Buffered Aloha with $K$-Exponential Backoff
Part II: Delay Analysis

Lin Dai, Member, IEEE, and Tony T. Lee, Fellow, IEEE

Abstract—This paper presents the delay analysis for buffered Aloha networks with $K$-Exponential Backoff. Mean access delay and mean queuing delay are derived and demonstrated via the examples of Geometric Retransmission ($K=1$) and Exponential Backoff ($K=\infty$). The comparison shows that higher delay is incurred with Geometric Retransmission when the aggregate input rate is small, and the delay gap is enlarged as the number of nodes $n$ increases. With a high traffic input rate, however, the delay performance with Exponential Backoff severely deteriorates. The mean queuing delay will be unbounded if the aggregate input rate exceeds 0.3.

We also extend the analysis to the contention-window-based backoff model which is widely adopted in practical MAC protocols. It will be revealed that both the retransmission-probability-based and the contention-window-based models exhibit the same stable region and achieve similar queuing performance in most cases, which justifies the intuition that was taken but remained unverified in previous studies: the retransmission-probability-based backoff model can serve as a good approximation of the contention-window-based one.

Index Terms—Slotted Aloha, exponential backoff, geometric retransmission, stability, queuing delay, access delay, contention window

I. INTRODUCTION

This is Part II of the paper series that studies the performance of buffered Aloha networks with $K$-Exponential Backoff. Stability and throughput analysis has been presented in Part I. Delay performance will be further investigated in this paper.

Delay analysis of buffered Aloha networks has long remained an open problem. In a buffered Aloha network, each node is equipped with an infinite or finite queue, and shares a common server. The total queuing delay of each packet consists of two parts: 1) waiting time, which is the time interval from the packet’s arrival to the instant that it becomes the head-of-line (HOL) packet; and 2) access delay, which is the time interval from the instant that it becomes the HOL packet to its successful transmission. Note that due to contention, the server may be paralyzed if the attempt rate is too high. In that case, the whole network will become unstable with unbounded delay.

Intuitively, a buffered Aloha system can be modeled as a multi-dimensional Markov chain with state vector $(C_1, C_2, \ldots, C_n)$, where $C_i$ represents the queue length of node $i, i=1,\ldots,n$ [1]. The ultimate solution for the case of $n=2$ has been characterized in [2]; nevertheless, the approach becomes intractable as the dimension exceeds three [3-6]. To simplify the analysis, various approximate methods have been developed to tackle the coupled queues [7-18]. Although most of them were customized for specific system configurations, the idea of decomposition gradually becomes a consensus: each node can be approximately treated as an independent queueing system with identically distributed service time [13-18]. The key to delay analysis is the characterization of service time distribution, which unfortunately remains quite elusive.

In fact, stability analysis is a prerequisite for delay analysis in random access networks. In contrast to traditional queueing systems, the service rate of each single queue in a random access network is determined by the aggregate activities of all the nodes. The queuing delay will become infinite if the system collapses, i.e., the aggregate service rate drops below the aggregate input rate. It is therefore important to characterize the system stable region first.

The stability analysis for $n$-node buffered Aloha networks with $K$-Exponential Backoff has been presented in Part I of the paper series, where the original Slotted Aloha (which is referred to as Geometric Retransmission) and Exponential Backoff are included as two special examples with the cutoff phase $K=1$ and $\infty$, respectively. Assume that each node is equipped with an infinite buffer and has Bernoulli arrivals with an input rate of $\lambda$ packets per time slot. We characterize the stable region $S$ of retransmission factor $q$, within which a network throughput of $\hat{\lambda}_{\text{out}}=\hat{\lambda}=n\lambda$ can be achieved for any aggregate input rate $0<\hat{\lambda}\leq\hat{\lambda}_{\text{max},S}=e^t$.

It will be further revealed in this paper that although the throughput stability can be always achieved with any $q\in S$, the delay performance is critically dependent on the value of retransmission factor $q$, and might be drastically different with various backoff schemes. For instance, with Geometric Retransmission ($K=1$), the mean queuing delay of each packet will quickly decrease as the retransmission probability increases, and is minimized when $q$ reaches the upper bound of the stable region $S_{\text{geo}}$. The maximum stable throughput $\hat{\lambda}_{\text{max},S}=e^t$ can be achieved with $q=1/n$, and the corresponding queuing delay linearly increases with the number of nodes $n$.

Manuscript received June 23, 2009.

Lin Dai is with the Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong (e-mail: lindai@cityu.edu.hk).

Tony T. Lee is with the Department of Information Engineering, Chinese University of Hong Kong, N.T., Hong Kong (e-mail: ttlee@ie.cuhk.edu.hk).
In contrast, the delay performance of Exponential Backoff \((K=\infty)\) networks is insensitive to the network population \(n\) when the aggregate input rate \(\lambda\) is small. Much lower queueing delay is therefore incurred compared to Geometric Retransmission, especially when the number of nodes \(n\) is large. Nevertheless, the delay performance deteriorates sharply as the traffic level increases. Our analysis shows that with \(q \in S^0\), the Exponential Backoff system will become quasi-stable if the aggregate input rate \(\lambda \geq 0.3\). The maximum stable throughput \(\lambda_{\text{max}} \approx e^{-1}\) is achieved at the cost of infinite mean queueing delay.

Note that in spite of severely penalized delay performance, Exponential Backoff networks provide the best robustness: a stable throughput can be achieved with the retransmission factor \(q\) fixed to be \(1-e^{-1}\) regardless of the traffic input rate or the network population. With Geometric Retransmission, however, the stable region rapidly shrinks as the number of nodes \(n\) increases. Any slight change of network population \(n\) will lead to a risk of system collapse if the retransmission factor \(q\) is not updated accordingly.

The performance comparison of Geometric Retransmission \((K=1)\) and Exponential Backoff \((K=\infty)\) suggests that for general \(K\)-Exponential Backoff systems, the cutoff phase \(K\) can serve as a leverage to strike a balance between queueing performance and system robustness. Given the retransmission factor \(q\), increasing \(K\) can greatly improve the maximum stable throughput that a network can achieve. On the other hand, it may also impair the queueing delay performance when the aggregate input rate \(\lambda\) is high. The optimal value of cutoff phase \(K\) is subject to the system requirements such as the range of traffic input rate and the delay constraint.

Note that in practical MAC protocols, a contention-window mechanism is usually adopted \([8, 17-20]\), which is slightly different from the retransmission-probability-model supposed in most analytical studies \([2-7, 21-24]\). Our analysis will further demonstrate that both models indeed share the same stable region. Lower queueing delay is incurred by the window-based backoff model; nevertheless, the delay gap is quite small unless the number of nodes \(n\) is large. Nevertheless, our analysis shows that with increasing \(K\), the delay performance deteriorates sharply as the traffic level is subject to the system requirements such as the range of traffic input rate and the delay constraint.

The remainder of this paper is organized as follows. Section II presents the delay analysis of \(K\)-Exponential Backoff, where Geometric Retransmission and Exponential Backoff are investigated as exemplary cases, and the tradeoff between queueing delay performance and system robustness is characterized. The analysis is further extended to the contention-window-based backoff model in Section III. Conclusions are summarized in Section IV.

II. QUEUEING DELAY OF \(K\)-EXPONENTIAL BACKOFF

As demonstrated in Part I, an \(n\)-node buffered Aloha network can be modeled as a non-work-conservative \(n\)-queue-single-server system. Each node is supposed to be equipped with an infinite buffer and the coupled queues are decomposed into independent FIFO queues with Bernoulli arrivals of rate \(\lambda\) packets per time slot. These decomposed Geo/G/1 queues are then hinged together by the probability of success \(p\) of the HOL packets. Fig. 1 presents the phase model that was established in Part I to describe the state transition process of each individual HOL packet. With \(K\)-Exponential Backoff, the transmission probability of a HOL packet in phase \(i\) is \(q_i\), \(i=0,1,\ldots, K\).

\[
\begin{array}{ccccccccccc}
\text{Fig. 1. State transition diagram of HOL packets.}
\end{array}
\]

It can be seen from Fig. 1 that the service time \(X\) of a HOL packet is the sum of the time spent in successive phases, which is determined by both the probability of success \(p\) and the retransmission factor \(q\). Let \(Y_i\) denote the time spent from the beginning of phase \(i\) until the service completion, and \(Y_0\) be the time spent in phase \(i\). We have

\[
X_i = \begin{cases} Y_i & \text{with probability } p, \\ Y_i + X_{i+1} & \text{with probability } 1-p, \end{cases} \quad i=0,1,\ldots, K-1, \quad X_K = Y_K.
\]

With \(K\)-Exponential Backoff, the sojourn time at phase \(i\), \(Y_i\), is a geometric distributed random variable with the probability generating function

\[
G_{Y_i}(z) = \frac{pq^i z}{1-(1-p)q^i z}, \quad i=0,1,\ldots, K-1
\]

\[
G_{Y_0}(z) = \frac{pq^0 z}{1-(1-p)q^0 z}.
\]

From (3), we immediately have

\[
G_{Y_0}(1) = 1 + \frac{1-p}{q(1-p)} \left( \frac{1-p}{q(1-p)} \right)^K = \left( \frac{1-p}{q} \right)^K.
\]

and

\[
G_{Y_0}(1) = \frac{2(1-p)}{q(1-p)} + \frac{2(1-p)q^K}{q(1-p)} \left( \frac{1-p}{q} \right)^K - \frac{2(1-p)}{q^2(1-p)} \left( \frac{1-p}{q} \right)^K.
\]

The mean and variance of the service time \(X\) are then given by

\[
E[X] = G_{Y_0}'(1), \quad \text{and} \quad \text{var}[X] = G_{Y_0}'(1) + G_{Y_0}(1) - G_{Y_0}'(1)^2,
\] respectively.
Finally, the mean queueing delay of input packets can be obtained by substituting (4-5) into the Pollaczek-Kintchine (P-K) formula given in Appendix I:

\[ E[T] = G_{x_{nu}}(1) + \frac{\lambda G_{x_{nu}}^n(1)}{2(1 - \lambda G_{x_{nu}}^n(1))}. \]  

(8)

It can be seen from (6) and (8) that mean access delay \( E[X] \) and mean queueing delay \( E[T] \) are both determined by the probability of success \( p \) and the retransmission factor \( q \). The stable region \( S \) of retransmission factor \( q \) has been characterized in Part I, within which a stable throughput can be achieved. Apparently, the mean queueing delay becomes infinite if the retransmission factor \( q \) falls outside the stable region \( S \).

Define \( S_D \) as the delay-stable region of retransmission factor \( q \), within which the mean queueing delay of input packets is finite:

\[ S_D = \{ q | E[T] < \infty \}. \]  

(9)

It is plain to see that \( S_D \) is a subset of the stable region \( S \), i.e., \( S_D \subseteq S \), because a stable throughput is a necessary but not sufficient condition for finite mean queueing delay \( E[T] < \infty \). We are only interested in the delay performance within the delay-stable region \( S_D \). We will take the examples of Geometric Retransmission \( (K=1) \) and Exponential Backoff \( (K=\infty) \) to demonstrate the above results.

A. Queueing Delay of Geometric Retransmission

When the cutoff phase \( K=1 \), the first and second moments of service time \( X \) can be obtained from (4-5) as

\[ G_{x_{nu}}(1) = 1 + \frac{1 - p}{pq}, \]  

(10) and

\[ G_{x_{nu}}(1) = \frac{2(1 - p)}{(pq)^2}. \]  

(11)

The mean access and mean queueing delay of Geometric Retransmission are then given by

\[ E[X]_{\text{Geo}} = 1 + \frac{1 - p}{pq}, \]  

(12)

\[ E[T]_{\text{Geo}} = 1 + \frac{1}{pq} - \frac{(1 - p)}{pq} \frac{1}{q}. \]  

(13)

It has been demonstrated in Part I that an \( n \)-node network with Geometric Retransmission has a stable throughput if and only if the retransmission factor \( q \) is selected from the stable region

\[ S_{\text{Geo}} = \left\{ \frac{\hat{\lambda}(1 - p_L)}{p_L(n - \hat{\lambda})}, -\ln p_S \right\}, \]  

(14)

and the probability of success \( p \) will converge to the desired stable point \( p_L \). Delay stability can be achieved within the stable region \( S_{\text{Geo}} \) except when \( q \) reaches the lower bound, with which the offered load of each queue \( \rho = 1 \). As a result, the delay-stable region of Geometric Retransmission is given by

\[ S_{D_{\text{Geo}}} = \left\{ \frac{\hat{\lambda}(1 - p_L)}{p_L(n - \hat{\lambda})}, -\ln p_S \right\}, \]  

(15)

and the corresponding maximum stable throughput that the network can achieve with finite mean queueing delay, \( \hat{\lambda}_{\text{max}_{\text{Geo}}} \), is given by

\[ \hat{\lambda}_{\text{max}_{\text{Geo}}} = \hat{\lambda}_{\text{max}_{\text{Geo}}} = e^{-\hat{\lambda}}. \]  

(16)

The desired stable point \( p_L \) is solely determined by the aggregate input rate \( \hat{\lambda} = n \hat{\lambda} \). According to (12-13), both mean access delay \( E[X] \) and mean queueing delay \( E[T] \) will decrease as the retransmission factor \( q \) increases inside the delay-stable region \( S_{\text{Geo}} \). The minimum mean access delay and the minimum mean queueing delay of Geometric Retransmission can be then obtained as

\[ \min_{q \in S_{\text{Geo}}} E[X]_{\text{Geo}} = 1 + \frac{1 - p_L}{p_L} \frac{n}{-\ln p_S}, \]  

(17)

\[ \min_{q \in S_{\text{Geo}}} E[T]_{\text{Geo}} = 1 + n \left( \frac{1}{p_L} \ln p_S - \frac{\hat{\lambda}(1 - p_L) \ln p_S}{1 - \frac{1}{\hat{\lambda}} n} \right), \]  

(18)

which can be achieved when

\[ q = -\frac{1}{n} \ln p_S = -\frac{1}{n} W_\lambda (-\hat{\lambda}). \]  

(19)

It is clear from (17-18) that both the minimum mean access delay and the minimum mean queueing delay increase linearly with the number of nodes \( n \). They also increase with the aggregate input rate \( \hat{\lambda} \).

Fig. 2 shows the minimum mean access delay and minimum mean queueing delay versus \( \hat{\lambda} \) under different values of \( n \). It can be seen from Fig. 2 that the queueing delay is close to the access delay only when traffic is light. A sharp increase of delay (both access delay and queueing delay) can be observed when \( \hat{\lambda} \) approaches the maximum \( \hat{\lambda}_{\text{max}_{\text{Geo}}} = e^{-1} \), and the delay is doubled when the number of nodes \( n \) increases from 50 to 100. The simulation results presented in Fig. 2 perfectly agree with the analysis.
To achieve the best delay performance, however, the retransmission factor \( q = -\frac{1}{\lambda} W_{\lambda}(-\hat{\lambda}) \) needs to be adjusted whenever the traffic level changes. Even a slight increment of \( \hat{\lambda} \) could lead to system breakdown if \( q \) is not updated in time. In practice, it is desired that the network can operate in stable states under any traffic input rate \( 0 < \hat{\lambda} \leq \hat{\lambda}_{\text{max,} \hat{\lambda}}^{\text{Geo}} = 0.4 \). According to (14), this can be achieved with the retransmission factor \( q \) set to be \( q_{\text{Geo}}^{\text{max}} = 1/n \).

With \( q = q_{\text{Geo}}^{\text{max}} = 1/n \), the corresponding mean access delay and mean queuing delay are given by:

\[
E[X]^{\text{Geo}}_{q^{\text{max}}_n} = 1 + n \left( 1 - \frac{1}{p_L} \right), \quad (20)
\]

\[
E[T]^{\text{Geo}}_{q^{\text{max}}_n} = 1 + n \left( \frac{1}{p_L} - \frac{\hat{\lambda}(1-p)_{\lambda \rightarrow \hat{\lambda}}}{1-1/n} \right) \quad (21)
\]

Both of them again increase with the number of nodes \( n \) and the aggregate input rate \( \hat{\lambda} \).

The minimum mean queuing delay and the mean queuing delay with \( q = q_{\text{Geo}}^{\text{max}} \) are presented in Fig. 3. It can be clearly seen that the delay gap increases with the aggregate input rate \( \hat{\lambda} \). A closer look at (18) and (21) further suggests that the gap also increases with the number of nodes \( n \). As we have illustrated, \( q_{\text{Geo}}^{\text{max}} = 1/n \) is a more practical choice. Part of the delay performance has to be sacrificed in return of better system robustness.

\[ G_{s_0}^*(1) = 1 + \frac{1-p}{q - (1-p)} \quad (22) \]

\[ G_{s_0}^*(1) = \frac{2(1-p)q}{[q - (1-p)][q^2 - (1-p)]} \quad (23) \]

if the retransmission factor satisfies

\[ q > 1/\hat{\lambda} \quad (24) \]

Otherwise the second moment of \( X \) will become infinite. The corresponding mean access and mean queuing delay of Exponential Backoff are given by

\[ E[X]^{\text{Exp}} = 1 + (1-p)/(p + q - 1), \quad (25) \]

\[ E[T]^{\text{Exp}} = 1 + \frac{1-p}{p + q - 1} + \frac{\hat{\lambda}(1-p)q}{(p + q - 1)(p + q^2 - 1)} \quad (26) \]

\[ B. I Delay-Stable Region of Exponential Backoff \]

The stable region of Exponential Backoff has been characterized in Part I. In particular, with a retransmission factor \( q \) selected from the absolute-stable region \( S_{\lambda}^{\text{Exp}} \), the network is guaranteed to operate at the desired stable point \( p_L \). When the network shifts to the undesired stable point \( p \approx 1-q \), a stable throughput can still be achieved if the retransmission factor \( q \) is selected from the stable region \( S_{\lambda}^{\text{Exp}} \). The complete stable region is therefore given by

\[ S_{\lambda}^{\text{Exp}} = S_{\lambda}^{\text{Exp}} \cup S_{\lambda}^{\text{Exp}} = [1 - p_L, 1 - p_S] \quad (27) \]

If stable throughput is the only concern, it does not matter whether the probability of success \( p \) converges to \( p_L \) or \( p_S \); throughput stability can always be achieved as long as the retransmission factor \( q \) is selected from the stable region \( S_{\lambda}^{\text{Exp}} \). Nevertheless, the delay performance is drastically different at the above two stable points. It is shown in Appendix II that when the probability of success \( p \) converges to the undesired stable point \( p_S \), (24) cannot be satisfied, implying that the mean queuing delay will become unbounded.

Similar to Geometric Retransmission, the Exponential Backoff system is guaranteed to operate at the desired stable point \( p_L \) if the retransmission factor \( q \) is selected from the absolute-stable region \( S_{\lambda}^{\text{Exp}} \), at which delay stability is achievable. Unfortunately, the absolute-stable region of Exponential Backoff will rapidly shrink as the number of nodes \( n \) increases. With \( n = 100 \), for instance, \( S_{\lambda}^{\text{Exp}} \) becomes an empty set when the aggregate input rate \( \hat{\lambda} \) exceeds 0.045.

To characterize the delay-stable region of Exponential Backoff with a large number of nodes \( n \), we need to determine how the probability of success \( p \) varies with the retransmission factor \( q \) outside the absolute-stable region, which is difficult to be obtained. Nevertheless, it is shown in Appendix III that with

\[ \frac{-\ln p_S}{1 - \ln p_S} \leq q \leq 1 - p_S \quad (28) \]

the Exponential Backoff system will operate at the undesired stable point \( p_S \) with high probability, at which the mean queuing delay is unbounded. Moreover, if the system operates at the desired stable point \( p_L \), we can see from (24) that with

\[ 1 - p_L \leq q \leq \sqrt{1 - p_L} \quad (29) \]

the mean queuing delay will also become infinite.

Fig. 3. Mean queuing delay with Geometric Retransmission under different values of retransmission factor \( q \). \( n = 50 \).
As shown in Fig. 4, despite a stable throughput, the mean queueing delay will grow unboundedly if the retransmission factor $q$ falls into the shadowing area, indicating that the network is quasi-stable.

By combining (28) and (29), we can conclude that for large number of nodes $n$, the delay-stable region of Exponential Backoff should satisfy:

$$\hat{\lambda}_{\text{max,D}}^\text{Exp} \subseteq \left(\sqrt{1-p_L}, \frac{-\ln p_S}{1-\ln p_L}\right).$$  \hspace{1cm} (30)

It can be readily seen from (30) and Fig. 4 that for Exponential Backoff, the maximum stable throughput with finite mean queueing delay, $\hat{\lambda}_{\text{max,D}}^\text{Exp}$, satisfies

$$\hat{\lambda}_{\text{max,D}}^\text{Exp} < \hat{\lambda}_0 < \lambda_{\text{max,S}}^e = e^{-1},$$  \hspace{1cm} (31)

where $\hat{\lambda}_0$ is the root of the following equation

$$\sqrt{1-p_L} = \frac{-\ln p_S}{1-\ln p_L}.\hspace{1cm} (32)$$

Infinite mean queueing delay will be incurred with Exponential Backoff when the aggregate input rate $\hat{\lambda}$ exceeds $\hat{\lambda}_0=0.3$.

It has been shown in Fig. 17 of Part I that with $\hat{\lambda}=0.3$, the Exponential Backoff system will operate at the undesired stable point $p_L$, with any retransmission factor $q \in S^\text{Exp}$. As we have mentioned in Section IV. B, Part I, “capture phenomenon” [25-27] occurs when the Exponential backoff system operates at the undesired stable point $p_L$. In that case, many nodes will be pushed to large phases with extremely small retransmission probabilities such that the node who once succeeds can dominate the channel for a long time and produce a continuous stream of packets. The network throughput can still strike a balance, but the output process is no longer stationary. The queue length of each single node will vary violently and fail to converge as time grows.

Fig. 5 plots the mean queueing delay of Exponential Backoff with the traffic input rate $\hat{\lambda}=0.3$ and the retransmission factor $q=1-e^{-1}$. It can be clearly seen that the mean queueing delay constantly varies with time. It rapidly mounts up once some HOL packets are blocked in deep phases, and starts declining when any of those nodes finally captures the channel and releases its accumulated packets. As time elapses, the mean queueing delay fails to follow any bound, because the retransmission probability of each HOL packet can be arbitrarily small. The second moment of service time in this case has indeed diverged. As a result, the mean queueing delay grows unboundedly, although a stable throughput can still be achieved in the long run.

B.2 Delay Performance with $q = q_m^\text{Exp} = 1-e^{-1}$

In contrast to Geometric Retransmission, the optimal delay performance of Exponential Backoff cannot be obtained as the delay-stable region with a large number of nodes $n$ is unknown. In the following, we will focus on the delay performance when $q = q_m^\text{Exp} = 1-e^{-1}$, with which a stable throughput can be achieved under any traffic input rate $0 < \hat{\lambda} \leq \hat{\lambda}_{\text{max,S}}^e = e^{-1}$.

Suppose that the system operates at the desired stable point, i.e., $p = p_L$, with $q = q_m^\text{Exp}$ and $\hat{\lambda} < \hat{\lambda}_0$. The mean access and queueing delay can be obtained from (25-26) as

$$\mathbb{E}[X^\text{Exp}_{k=n}] = \frac{1}{p_L-e^{-1}},\hspace{1cm} (33)$$

$$\mathbb{E}[T^\text{Exp}_{k=n}] = \frac{1-e^{-1}}{p_L-e^{-1}} + \frac{\hat{\lambda}(1-p_L)(1-e^{-1})}{n(p_L-e^{-1})(1-p_L)(1-e^{-1})+\hat{\lambda}(1-e^{-1})p_L-2e^{-1}+e^{-2}}.\hspace{1cm} (34)$$

It is interesting to see from (34) that with a large number of nodes $n$,

$$\mathbb{E}[T^\text{Exp}_{k=n}] \approx \mathbb{E}[X^\text{Exp}_{k=n}] = \frac{1}{p_L-e^{-1}},\hspace{1cm} (35)$$

implying that the delay does not vary with the number of nodes $n$ if the system operates at the desired stable point $p_L$. This can be clearly observed in Fig. 6.
Note that the gap between simulation and analytical results becomes noticeable as the aggregate input rate  \( \hat{\lambda} \) approaches  \( \hat{\lambda}_0 \). A closer observation indicates that the correlation among queues will become significantly high as the traffic input rate  \( \hat{\lambda} \) increases. The mounting correlation will lead to a departure of the probability of success from the theoretical value of  \( p_s \), which is derived based on the assumption of independent queues. When the aggregate input rate  \( \hat{\lambda} \) exceeds  \( \hat{\lambda}_0 \), the probability of success will converge to the undesired stable point  \( p_{sp} \), and the mean queueing delay will grow unboundedly.

It can be seen from Fig. 6 that the delay performance of Exponential Backoff is highly polarized. When the traffic input rate  \( \hat{\lambda} \) is lower than  \( \hat{\lambda}_0 \), significant improvement can be observed compared to Geometric Retransmission. The delay gap becomes even larger as the number of nodes  \( n \) increases, because the mean queueing delay with Geometric Retransmission linearly increases with  \( n \). Nevertheless, Geometric Retransmission has much better queueing performance when the traffic level is high. Finite mean queueing delay is incurred over the whole traffic range  \( 0 < \hat{\lambda} \leq \hat{\lambda}_{\text{max},D} = e^{-1} \).

Although the Exponential Backoff system becomes quasi-stable when traffic is heavy, it provides the best robustness. As shown in Fig. 4, the stable region of Exponential Backoff does not vary with the number of nodes  \( n \). The retransmission factor  \( q \) can be fixed to be  \( d_{\text{in}}^{\text{Exp}} = e^{-1} \) no matter how many nodes the network has. In contrast, the stable region of Geometric Retransmission will quickly shrink as the network population  \( n \) increases. According to (14), the retransmission factor  \( q \) has to be adjusted in the order of  \( 1/n \) to achieve stability. This indicates that any increase of number of nodes  \( n \) may lead to system collapse if  \( q \) is not updated in time.

Realtime feedback on the total number of nodes is critical for stabilizing the Geometric Retransmission systems.

### C. Tradeoff between Queueing Performance and Robustness

From the comparison on stability and delay performance of Geometric Retransmission (\( K=1 \)) and Exponential Backoff (\( K=\infty \)), we can see that with a small cutoff phase  \( K \), the system has good queueing performance but is prone to be unstable. Intuitively, nodes have to back off to larger phases to alleviate contention. If the cutoff phase  \( K \) is small, all the nodes will quickly reach the maximum phase  \( K \) and the retransmission probabilities cannot be reduced any more. As a result, the network has very limited capability to absorb the growing contention and will quickly collapse as the number of nodes  \( n \) or traffic input rate  \( \hat{\lambda} \) increases. The positive side, however, is that the queueing performance is quite good because the retransmission probability of each HOL packet is lower bounded by  \( q^K \).

With  \( K \) increasing, nodes will have more space to lower down their transmission requests as contention grows. An infinite cutoff phase  \( K \) indicates that nodes can always back off to deeper phases to make the attempt rate arbitrarily small until the network is stabilized. That is why the stable region of Exponential Backoff does not vary with the number of nodes  \( n \). Nevertheless, some HOL packets may suffer from extremely long delay due to the small retransmission probabilities, and the queueing delay may become unbounded as the variance of service time diverges when the traffic input rate is too high.

For  \( K \)-Exponential Backoff with  \( 1 < K < \infty \), the cutoff phase  \( K \) can serve as a leverage to achieve a tradeoff between queueing performance and system robustness. For illustration, let us fix the retransmission factor  \( q \) at  \( q^{\text{Exp}} = 1-e^{-1} \). As we have shown in Section II. B, with Exponential Backoff, the maximum stable throughput that a network can achieve with  \( q=1-e^{-1} \) is

\[
\hat{\lambda}_{\text{max},S,q^{\text{Exp}}} = e^{-1}.
\]

With Geometric Retransmission, however, the corresponding maximum stable throughput is given by

\[
\hat{\lambda}_{\text{max},S,q^{\text{Geo}}} = n(1-e^{-1}) \exp\{n(1-e^{-1})\}
\]

According to Eqs. (50) and (52) in Part I, which is close to zero when the number of nodes  \( n \) is large. In other words, the Geometric Retransmission system will collapse almost under any non-zero traffic input rate  \( \hat{\lambda} > 0 \) if  \( q = q^{\text{Geo}} = 1-e^{-1} \). It is clear that the Exponential Backoff and Geometric Retransmission systems provide the highest and the lowest robustness performance, respectively.

With an arbitrary cutoff phase  \( 1 < K < \infty \), the maximum stable throughput  \( \hat{\lambda}_{\text{max},S,q^{\text{Geo}}} \) can be derived according to Eqs. (50) and (82) in Part I. Since it does not have an explicit expression, we present the numerical results in Fig. 7.

As we can see from Fig. 7, the maximum stable throughput  \( \hat{\lambda}_{\text{max},S,q^{\text{Geo}}} \) sharply increases with the cutoff phase  \( K \) and quickly approaches  \( e^{-1} \). When  \( K \) is small, the system performance can be significantly improved even with a slight increment of  \( K \).
With a high traffic input rate, however, the queueing performance may start deteriorating as the cutoff phase $K$ increases. Suppose that the system operates at the desired stable point $p_\ell$ with $q=d_w^{\text{Exp}}$. We can see from (32) that when the traffic input rate $\hat{\lambda}$ exceeds $\hat{\lambda}_0$, we have
\[
\sqrt{1-p_\ell} > q = 1 - e^{-1}.
\]
According to (5), the second moment of service time will exponentially grow with the cutoff phase $K$ and finally become unbounded as $K=\infty$.

Let us set the delay constraint to be $G^*_K(1) < 1000$. The corresponding maximum stable throughput with the delay constraint, $\hat{\lambda}_{\text{K-Exp}}^{\text{max}}$, under different values of $K$ is plotted in Fig. 7. It can be clearly seen that despite the boost in the maximum stable throughput $\hat{\lambda}_{\text{K-Exp}}^{\text{max}}$, a further increase of cutoff phase $K$ may instead jeopardize the system performance. The highest maximum stable throughput with the delay constraint $\hat{\lambda}_{\text{K-Exp}}^{\text{max}}$, is achieved at around 0.34 when $K=19$. Note that with a more stringent delay constraint, $\hat{\lambda}_{\text{K-Exp}}^{\text{max}}$ may be even lower.

The above analysis is verified by simulation results presented in Table I. As we can see, the delay performance is not sensitive to the cutoff phase $K$ when the traffic rate is smaller than $\hat{\lambda}_0 \approx 0.3$. This is because the HOL packets are seldom pushed into deep phases when the traffic input rate is low. The maximum number of phase $K$ will not affect the delay performance as long as it is large enough to ensure a stable throughput.

With a high traffic rate, i.e., $\hat{\lambda} > \hat{\lambda}_0$, however, the mean queueing delay will rapidly increase with the cutoff phase $K$, and finally become infinite when $K=\infty$. As our analysis has revealed, although a larger cutoff phase $K$ leads to a higher maximum stable throughput, an excessively large $K$ will impair the queueing performance.

### III. Window-Based $K$-Exponential Backoff

The backoff model studied in Section II is slightly different from the contention-window-based backoff mechanism adopted in practical systems (such as Ethernet and IEEE 802.11 networks), where a HOL packet in phase $i$ that enters a collision, or jump back to phase 0 if it is successfully transmitted, $i=0,1,\ldots,K-1$. At each phase $i$, the HOL packet will randomly select a value from $\{0,\ldots, W_i - 1\}$, where $W_i$ is the contention window size of phase $i$. It then counts down at each time slot and retransmits when the counter is zero.

The sojourn time at any phase $i=0,1,\ldots,K-1$, $Y_i$, is a uniformly distributed random variable with state space $\{1,\ldots, W_i\}$. At phase $K$, the HOL packet will stay until it is successfully transmitted. The sojourn time at phase $K$, $Y_K$, is therefore a renewal process and the inter-renewal interval follows the uniform distribution with state space $\{1,\ldots, W_K\}$.

Note that the probability-based backoff model discussed in Section II can be also regarded as a special case of the above semi-Markov chain. The sojourn time at each phase $i$, $Y_i$, is a geometric distributed random variable with parameter $q^i$ when $i=0,1,\ldots,K-1$ and $pq^K$ when $i=K$.

### Table I. Mean Queueing Delay with $K$-Exponential Backoff

| $\hat{\lambda}$ | $K=10$ | $K=12$ | $K=20$ |
|-----------------|--------|--------|--------|
| 0.05            | 1.2    | 1.2    | 1.2    |
| 0.1             | 1.5    | 1.5    | 1.5    |
| 0.15            | 2.1    | 2.1    | 2.1    |
| 0.2             | 3.4    | 3.4    | 3.4    |
| 0.25            | 7.3    | 7.8    | 8.4    |
| 0.3             | $\infty$ | 39    | 185    |
| 0.35            | $\infty$ | $\infty$ | 5,562  |
The phase distribution of HOL packets is then given by
\[ \tilde{f}_i = \frac{f_i \cdot \tau_i}{\sum_{i=0}^{K} f_i \cdot \tau_i}, \quad i=0,1,\ldots, K. \] (41)

It can be seen from (39-41) that if the contention window size is set to be
\[ W_i = 2/q - 1, \quad i=0,1,\ldots, K, \] (42)
the phase distribution can be obtained as
\[ \tilde{f}_0 = \frac{q}{p + q - 1} \left( \frac{q}{p + q - 1} \right)^K, \] (43)
\[ \tilde{f}_i = \tilde{f}_0 \left( \frac{1}{1-p} \right)^i, \quad i=1,\ldots, K-1, \] (44)
\[ \tilde{f}_K = \tilde{f}_0 \left( \frac{1}{1-p} \right)^K / p, \] (45)
which is consistent with Eqs. (3-5) in Part I derived based on the probability-based backoff model. In fact, the phase distribution of HOL packets is determined by the first moment of the sojourn time at each phase. Although different backoff models have distinct distributions of sojourn time \( Y_i \), they follow the same phase distribution as long as the mean sojourn time \( E[Y_i] \) is equal in both cases.

The stability analysis of probability-based \( K \)-Exponential Backoff has been presented in Part I of the paper series. The counterparts of Theorems 1 and 2 in the window-based case are shown in Appendix IV for demonstration. It can be easily checked that all the analytical results presented in Part I hold true for window-based \( K \)-Exponential Backoff, if the contention window size is set as (42). Both backoff models share the same stable region if they have the same mean sojourn time at each phase.

Note that the contention window size \( W_i \) should be selected as an integer in practical systems, i.e.,
\[ W_i = \left\lceil \frac{2}{q'} - 1 \right\rceil, \quad i=0,1,\ldots, K. \] (46)

Here we ignore the difference of (42) and (46) for the sake of discussion. The simulation results will show that this approximation does not affect the conclusion.

### B. Delay Comparison of Window-based Backoff Model and Probability-based Backoff Model

Let \( X_i \) be the time spent from the beginning of phase \( i \) until service completes. According to Fig. 8, the probability generating function of \( X_i \) can be written as
\[
\begin{align*}
G_{x_i}(z) &= p G_{x_i}(z) + (1-p)G_{x_i}(z)G_{x_{i-1}}(z), \quad i = 0,1,\ldots, K-1 \\
G_{x_K}(z) &= G_{x_K}(z).
\end{align*}
\] (47)

where \( G_{x_i}(z) \) is the probability generating function of the sojourn time \( Y_i \). It can be obtained from (47) that
\[ G_{x_i}^{'}(1) = \sum_{j=0}^{K} (1-p)^j G_{x_i}^{'}(1) \] (48)
and
\[ G_{x_i}^{*}(1) = \sum_{j=0}^{K} (1-p)^j \left( G_{x_i}^{'}(1) + 2G_{x_i}^{'}(1)G_{x_{j+i}}(1) \right). \] (49)

The mean queueing delay \( E[T] \) can be then obtained by substituting (48-49) into (8).

The mean queueing delay is dependent on both the first and second moments of the sojourn time \( Y_i \). The probability generating function of the sojourn time \( Y_i \) in the window-based backoff model is given by
\[ G_{y_i}(z)_{\text{win}} = \frac{1}{W_i} \left( 1 - z^{-W_i} \right), \quad i = 0,1,\ldots, K-1 \] (50)
and
\[ G_{y_K}(z)_{\text{win}} = \frac{p}{1-(1-p)z^{-W_K}}. \] (51)

The first and second moments of \( Y_i \) can be then obtained as
\[
\begin{align*}
G_{y_i}^{'}(1)_{\text{win}} &= \frac{W_i+1}{2}, \quad i = 0,1,\ldots, K-1 \\
G_{y_K}^{'}(1)_{\text{win}} &= \frac{W_K+1}{2p}.
\end{align*}
\] (52)
and
\[
\begin{align*}
G_{y_i}^{*}(1)_{\text{win}} &= \frac{(W_i+1)(W_i-1)}{3}, \quad i = 0,1,\ldots, K-1 \\
G_{y_K}^{*}(1)_{\text{win}} &= \frac{(W_K+1)(W_K-1) + (1-p)(W_K+1)^2}{3p}. \quad i = 0,1,\ldots, K-1
\end{align*}
\] (53)

When the contention window size \( W_i \) is set according to (42), it can be obtained from (53) that the second moment of \( Y_i \), \( i=0,1,\ldots, K-1 \), in the window-based backoff model is given by
\[ G_{y_i}^{*}(1)_{\text{win}} = \frac{4}{3} - \frac{q'}{q'^2} < \frac{2(1-q''^2)}{q''^2} = G_{y_i}^{*}(1)_{\text{prob}} \] (54)
which is smaller than that in the probability-based model. Similarly, we have
\[ G_{y_i}^{*}(1)_{\text{win}} = \frac{2}{pq'^2} \left( 1 - \frac{2pq'^k}{3} \right) < \frac{2(1-pq'^k)}{(pq'^2)} = G_{y_i}^{*}(1)_{\text{prob}} \] (55)

Therefore, lower queueing delay can be expected in the window-based case. In the following, we will take the examples of Geometric Retransmission (\( K=1 \)) and Exponential Backoff (\( K=\infty \)) to demonstrate the above results.
1) Window-based Geometric Retransmission

The first and second moments of service time \( X \) of window-based Geometric Retransmission can be obtained by combining (48-49), (52-53), (42) and \( K=1 \):

\[
G'_{N_0}(1)_{w_1} = 1 + \frac{1 - p}{pq} \tag{56}
\]
and

\[
G^*_{N_0}(1)_{w_1} = \frac{2(1 - p)}{(pq)^2} \left(1 - \frac{p + pq}{3} \right) \tag{57}
\]

Note that the window-based Geometric Retransmission has the same stable region as the probability-based one. With \( q = q^*_{\text{Geo}} = 1/n \), or, equivalently,

\[
W_n = 1 \quad \text{and} \quad W^{\prime}_i = 2n - 1, \tag{58}
\]
the corresponding mean access delay and mean queueing delay are given by:

\[
E[X]_{W_{n-2n-1}}^{\text{Geo}} = 1 + n - \frac{1 - p}{pL}, \tag{59}
\]

\[
E[T]_{W_{n-2n-1}}^{\text{Geo}} = 1 + n \left( \frac{1}{pL} - \frac{1}{1 - \frac{1}{\lambda}} \right) \left( 1 - \frac{\lambda - 1}{3(n - \lambda)} \right) \tag{60}
\]

It can be seen from (59-60) and (20-21) that despite the same access delay, lower queueing delay will be incurred in the window-based backoff model, and the delay gap will be enlarged as the traffic input rate \( \hat{\lambda} \) or the number of nodes \( n \) increases. This can be clearly observed in Fig. 9. Both backoff models have similar delay performance when the traffic input rate \( \hat{\lambda} \) is low. The delay gap becomes noticeable with a large number of nodes \( n \) and a high input rate \( \hat{\lambda} \). Nevertheless, the difference is still quite small compared to the mean queueing delay \( E[T] \) incurred.

2) Window-based Exponential Backoff

The first and second moments of service time \( X \) of window-based Exponential Backoff can be obtained by combining (48-49), (52-53), (42) and \( K=\infty \):

\[
G'_{N_0}(1)_{w_1} = 1 + \frac{1 - p}{q - (1 - p)} \tag{61}
\]
and

\[
G^*_{N_0}(1)_{w_1} = \frac{2(1 - p)q}{q(1 - p)} \left[ \frac{2 + q}{3} \right] \tag{62}
\]

if \( q > \sqrt{1 - p} \). With \( q = q_{\text{Exp}} = 1 - e^{-\lambda} \), or, equivalently,

\[
W^{\prime}_i = 2(1 - e^{-\lambda})^{-1}, \quad i = 0, 1, \ldots, \tag{63}
\]

and \( \hat{\lambda} < \hat{\lambda}_{\rho} \approx 0.3 \), the mean access delay and mean queueing delay of window-based Exponential Backoff are given by

\[
E[X]_{W_{n-2n-1}}^{\text{Exp}} = \frac{1 - e^{-\lambda}}{pL - e^{-\lambda}} \tag{64}
\]

\[
E[T]_{W_{n-2n-1}}^{\text{Exp}} = \frac{1 - e^{-\lambda}}{pL - e^{-\lambda}} \left( 1 - e^{-\lambda} / 3 \right) \tag{65}
\]

As shown in Fig. 10, the delay gap between two backoff models is negligible with a traffic input rate \( \hat{\lambda} < \hat{\lambda}_{\rho} \approx 0.3 \). In contrast to Geometric Retransmission, the gap does not increase with the number of nodes \( n \) in the Exponential Backoff case, because the delay is not sensitive to the change of network population in both models.

When the traffic input rate \( \hat{\lambda} \) is close to \( \hat{\lambda}_{\rho} \approx 0.3 \), the delay performance rapidly deteriorates and the system becomes quasi-stable as \( \hat{\lambda} \) exceeds 0.3. The capture phenomenon will occur regardless of which backoff model is adopted.
It can be clearly seen from the above analysis that similar queueing performance is achieved in both backoff models. In fact, it has been shown in (54-55) that the difference of the second moment of sojourn time at phase \( i \) in the probability-based and the window-based models, i.e.,

\[ G''_p(\lambda) - G''_w(\lambda) \]

becomes significant only when the phase number \( i \) is large. With Geometric Retransmission, the maximum number of phases is \( K=1 \). In the case of Exponential Backoff, despite an infinite cutoff phase \( K \), the HOL packets usually stay at small phases when the traffic input rate \( \lambda \) is low. As a result, the delay gap between the two backoff models remains quite small all the while.

With \( K \)-Exponential backoff (\( 1<K<\infty \)), similar queueing delay performance should be achieved in both backoff models when the cutoff phase \( K \) is small. With a large \( K \) and a high traffic input rate \( \lambda \), however, some HOL packets will be pushed into deep phases and the difference of the second moment of sojourn time of the two backoff models \( G''_p(\lambda) - G''_w(\lambda) \) will grow exponentially with the phase number \( i \). Much lower queueing delay is therefore expected to be incurred by the window-based backoff model.

This can be clearly observed in Table II. Both the window-based and probability-based models share the same stable regions, and achieve almost the same queueing delay performance when the traffic input rate \( \lambda \) is low. With \( \lambda > 0.3 \), the delay gap becomes significant especially when the cutoff phase \( K \) is large.

### Table II. Mean Queueing Delay with Contention-Window-Based and Retransmission-Probability-Based K-Exponential Backoff

\( n=100, \quad q=1-e^{-1} \) and \( W_i = [2/(1-e^{(-\lambda)})]^{-1}, \quad i=0,1,...,K \).

| \( \lambda \) | \( K=10 \) | \( K=12 \) | \( K=20 \) |
|---|---|---|---|
| \( W_i \) | \( q' \) | \( W_i \) | \( q' \) | \( W_i \) | \( q' \) |
| 0.05 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| 0.1 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| 0.15 | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 |
| 0.2 | 2.9 | 3.4 | 2.9 | 3.4 | 2.9 | 3.4 |
| 0.25 | 5.7 | 7.3 | 5.9 | 7.8 | 6.2 | 8.4 |
| 0.3 | \( \infty \) | 25.6 | 39 | 75.5 | 185 |
| 0.35 | \( \infty \) | \( \infty \) | \( \infty \) | \( 4,520 \) | \( 5,562 \) |

### IV. Conclusions

The delay analysis of buffered Aloha networks with \( K \)-Exponential Backoff was presented in this paper. Geometric Retransmission (\( K=1 \)) and Exponential Backoff (\( K=\infty \)) are demonstrated as exemplary cases, and the performance comparison shows that although both of them can achieve a stable throughput of \( 0<\hat{\lambda} \leq e^{-1} \) with the retransmission factor \( q \) properly selected from the corresponding stable region, the delay performance is quite different. With Geometric Retransmission, finite mean queueing delay can be achieved within the whole traffic range \( 0<\hat{\lambda} \leq e^{-1} \), and both the mean access delay and the mean queueing delay linearly increase with the number of nodes \( n \). The delay performance of Exponential Backoff is not sensitive to the change of network population; nevertheless, it is severely penalized when traffic is heavy. With the aggregate input rate \( \hat{\lambda} > 0.3 \), the network will shift to the undesired stable point, and the variance of service time will grow unboundedly as time elapses, leading to infinite mean queueing delay. The cutoff phase \( K \) can be carefully tuned to achieve a good tradeoff between queueing performance and system robustness in a general \( K \)-Exponential backoff network.

The delay analysis was further extended to the contention-window-based backoff case. The performance comparison indicates that the retransmission-probability-based backoff model can serve as a good analytical model of the practical contention window mechanism, as it follows the same stable region and provides a tight delay upperbound.

### APPENDIX I. POLLACZEK-KINTCHINE FORMULA OF GEO/G/1

Let \( A \) represent the number of arrivals during service time \( X \). The probability generating function of \( A \) is then given by

\[
M_A(z) = \sum_{k=0}^{\infty} z^k \Pr \{ A = k \} = \sum_{n=1}^{\infty} \frac{z^n}{n!} \lambda^n (1-\lambda)^n P_n(z),
\]

where \( \lambda \) is the input rate.

Let \( C_n \) be the number of customers left behind by the departure of the \( n \)-th customer, and \( A_n \) be the number of customers arriving during the service of the \( n \)-th customer. The length of the buffer queue is given by the Lindley equation:

\[
C_n = C_{n-1} - I(C_n) + A_{n+1},
\]

where

\[
I(C_n) = \begin{cases} 0 & \text{if } C_n = 0 \\ 1 & \text{if } C_n > 0 \end{cases}
\]

is the indicator function. Let \( C = \lim_{n \to \infty} C_n \). We have

\[
M_C(z) = \mathbb{E}[z^{-I(C)-1}] = \left[ \left( 1 - \frac{1-z}{z} \right) P_0 + \frac{1}{z} M_z(z) \right] M_A(z).
\]

Dividing by \( z - M_A(z) \),

\[
M_C(z) = \frac{P_0(z-1)M_A(z)}{z-M_A(z)},
\]

where \( P_0 = 1-\rho \) and \( \rho \) is the offered load.

Let \( T \) denote the queueing delay of each packet. We have

\[
\Pr \{ C = k \} = \sum_{n=1}^{\infty} \Pr \{ A = k, T = n \} \Pr \{ T = n \}.
\]

The probability generating function of \( C \) can then be written as

\[
M_C(z) = \sum_{k=0}^{\infty} \frac{z^k}{k!} \lambda^k (1-\lambda)^k P_k(z) \Pr \{ T = n \}
= \sum_{n=1}^{\infty} \frac{z^n}{n!} \lambda^n (1-\lambda)^n z^n
= \sum_{n=1}^{\infty} \Pr \{ T = n \} (1-\lambda + \lambda z)^n = M_A(1-\lambda + \lambda z).
\]
Combining (66), (69) and (71), we have
\[ M_f(z) = \frac{p_0(z-1)M_f(z)}{(z-1)+\lambda[1-M_f(z)]}. \] (72)

From (72) finally we can obtain the mean queueing delay of a Geo/G/1 queue as
\[ E[T] = E[X] + \frac{\lambda[\text{var}[X] + E[X]^2 - E[X]]}{2(1-\lambda E[X])}. \] (73)

**APPENDIX II. q < \sqrt{1-p_A^{\text{Exp}}} FOR LARGE n**

It has been derived in Appendix II of Part I that the undesired stable point of Exponential Backoff is given by
\[ p_A^{\text{Exp}} = \frac{n(1-q)/q}{W_0((n/q-1)\exp(n/q))} = \frac{n(1-q)}{n+q\ln(1-q)}. \] (74)

For given \(0 < q < 1\), as long as the number of nodes \(n\) satisfies
\[ n > -(1+q)\ln(1-q), \] (75)

we have
\[ \sqrt{1-p_A^{\text{Exp}}} = \sqrt{\frac{n+\ln(1-q)}{n+q\ln(1-q)}} > q. \] (76)

**APPENDIX III. LOWER BOUND OF INSTANTANEOUS ATTEMPT RATE OF EXPONENTIAL BACKOFF WITH A LARGE NUMBER OF NODES n**

In an \(n\)-node buffered Exponential Backoff network, suppose that there are totally \(n_b\) backlogged HOL packets at time slot \(t\), with \(n_i\) packets in phase \(i\), \(i=1, 2, \ldots\). The attempt rate \(G_i\) is given by
\[ G_i = (n-n_b)\lambda + \sum_{j=0}^i n_j q^j \geq \sum_{j=0}^i n_j q^j, \] (77)
where
\[ \sum_{j=0}^i n_j = n_b. \] (78)

Let \(\phi\) denote the probability that there is a phase-i HOL packet in the node’s buffer, \(i=1, 2, \ldots\). We know from the Markov chain shown in Fig. 2 of Part I that
\[ \phi = \rho f_i, \quad i=1, 2, \ldots \] (79)
where \(\rho\) is the offered load of each queue and \(f_0, f_1, \ldots\) are the limiting probabilities, and we have
\[ f_0 \geq f_1 \geq f_2 \geq \cdots \] (80)

According to the law of large numbers, at any time slot \(t\), the number of backlogged HOL packets in phase \(i\), \(n_i\), satisfies
\[ \frac{n_i}{n} \rightarrow \phi \quad \text{for} \ n \rightarrow \infty, \] (81)
and
\[ \frac{n_b}{n} \rightarrow \rho(1-f_0) \quad \text{for} \ n \rightarrow \infty. \] (82)

By combining (79-81), we can conclude that for large \(n\), there is a high probability that
\[ n_i \geq n_j \geq n_k \geq \cdots \] (83)

As a result, the attempt rate \(G_i\) is lower bounded by
\[ G_i \geq \sum_{j=i}^n n_j q^j \geq \sum_{j=i}^n q^j \to \frac{q-q^i}{1-q} \quad \text{for} \ n \to \infty. \] (84)

With a large number of nodes \(n\), if the retransmission factor \(q\) satisfies (28), we can see from (84) that the attempt rate \(G_i\) will exceed \(-\ln p_S\) with high probability, and the probability of success \(p\) will converge to the undesired stable point \(p_d\).

**APPENDIX IV. PROOFS OF THEOREMS 1 AND 2**

We need the following lemma before presenting the proofs of Theorems 1 and 2.

**Lemma 1.** For buffered Aloha with window-based \(K\)-Exponential Backoff, the probability that a phase-i HOL packet has a transmission request, \(i=0, 1, \ldots, K\), is given by
\[ r_i = \frac{2}{W_i+1}. \] (85)

Proof: In the window-based backoff model, each HOL packet in phase \(i\) chooses a random value from \(\{0, \ldots, W_i\}\), \(i=1, \ldots, K\), with equal probability, and sends a transmission request when counting down to zero. Let \(J_i^*\) denote the residual time at time slot \(t\) of a phase-i HOL packet. The transition process of \(\{J_i^*\}\) can be described by the Markov chain shown in Fig. 11.

![State transition diagram of the residual time of a phase-i HOL packet.](image)

Fig. 11. State transition diagram of the residual time of a phase-i HOL packet.

The limiting probabilities of the above Markov chain can be obtained as
\[ \pi_i = \frac{2}{1+W_i} \frac{\text{W}_i-k}{\text{W}_i}, \quad k=0, 1, \ldots, \text{W}_i-1. \] (86)

The HOL packet will make a transmission request when the residual time becomes 0. As a result, the probability that a phase-i HOL packet has a transmission request, \(r_i\), is given by
\[ r_i = \pi_0 = \frac{2}{1+W_i}, \quad i=0, 1, \ldots, K. \quad \Box \]

According to (42) and Lemma 1, the request probability of a phase-i HOL packet is given by
\[ r_i = \frac{2}{W_i+1} = q^i, \quad i=0, 1, \ldots, K. \] (87)

The counterparts of Theorems 1 and 2 in Part I can be then presented in the following.

**Theorem 1.** For buffered Aloha with window-based \(K\)-exponential backoff \((1 \leq K \leq \infty)\), the probability of success \(p\) is given by
\[ p = \exp(-\lambda'/\rho), \] (88)
as the number of nodes \(n \to \infty\), where \(\lambda' = n\lambda\) is the aggregate input rate.

Proof: Each node in the network must be in one of the following states:
State 1: idle;
State 2: busy with a fresh HOL;
State 3: phase i and retransmitting, \(i=1,2,\ldots,K\);
State 4: phase i and not retransmitting, \(i=1,2,\ldots,K\).

The probability of a HOL packet being in phase i given that the node is busy is \(\tilde{f}_i\), where \(i=0,1,\ldots,K\). Suppose that the offered load per queue is \(\rho<1\). The probability of the above four states are given by:
1) \(\Pr\{\text{node is in State 1}\}=1-\rho\).
2) \(\Pr\{\text{node is in State 2}\}=\rho \tilde{f}_0\).
3) \(\Pr\{\text{node is in State 3}\}=\rho \tilde{f}_i r_i\), \(i=1,2,\ldots,K\);
4) \(\Pr\{\text{node is in State 4}\}=\rho \tilde{f}_i (1-r_i)\), \(i=1,2,\ldots,K\).

When a node successfully transmits a packet, its \(n\)-1 interfering nodes must be either in State 1 or State 4. The probability of success \(p\) in steady-state conditions can then be written as
\[
p = (\Pr\{\text{node is in State 1}\} + \sum_{i=2}^{K} \Pr\{\text{node is in State 4, phase } i\})^{-1} = (1 - \rho + \sum_{i=2}^{K} \rho \tilde{f}_i (1-r_i))^{-1}.
\]  
Note that the offered load \(\rho = \lambda / \tilde{f}_0\) as \(r_0=1\).

By combining (43-45), (87) and (89-90), we have
\[
p = (1 - \lambda / p)^{-1} \to \exp(-\lambda / p) \quad \text{for } n \to \infty.
\]

**Theorem 2.** For buffered Aloha with window-based K-Exponential Backoff (1\(\leq K\leq\infty\)), if \(q \leq -\lambda / \ln p\), then at any time slot \(t\), \(G_i \leq -\ln p_s\).

Proof: Suppose that there are totally \(n_b\) backlogged HOL packets at time slot \(t\), with \(n_i\) packets in phase \(i\), \(i=1,\ldots,K\). The attempt rate \(G_i\) is then given by
\[
G_i = (n-n_b)\lambda + \sum_{i=1}^{K} n_i r_i.
\]

According to (86), (91) can be written as
\[
G_i = (n-n_b)\lambda + \sum_{i=1}^{K} n_i q_i \leq (n-n_b)\lambda + n_i q'\quad (92)
\]
1) If retransmission factor \(q'\leq\lambda\), the attempt rate \(G_i\) is bounded by
\[
G_i \leq \lambda < -\ln p_s.
\]
2) If retransmission factor \(q'\geq\lambda\), the attempt rate \(G_i\) is bounded by
\[
G_i \leq n q' \leq -\ln p_s.
\]
Hence, the theorem is established by combining (93) and (94). □

**References**

[1] J. Y. Hui, *Switching and Traffic Theory for Integrated Broadband Networks*, Kluwer Academic Publishers, 1990.

[2] B. Tsybakov and W. Mikhaii, “Ergodicity of slotted Aloha system,” *Probl. Inform. Transmission*, vol. 15, no. 4, pp. 73-87, 1979.

[3] V. Anantharam, “The stability region of the finite-user slotted ALOHA protocol,” *IEEE Trans. Inf. Theory*, vol. 37, pp. 553-540, May 1991.

[4] W. Szpankowski, “Stability conditions for some multiqueue distributed systems: Buffered random access systems,” *Adv. Appl. Prob.*, vol. 26, pp. 498-515, 1994.

[5] W. Luo and A. Ephremides, “Stability of N interacting queues in random-access systems,” *IEEE Trans. Inf. Theory*, vol. 45, pp. 1579-1587, Jul. 1999.

[6] J. Luo and A. Ephremides, “On the throughput, capacity, and stability regions of random multiple access,” *IEEE Trans. Inf. Theory*, vol. 52, pp. 2593-2607, June 2006.

[7] S. Lam and L. Kleinrock, “Packet switching in a multi-access broadcast channel: Dynamic control procedures,” *IEEE Trans. Commun.*, vol. COM-23, pp. 891-904, 1975.

[8] B.-J. Kwak, N.-O. Song, L. E. Miller, “Performance analysis of exponential backoff,” *IEEE/ACM Trans. Networking*, pp. 343-355, April 2005.

[9] M. J. Ferguson, “An approximate analysis of delay for fixed and variable length packets in an unslotted ALOHA channel,” *IEEE Trans. Commun.*, vol. COM-25, No. 7, pp. 644-654, July 1977.

[10] S. S. Kamal and S. A. Mahmoud, “A study of users’ buffer variations in random access satellite channels,” *IEEE Trans. Commun.*, vol. COM-27, pp. 857-868, June 1979.

[11] K. K. Mittal and A. N. Venetsanopoulos, “Buffer analysis in random multiple access broadcast communications systems,” *Can. Elec. Eng. J.*, 1984.

[12] L. Georgiadis, L. F. Merakos and P. Papantoni-Kazakos, “A method for the delay analysis of random multiple-access algorithms whose delay process is regenerative,” *IEEE JSAC*, vol. SAC-5, no. 6, pp. 1051-1087, July 1987.

[13] T. K. Apostolopoulos and E. N. Protonotarios, “Queueing analysis of buffered slotted multiple access protocols,” *Comput. Commun.*, vol. 8, no. 1, pp. 9-21, 1985.

[14] T. Takine, Y. Takahashi and T. Hasegawa, “An approximate analysis of a buffered CSMA/CD,” *IEEE Trans. Commun.*, vol. 36, no. 8, pp. 932-941, Aug. 1988.

[15] A. Ganz and I. Chlamtac, “A linear solution to queueing analysis of synchronous finite buffer networks,” *IEEE Trans. Commun.*, vol. 38, no. 4, pp. 440-446, Apr. 1990.

[16] T. Wan and A. U. Sheikh, “Performance and stability analysis of buffered slotted ALOHA protocols using tagged user approach,” *IEEE Trans. Veh. Technol.*, vol. 49, no. 2, pp. 582-593, Mar. 2000.

[17] C. H. Foh, M. Zukerman and J. W. Tantra, “A Markovian framework for performance evaluation of IEEE 802.11,” *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, pp. 1276-1285, Apr. 2007.

[18] O. Tickoo and B. Sikdar, “Modeling queueing and channel access delay in unsaturated IEEE 802.11 random access MAC based wireless networks,” *IEEE/ACM Trans. Networking*, vol. 16, no. 4, pp. 878-891, Aug. 2008.

[19] A. Bianchi, “Performance analysis of the IEEE 802.11 distributed coordination function,” *IEEE J. Select. Areas Commun.*, vol. 18, no. 3, pp. 535-547, Mar. 2000.

[20] B. M. Metcalfe and D. R. Boggs, “Ethernet: distributed packet switching for local computer networks,” *Commun. ACM*, pp. 395-404, Jul. 1976.

[21] D. Aldous, “Ultimate instability of exponential back-off protocol for acknowledgement-based transmission control of random access communication channels,” *IEEE Trans. Inf. Theory*, pp 219-223, 1987.

[22] J. Goodman, A. G. Greenberg, N. Madras and P. March, “Stability of binary exponential backoff,” *J. ACM*, vol. 35, pp. 579-602, 1988.

[23] J. Hastad, T. Leighton and B. Rogoff, “Analysis of backoff protocols for multiple access channels,” *SIAM J. Comput.*, vol. 25, pp. 740-744, 1996.

[24] H. Al-Ammal, L. A. Goldberg and P. MacKenzie, “An improved stability bound for binary exponential backoff,” *Theory Comput. Syst.*, vol. 30, pp. 229-244, 2001.

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

[26] K. K. Ramakrishnan and H. Yang, “The Ethernet capture effect: analysis and solution,” *Proc. 19th Local Computer Networks Conf.*, Oct. 1994.

[27] K. Medepalli and F. A. Tobagi, “On optimization of CSMA/CA based wireless LANs: Part I – impact of Exponential Backoff,” *IEEE Globecom*, 2006.