Optimal controlled ALOHA for two-way data communication in a cable television network

Vleuten, van der, R.J.; van Etten, W.C.; van den Boom, H.P.A.

Published in:
IEEE Transactions on Communications

DOI:
10.1109/26.297854

Published: 01/01/1994

Document Version
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the author's version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Citation for published version (APA):
Optimal Controlled ALOHA for Two-Way Data Communication in a Cable Television Network

René J. van der Vleuten, Student Member, IEEE, Wim C. van Etten, Senior Member, IEEE, and Henricus P.A. van den Boom

Abstract—Based on a unified analysis of both pure and slotted ALOHA systems, a new control algorithm for ALOHA systems is proposed. A feature of great practical importance is the algorithm’s automatic adjustment to changes in average traffic intensity or the (finite or infinite) number of active stations in the system. In addition, the algorithm has a low-complexity implementation.

Computer simulations, concentrating on the use for two-way data communication in a cable television network, have demonstrated that the practical performance of the algorithm closely approximates the theoretical optimum, even under extremely heavy traffic load conditions. Furthermore, dynamic performance simulations have shown that the algorithm assures swift recovery from overload situations.

I. INTRODUCTION

The ALOHA protocol (see, e.g., [1], [2]) can be applied in a system in which all stations have access to a single common transmission channel. Using the channel, all stations send their information to a central point (the center). In pure ALOHA, each station in the network is allowed to send a packet at an arbitrary moment. In slotted ALOHA, time is divided into slots (the size of which is slightly larger than the packet size, taking into account the differences in propagation times between the stations and the center) and a station wanting to send a packet has to wait for the beginning of the next slot. The center, using a separate channel designated for this purpose, sends an acknowledgment of each correctly received packet to the station that sent that packet. All stations operate independently of each other, so several stations can send a packet at the same time, causing collisions and destroying the colliding packets. A station knows its packet has suffered from a collision if it does not receive an acknowledgment within a certain time, the roundtrip time. In that case, it sends the packet once again, after having waited a random time (in order to prevent packets from the same senders to collide over and over again). A generally accepted retransmission strategy is to uniformly distribute the random waiting time until retransmission, between 0 and L, minimizing the number of repeated collisions [3]. While a station has a packet for retransmission it does not generate new packets.

The fact that transmission errors could occur on the channel will be neglected, assuming that the bit error rate on the channel is so low that transmission errors are negligible compared to those caused by collisions.

In order to improve the performance of the system and to guarantee its stability, a control algorithm is applied. The control is located at the center and broadcasts the value for the maximum waiting time until retransmission, L, that is to be used by all stations.

In Section II, the general principle of control algorithms for the ALOHA system will be discussed and the theoretical foundation for a new and optimal control algorithm will be laid. Section III describes the control algorithm and its implementation and in Section IV the results of performance simulations of the algorithm are reported. Finally, Section V contains the conclusions.

II. THE THROUGHPUT AND STABILITY OF AN ALOHA SYSTEM

The following assumptions are made.

- All packets are of equal length, maximizing the throughput (see, e.g., [4]).
- The number of stations is infinitely large.
- All stations together generate new packets forming a Poisson process with an average of S packets per packet time (slot).
- Apart from new packets, the stations also generate retransmissions of previously collided packets. It is assumed that the total of retransmitted and new packets also forms a Poisson process, with an average of G packets per packet time. This assumption is a good approximation, as has been confirmed by experiments as well as theory [5].

Under these assumptions the following well-known equations for the throughput, S, of the system in the equilibrium state may be derived (see, e.g., [1]):

\[
S = Ge^{-2G},
\]

(1)

for pure ALOHA, and

\[
S = Ge^{-G},
\]

(2)

for slotted ALOHA. The maximum throughput, equal to \(1/(2e) \approx 0.184\) for pure ALOHA and \(1/e \approx 0.368\) for slotted ALOHA, is achieved at \(G = 1/2\) or \(G = 1\), respectively.

0090-6778/94$04.00 © 1994 IEEE
Based on (1) or (2) only, it is impossible to design a control algorithm. Furthermore, a real system comprises only a finite number of stations. In most cases, this number is so large that it may be presumed to be infinite. For (temporarily) heavy channel traffic, however, that approximation may not be sufficiently accurate. For these reasons, the second assumption will be dropped in the following (while maintaining the others), i.e. the number of stations will be assumed to be finite.

Suppose there are \( N \) stations, each generating new packets according to a Poisson process with an average of \( p \) packets per packet time. If a packet collides it will be retransmitted after an average waiting time of \( \delta \) packet times, where \( \delta \) is the sum of the roundtrip time, \( R \) (the fixed time that elapses between the successful transmission of a packet and the reception of the acknowledgment for that packet), and the average waiting time until retransmission (which equals \( L/2 \) since the distribution is uniform between 0 and \( L \)). Thus,

\[
\delta = R + \frac{L}{2}, \tag{3}
\]

where \( R \) and \( L \) are normalized with respect to the packet time. The collided packet will generate an average retransmission traffic of \( 1/6 \) packets per packet time.

By assuming that \( n \) of the \( N \) stations have a packet for retransmission the following expression can be stated:

\[
G = (N - n)p + \frac{n}{\delta}, \tag{4}
\]

where the channel traffic, \( G \), is the sum of the stream of newly generated packets, \( (N - n)p \), and the stream of retransmitted packets, \( n/\delta \). In equilibrium, the incoming and outgoing traffic of the system have to be equal, so

\[
S = (N - n)p. \tag{5}
\]

For pure ALOHA, the equation for the throughput is obtained by subsequently substituting (5) into (4), giving

\[
G = S + \frac{n}{\delta}, \tag{6}
\]

and (6) into (1), resulting in

\[
S = (S + \frac{n}{\delta})e^{-2(S + \frac{n}{\delta})}. \tag{7}
\]

For slotted ALOHA, a similar expression is obtained by substituting (6) into (2). Equation (7) can be solved numerically using the Newton-Raphson algorithm. The throughput in the equilibrium state can be calculated—for each value of \( N \), \( p \), \( R \), and \( L \)—by determining the intersection between the throughput-backlog curve given by (7) and the load line, which is given by (5). Then, \( n \), the average number of stations having a packet for retransmission, is known too. In the following, \( n \) will be referred to as the backlog.

It is well known that an ALOHA system is inherently unstable, i.e. in the long run the system will end up in a state of low throughput and high delay. In case of a finite number of stations, the system can be made stable by choosing \( L \) appropriately large [1], [6], [7], [8]. The system, therefore, has to be adapted to the worst case, meaning that under normal conditions the delay is much larger than required. It would be desirable that the average delay be small if the backlog is small. If the backlog increases the average delay has to increase as well, in order to guarantee the stability of the system. In other words, \( L \) is adjusted to the momentary intensity of the channel traffic. This is the principle of controlled ALOHA.

From literature many control methods are known [1], [6], [9]-[13]. All methods are based on the same principle: it is assumed that the momentary value of the backlog is known (possibly obtained by estimation) and from that value a value for \( L \) is chosen to maximize the momentary throughput. The accuracy with which \( L \) is adjusted varies from roughly to “optimal.”

At this point, a comment on optimal control is appropriate. Strictly speaking, the algorithms that were mentioned in the previous paragraph are not optimal [14]. An optimal control algorithm has to take into account the whole past of the channel in order to be able to estimate the momentary value of the backlog as accurately as possible [15]. In practice, only the recent history of the channel is relevant. Furthermore, for optimal control \( L \) has to be adjusted almost continuously, choosing \( L \) such that it does not maximize the momentary throughput but the throughput during the interval until its next adjustment. The improvement that is obtained hereby is negligible in practical situations. In conclusion, it is very difficult, and not practical, to realize optimal momentary control.

III. THE CONTROL ALGORITHM

The important performance measures for the system are throughput and stability, which have already been discussed, and average packet delay, which will be dealt with here.

The delay of a packet is defined here as the time that elapses between the first transmission of the packet and the (successful) last (re-)transmission of that packet, i.e., a packet that is successfully transmitted the first time has a delay of 0. A packet that collides when it is transmitted for the first time, but is successfully transmitted the second time, has an average delay of \( \delta \) packet times, where \( \delta \) is defined by (3). The average delay of all packets, \( D \), is the product of \( \delta \) and the average number of retransmissions per packet. Since the average number of transmissions per packet is equal to \( G/S \), the average number of retransmissions per packet equals \( G/S - 1 \). So, for the average delay the following, very simple, formula can be derived (using (6)):

\[
D = \left( \frac{G}{S} - 1 \right) \delta = \frac{n}{S}. \tag{8}
\]

The strategy of the control algorithm is to maximize the momentary throughput, minimizing the average delay. If there is a backlog, it will be reduced as fast as possible (within the limits of the ALOHA protocol). For pure ALOHA, the condition for maximum throughput, \( S = 1/(2e) \), is \( G = 1/2 \). Substituting these values into (6) and rewriting this equation yields

\[
\frac{n}{\delta} = e - 1. \tag{9}
\]
Combining (9) with (3) shows that the throughput is maximized by taking
\[ L = \frac{4ne}{e - 1} - 2R. \]  
(10)

For slotted ALOHA, the equivalent equations of (9) and (10) are
\[ n = \frac{e - 1}{\delta}, \]  
(11)
and
\[ L = \frac{2ne}{e - 1} - 2R. \]  
(12)

So, the control algorithm can maximize the throughput by satisfying (10) for pure ALOHA or (12) for slotted ALOHA. As an example (for pure ALOHA) Fig. 1 shows how the adjustment of \( L \) changes the throughput-backlog curve. For this example the system was assumed to have been in the state corresponding to the middle intersection. It can be observed that after the adjustment of \( L \) the throughput reaches its maximum value and the system has become stable again (there is only a single intersection between the curve and the load line).

In order to be able to set \( L \) to its optimal value, the control algorithm has to know the momentary value of the backlog, \( n \). This value can be estimated, assuming the system is in equilibrium. Rewriting (5) gives \( n = N - S/p \) (\( N \) and \( p \) are constants and \( S \) can be measured by the center), but this estimate is inaccurate since it does not take into account the possible fluctuations in \( p \) (or in \( N \)); in practice, \( p \) will not be constant and the number of active stations may change, too. A good estimate can be made on the basis of (6). Rewriting gives \( n = (G - S)\delta \). The estimate adapts itself to changing values of \( p \) or \( N \), but it does require knowledge of \( G \).

For pure ALOHA, there are two ways to estimate \( G \). If the channel traffic is a Poisson process with traffic density \( G \) then from probability theory it is known that the time between two subsequent packet transmissions is exponentially distributed with parameter \( G \). Therefore, the average time between two subsequent generations equals \( 1/G \). Because the exponential distribution has the memoryless property, the average time between the end of a transmission and the beginning of the next transmission also equals \( 1/G \). Thus, \( G \) can be computed by measuring the average length of the periods during which the channel is idle. A simpler method (which is also applicable to slotted ALOHA) is to measure the total time during which the channel is idle. Because of the Poisson distribution, the probability of the channel being idle during one packet time (one slot) equals \( \exp(-G) \), implying that the fraction of time during which the channel is idle also equals \( \exp(-G) \).

The optimal adjustment of \( L \) (which automatically adapts to changes in \( N \) or \( p \)) is a new feature of the proposed algorithm. Previously reported algorithms inaccurately adjust \( L \) or are not adaptive, guaranteeing system stability only for a finite number of stations. As a comparison with other control algorithms, combining (3), (6), and (10) results in the following update equation for \( L \) (for pure ALOHA):
\[ L := \frac{2e}{e - 1}(G - S)L + 2R\left(\frac{2e}{e - 1}(G - S) - 1 \right). \]  
(13)

When \( R \) is small compared to \( L/2 \), (13) resembles the update equation of an exponential-backoff control algorithm (see, e.g., [13]), with an optimally adjusted backoff coefficient: \( 2e(G - S)/(e - 1) \).

In summary, the proposed control algorithm continuously executes the following steps:
1) During a certain time, called the control interval, the throughput, \( S \), and the fraction of time during which the channel is idle are measured by the center.
2) The channel traffic, \( G \), is estimated as the negated natural logarithm of the measured fraction of time during which the channel is idle. Based on \( S \) and the estimated \( G \), the backlog, \( n \), is estimated using (6).
3) Based on the estimated \( n \), \( L \) is made equal to its optimal value (given by (10) or (12)), unless \( L \) would become smaller than a certain minimum value—in which case it is set to this minimum value—or larger than a certain maximum value—in which case it is set to this maximum value.

The use of a minimum and maximum value for \( L \) will now be motivated. If both backlog and channel traffic are low, the function of the minimum value for \( L \) is to prevent the control algorithm from increasing the backlog and generating useless retransmissions in order to try to make \( G \) equal to 1/2 or 1. The use of a minimum value for \( L \) implies that, under normal operation, the average delay will be slightly larger than required. On the other hand, it provides a buffer against fluctuations of the intensity of the offered traffic. The minimum value for \( L \) depends on two factors: the average intensity of the channel traffic and the size of the control interval. As the traffic intensity and the size of the control interval increase, the minimum value for \( L \) has to increase, too. A suitable value may be determined in practice, or, in case one wants to impose strict restraints, a limit can be put on the probability of the backlog increasing to a certain value within a single control interval.

The maximum value for \( L \) prevents it from "exploding" in case of a large pulse in the traffic being offered to the system.
The explosion is caused by the fact that in the aforementioned case there is a large discrepancy between the estimated value of the backlog and its real value. The backlog is estimated incorrectly because the system is not in equilibrium. A suitable maximum value for $L$ is found by substituting $n = N$ into (10) or (12).

IV. PERFORMANCE SIMULATIONS

As a practical example, in this section the application of the ALOHA protocol for two-way data communication in a cable television network will be discussed. Most modern cable television networks have a hierarchical structure (which can be represented by a tree), as shown in Fig. 2. In general, the networks are used for the distribution of radio and television programs from the head end (the highest level in the tree) to the individual subscriber (the lowest level in the tree). The last starpoint before the subscriber is called a mini starpoint. For reasons of privacy and security, each subscriber has an individual connection to one of the mini starpoints.

In general, cable networks are very expensive. For economic reasons it is therefore interesting to look for a more intensive and varied use of those networks. Enabling two-way data communication between the subscriber and the head end offers prospects for new services like, e.g., pay TV, tele-shopping, alarm and guard facilities, or interactive videotex [16]. The services offered, as well as the system architecture, could be similar to those of the INDAX system [17].

Because of the nature of the channel traffic (short bursts at random time intervals) and the structure of the network (many stations connected to the common channel leading to the head end), the ALOHA protocol is appropriate. Since the head end uses a separate contention-free channel for sending the acknowledgments of correctly received packets as well as the value for $L$, the control does not suffer from network congestion. In our system design the mini starpoints, each connected to approximately 20 subscribers, are the actual stations in the ALOHA system.

The main parameters of the system used in the simulations are listed in Table I. The number of stations (mini starpoints) corresponds with $5000 \cdot 20 = 100,000$ cable television network subscribers. From Table I the roundtrip propagation time can be computed as $20 \text{ km}/v_c \approx 1 \cdot 10^{-4} \text{ s}$ (the velocity of electromagnetic waves in a coaxial cable, $v_c$, equals about $2 \cdot 10^8 \text{ m/s}$). Since a packet time equals about $6 \cdot 10^{-4} \text{ s}$, the roundtrip propagation time equals approximately 0.2 packet times. Taking a packet processing time at the head end of about 1 packet time, the total roundtrip time $R \approx 1.2$ packet times. So, in the cable television network for which the ALOHA protocol is intended the roundtrip time, $R$, is negligible compared to the average waiting time until retransmission, $L/2$. Therefore, $R$ was set to 0 in the simulations. Since the roundtrip propagation time is less than 1 packet time, in this case the capacity of the system can be increased (by a factor of 2/1.2) by using slotted ALOHA instead of pure ALOHA. However, for a cable television network it may be more cost-effective to increase capacity by increasing the data transmission rate since the available bandwidth is not the limiting factor.

In order to verify the theoretical results, simulations were performed for various traffic intensities. In addition, as a test for stability, the dynamic system performance has been measured.

A. The Performance for Various Traffic Intensities

For pure ALOHA, simulations were performed for three traffic intensities $N_p$: normal traffic ($N_p = 0.02$), heavy traffic ($N_p = 0.1$), and very heavy traffic ($N_p = 0.175$). Since for $N$ the practical value of 5000 was used, the values for $p$ were $4 \cdot 10^{-6}$, $2 \cdot 10^{-5}$, and $3.5 \cdot 10^{-5}$, respectively. For the minimum value of $L$ an arbitrary value of 100 was used. The control interval (the time between two subsequent adjustments of $L$) was set to 1500 packet times, corresponding with a control interval in the order of one second in practice since there are approximately 1700 packets/s (see Table I). Without control the system would be unstable in all three cases; especially for $N_p = 0.175$ the situation is very critical. In Table II, the theoretical values (for $L = 100$) and simulation results for $G$, $S$, $n$, and $D$ are listed. For $N_p = 0.02$ and 0.1 the results agree very well with the theoretically expected values.

For $N_p = 0.175$ two columns are given, the first of which was measured with the normal control interval of 1500. It can be seen that the system reacts too slowly to statistical fluctuations of the intensity of the offered traffic: the backlog rises to a high value and the control algorithm has to "use the safety-brake" to render the system stable again. The average value for $L$ was 526 [from (6)]. The results in the second column were measured with a control interval of 150. The algorithm reacts faster, leading to a lower average backlog and a smaller average delay. The average value for $L$ was 131, in this case. Equations (6) and (8) remain valid in both columns; (1), however, is not valid. The conclusion is that for very heavy traffic a smaller control interval should be used.
than for normal traffic. The simulations then agree well with the theoretical model.

For slotted ALOHA, simulations were performed for heavy traffic ($N_p = 0.2$), very heavy traffic ($N_p = 0.35$), and extremely heavy traffic ($N_p = 0.3675$), testing the system to the limit. As a minimum value for $L$, again 100 was used. The theoretical and measured values of $G$, $S$, $n$, and $D$ are listed in Table III. Simulations were performed with control intervals of 1500 and 150 packet times. The measured values agree very well with the theoretical ones for $N_p = 0.2$ and $N_p = 0.35$ (for a control interval of 150). For $N_p = 0.3675$ the agreement is quite good, especially when considering the fact that the system operates at 99.9% of its theoretically maximal throughput.

For slotted ALOHA, too, the conclusion is that for very heavy traffic a smaller control interval should be used than for normal traffic.

Because of the similarity of the new control algorithm to an exponential-backoff control algorithm, as expressed by (13), we also performed simulations with the binary exponential-backoff algorithm (used in Ethernet) (see, e.g., [1]). For this algorithm, each station uses its own, local, value for $L$. The first time a new packet is transmitted, a station uses a certain minimal value for $L$. If its packet collides on the first transmission, the station doubles its value for $L$ at each successive retransmission, until the packet is transmitted successfully. Table IV lists the measured values of $G$, $S$, $n$, and $D$ for a slotted system, using a minimal value of 50 for $L$ (which resulted in a higher performance than a minimal value of 100). As can be seen, except for $N_p = 0.2$ where our choice of a minimal value for $L$ of 100 for the ALOHA system is rather conservative, the new control algorithm clearly outperforms the binary exponential-backoff algorithm. This result is not surprising, since our algorithm uses an optimized backoff coefficient whereas the binary exponential-backoff algorithm is not adaptive.

B. The Dynamic System Performance

Pulse response simulations were only carried out for pure ALOHA since they were intended to test stability rather than to produce quantitative results.
In [18] the use of traffic overload pulse response was suggested as a tool for the measurement of dynamic system performance. It was found that the backlog fall time, defined as the duration of the 90% to 10% drop in the ensemble average backlog waveform during the decreasing region of the applied traffic pulse, provides a compact summary of transient behaviour during temporary traffic overload. A related measure would be the backlog fall speed, expressed in stations per packet time and defined here as the quotient of the magnitude of the 90 to 10% drop and the backlog fall time. This normalized measure facilitates a comparison between different systems with different numbers of stations.

Simulations were performed for two different input traffic pulse shapes: exponential and uniform. The pulses are formed as the sum of the packets generated by the individual stations, which each send a packet after waiting for an average of 7500 packet times (approximately 5 s) from the beginning of the pulse. For both simulations, the value for \( N_p \) under normal conditions was set to \( 4 \cdot 10^{-6} \) (\( N_p = 0.02 \)), the minimum value for \( L \) was set to 100, the maximum value to 32 000, and the control interval was 150 packet times. The exponential input pulse and the resulting backlog waveform are shown in Fig. 3(a) and Fig. 3(b), respectively. The backlog fall speed in this case equals approximately 0.11 stations per packet time. The uniform pulse and its corresponding backlog wave-
René J. van der Vleuten was born in Valkenswaard, the Netherlands, in July 1965. In August 1989 he received the M.Sc. degree in electrical engineering from Eindhoven University of Technology, the Netherlands, and in November 1991 he received the Degree of Chartered Designer in the field of Information Technology from the Department of Electrical Engineering at Delft University of Technology, the Netherlands. He received the Ph.D. degree in electrical engineering from Delft University of Technology, in March 1994.

From October 1988 to July 1989, he was a guest at Philips Research Laboratories, Eindhoven, and in July and August 1988 he was a guest of the Technical University of Poznań, Poland. His interests include communication theory and image coding.

Dr. van der Vleuten has been awarded the second prize in the 1990 IEEE Region 8 Student Paper Competition. He was nominated for the 1989 Professor Bailer prize, awarded annually to a telecommunications graduate by the Royal Institute of Engineers in the Netherlands.

W. van Etten (M’80-SM’91) was born in Zevenbergen, The Netherlands, in 1942. He received the M.Sc. and Ph.D. degrees in electrical engineering from Eindhoven University of Technology, Eindhoven, The Netherlands, in 1969 and 1976, respectively.

From 1969 to 1970 he was employed by Philips Electronics, developing circuits for oscilloscopes. In 1970 he was appointed an assistant Professor at Eindhoven University of Technology, Faculty of Electrical Engineering. From 1970 to 1976 he was engaged in research on the transmission of digital signals via coaxial and multiwire cables. Since 1976 he has been involved with research and education on optical fiber communications. Presently, he is an Associate Professor at the Eindhoven University of Technology. He is author or coauthor of thirty papers in international journals. Moreover, he is coauthor of the book Fundamentals of Optical Fiber Communications (Englewood Cliffs, NJ: Prentice Hall), 1991.

Henricus P.A. van den Boom was born in Eindhoven, The Netherlands, in 1955. He received the M.Sc. degree in electrical engineering from Eindhoven University of Technology, Eindhoven, The Netherlands, in 1984.

Since then he has been an Assistant Professor at the Telecommunications Section of the Faculty of Electrical Engineering of the same university. He lectures in electrical transmission line theory and optical fiber communication systems. He has been involved in research on two-way CATV networks. His present research includes study of coherent optical communication systems.