \text{(A.2)}$$

(For simplicity, we only specify the range of k for which the probability is nonzero in the following derivations.) Further, define sequence $a_r(n)$ as

$$a_r(n) = \underbrace{a_1(n) * a_1(n) * \cdots * a_1(n)}_{r}. \quad \text{(A.3)}$$

According to the definition, $a_r(n),\ (r \ge 2)$ can be recursively derived as[3]

$$\begin{aligned} a_r(n) &= a_{r-1}(n) * a_1(n) \\ &= \sum_{j=n-\omega}^{n-1} a_{r-1}(j) \\ &= a_r(n-1) + a_{r-1}(n-1) \\ &\quad - a_{r-1}(n-\omega-1). \end{aligned} \quad \text{(A.4)}$$

The distribution of X_r is simply

$$P\{X_r = k\} = \frac{a_r(k)}{\omega^r}, \quad k = r, r+1, \ldots, r\omega \ \text{UB}. \quad \text{(A.5)}$$

For BEB policy, we define sequence $b_1(n)$ exactly the same as $a_1(n)$. Hence, the probability mass function of W_1 can be expressed as $P\{W_1 = k\} = b_1(n)/\omega$. Since under BEB policy, backoff range is doubled every time an unsuccessful transmission occurs, sequence $b_r(n)$ is defined as

$$b_r(n) = b_1(n) * [b_1(n) + b_1(n-\omega)] * \cdots * \left[\sum_{j=0}^{2^{r-1}-1} b_1(n - j\,\omega)\right] \quad \text{(A.6)}$$

which can also be recursively calculated[4]

$$\begin{aligned} b_r(n) &= b_{r-1}(n) * \left[\sum_{j=0}^{2^{r-1}-1} b_1(n - j\omega)\right] \\ &= \sum_{j=n-2^r\omega}^{n-1} b_{r-1}(j) \\ &= b_r(n-1) + b_{r-1}(n-1) \\ &\quad - b_{r-1}(n - 2^{r-1}\omega - 1). \end{aligned} \quad \text{(A.7)}$$

[3]Note that the elements in sequence $a_r(n)$ are actually the *polynomial coefficients* of function $f_r(t) = (1 + t + t^2 + \cdots + t^{\omega-1})^r$ [10, pp. 77–78], where element $a_r(k),\ (r \le k \le r\omega)$ corresponds to the coefficient of t^{k-r}. Based on this observation, all elements in $a_r(n)$ can be derived, theoretically, by differentiating $f_r(t)$ with respect to t. But this approach is very complicated and unpractical when ω and r are large.

[4]Incidentally, sequence $b_r(n)$ can also be shown as the coefficients of a function, say $g_r(t) = \left(\sum_{j=0}^{\omega-1} t^j\right)^r \prod_{j=0}^{m-2} \left(1 + t^{2^j\omega}\right)^{r-j-1}$, where element $b_r(k)$ corresponds to the coefficient of t^{k-r}.

Therefore, the distribution of X_r is given by

$$P\{X_r = k\} = \frac{b_r(k)}{2^{r(r-1)/2}\omega^r}, \quad k = r, r+1, \ldots, (2^r - 1)\omega \ \text{BEB}. \quad \text{(A.8)}$$

Consider finally the GB policy. Let sequence $c_1(n)$ be a discrete unit step function whereby $c_1(n) = 1$ for $n \ge 1$ and $c_1(n) = 0$, otherwise. Hence, the probability mass function of W_i can be expressed as $P\{W_i = k\} = c_1(k)(1-q)^{k-1}q$. Similar to $a_r(n)$, sequence $c_r(n)$ is defined as

$$c_r(n) = \underbrace{c_1(n) * c_1(n) * \cdots * c_1(n)}_{r} \quad \text{(A.9)}$$

and can be recursively derived as

$$\begin{aligned} c_r(n) &= c_{r-1}(n) * c_1(n) \\ &= \sum_{j=r-1}^{n-1} c_{r-1}(j) \\ &= c_r(n-1) + c_{r-1}(n-1). \end{aligned} \quad \text{(A.10)}$$

The distribution of X_r is simply

$$P\{X_r = k\} = c_r(k) q^r (1-q)^{k-r}, \quad k = r, r+1, \ldots \ \text{GB}. \quad \text{(A.11)}$$

APPENDIX B
DERIVATION OF (21)

Conventionally, busy period analysis is used to derive throughput of random access protocols [1]–[4], [7]. Fig. 12 shows the busy and idle periods for slotted nonpersistent CSMA protocol. Let B and I denote the length of busy and idle periods, respectively. Follow the approach in [4, pp. 90–91],[5] the average busy period $\overline{B}$ is given by

$$\overline{B} = \frac{(1+a)T}{e^{-aG}} \quad \text{(B.1)}$$

and the average idle period $\overline{I}$ is

$$\overline{I} = \frac{aT}{1 - e^{-aG}}. \quad \text{(B.2)}$$

Consider a tagged packet (not shown in Fig. 12) accessing the channel. In a typical cycle shown in Fig. 12, there are $\overline{I}/(aT)$ points (marked with "◯") where the tagged packet will be successfully transmitted if it attempts (no other packets are arrived in the last slot before these points). At the points marked with "◊" (totally $\overline{B}((1+a)T/(aT) - 1)/((1+a)T)$ such points), the channel is sensed busy and no transmission (including the tagged packet) will take place. The remaining $\overline{B}/((1+a)T)$ points marked with "△" are the instants where the channel is

[5]A different approach, which offers the same results of throughput, p_s, p_b, and p_c, is presented in [7, pp. 312–315].

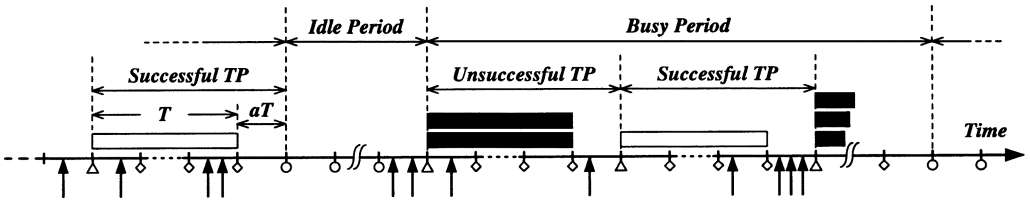

Fig. 12. Busy and idle periods in slotted nonpersistent CSMA protocol. Up arrows point at packet arrival instants. TP is the abbreviation for *transmission period*.

sensed idle but a collision will occur if the tagged packet is transmitted (one or more packets have already arrived in the last slot before these points). Based on the above analysis, the success probability p_s can be derived as

$$p_s = \frac{\text{average number of ``}\bigcirc\text{'' points in a cycle}}{\text{average number of ``}\bigcirc\text{'', ``}\Diamond\text{'', and ``}\triangle\text{'' points in a cycle}} = \frac{\dfrac{\overline{I}}{(aT)}}{\dfrac{(\overline{I}+\overline{B})}{(aT)}} = \frac{ae^{-aG}}{1+a-e^{-aG}}. \tag{B.3}$$

The throughput S is simply $S = G \cdot p_s$.

By the same argument

$$\begin{aligned} p_b &= P\{\text{channel is sensed busy}\} \\ &= \frac{\dfrac{\overline{B}}{(1+a)T}\left(\dfrac{(1+a)T}{aT}-1\right)}{\dfrac{(\overline{I}+\overline{B})}{(aT)}} \\ &= \frac{1-e^{-aG}}{1+a-e^{-aG}} \end{aligned} \tag{B.4}$$

$$\begin{aligned} p_c &= P\{\text{collision}\} \\ &= \frac{\dfrac{\overline{B}}{[(1+a)T]}}{\dfrac{(\overline{I}+\overline{B})}{(aT)}} \\ &= \frac{a(1-e^{-aG})}{1+a-e^{-aG}}. \end{aligned} \tag{B.5}$$

As a check, $p_s + p_b + p_c = 1$.

Recall R is the number of retransmissions needed for a successful transmission and K is the number of times the channel is sensed busy when accessed. Given $R = r$, the conditional probability $P\{K = k|R = r\}$ is given by

$$P\{K=k|R=r\} = \binom{r}{k}\left(\frac{p_b}{p_b+p_c}\right)^k\left(\frac{p_c}{p_b+p_c}\right)^{r-k}, \quad 0 \le k \le r. \tag{B.6}$$

Recall R has a geometric distribution with parameter p_s. The joint probability $P\{K = k, R = r\}$ is simply

$$\begin{aligned} P\{K=k,R=r\} &= P\{K=k|R=r\}P\{R=r\} \\ &= \binom{r}{k}p_b^k p_c^{r-k}p_s, \\ &\quad 0 \le k \le r. \end{aligned} \tag{B.7}$$

Equation (B.7) is used in the text as (21). As a check, the distribution of K is

$$\begin{aligned} P\{K=k\} &= \sum_{r=k}^{\infty} P\{K=k,R=r\} \\ &= \frac{p_b^k p_s}{(1-p_c)^{k+1}}, \quad k \ge 0 \end{aligned} \tag{B.8}$$

and the mean value of K is

$$E[K] = \frac{p_b}{p_s} = (E[R]+1)p_b \tag{B.9}$$

as expected.

REFERENCES

[1] L. Kleinrock and F. A. Tobagi, "Packet switching in radio channels: Part I—Carrier sense multiple-access modes and their throughput-delay characteristics," *IEEE Trans. Commun.*, vol. COM-23, pp. 1400–1416, Dec. 1975.

[2] L. Kleinrock, *Queueing Systems Volume II: Computer Applications*. New York: Wiley, 1976.

[3] F. A. Tobagi and V. B. Hunt, "Performance analysis of carrier sense multiple access with collision detection," *Comput. Networks*, vol. 4, pp. 245–259, Oct./Nov. 1980.

[4] R. Rom and M. Sidi, *Multiple Access Protocols: Performance and Analysis*. New York: Springer-Verlag, 1990.

[5] F. A. Tobagi, "Distribution of packet delay and interdeparture time in slotted ALOHA and carrier sense multiple access," *J. Assoc. Comput. Mach.*, vol. 29, pp. 907–927, Oct. 1982.

[6] S. L. Beuerman and E. J. Coyle, "The delay characteristics of CSMA/CD networks," *IEEE Trans. Commun.*, vol. 36, pp. 553–563, May 1988.

[7] J. L. Hammond and P. J. P. O'Reilly, *Performance Analysis of Local Computer Networks*. Reading, MA: Addison-Wesley, 1986.

[8] K.-C. Chen, "Medium access control of wireless LANs for mobile computing," *IEEE Network Mag.*, pp. 50–63, Sept./Oct. 1994.

[9] M. A. Marsan and F. Neri, "A simulation study of delay in multichannel CSMA/CD protocols," *IEEE Trans. Commun.*, vol. 39, pp. 1590–1602, Nov. 1991.

[10] L. Comtet, *Advanced Combinatorics: The Art of Finite and Infinite Expansions*. Amsterdam, The Netherlands: Reidel, 1974.

Yang Yang (S'99–A'99–M'02) received the B.Eng. and M.Eng. degrees in radio engineering from Southeast University, Nanjing, China, in 1996 and 1999, respectively, and the Ph.D. degree in information engineering from The Chinese University of Hong Kong, Hong Kong, in 2002.

From 1996 to 1999, he was a Research Assistant with the National Mobile Communication Laboratory at Southeast University, Nanjing, China. In August 2003, he joined the Department of Electronic and Computer Engineering, Brunel University, Uxbridge, U.K., as a Lecturer. Before that, he had spent one year in the Department of Information Engineering, The Chinese University of Hong Kong, as an Assistant Professor. His general research interests include radio channel characterization, dynamic resource allocation for integrated services, and performance evaluation of multiple access protocols.

Tak-Shing Peter Yum (S'76–A'78–SM'86) was born in Shanghai. He received the B.S., M.S., and Ph.D. degrees from Columbia University, New York, NY in 1974, 1975, and 1978, respectively.

He joined Bell Telephone Laboratories in April 1978 working on switching and signaling systems. In 1980, he accepted a teaching appointment at the National Chiao Tung University, Taiwan. Then, in 1982, he joined the Chinese University of Hong Kong, where he is now Professor of Information Engineering. He has published original research on packet switched networks with contributions in routing algorithms, buffer management, deadlock detection algorithms, message resequencing, and multiaccess protocols. He then branched out to work on the design and analysis of cellular network, lightwave networks, and video distribution networks. His recent works are on the technologies for the 3G and IP networks.