On a Self-Adjusting Capability of Random Access Networks

LEONARD KLEINROCK, FELLOW, IEEE, AND GIDEON Y. AKAVIA, MEMBER, IEEE

Abstract—**We consider a distributed communication network with many terminals which are distributed in space and wish to communicate with each other using a common radio channel. Choosing the transmission range in such a network involves the following tradeoff: a long range enables messages to reach their destinations in a few hops, but increases the amount of traffic competing for the channel at every point.**

We give a simple model for the per-hop delay in random access networks, analyze this tradeoff, and give the optimal transmission range. When choosing this optimal range, as a function of specified traffic and delay parameters, networks demonstrate an important self-adjusting capability. This capability to adjust to traffic makes heavily loaded *networks* far better than *centralized systems* (in which all messages must reach one common destination).

Dividing a terminal population into *power groups* can improve any random access system, especially when the traffic is split between groups in an appropriate way, which we demonstrate. But since networks are hurt by destructive interference less than centralized systems, it is harder to improve them. Using power groups can significantly improve centralized systems, but will lead to a smaller relative improvement in networks. Decomposing the system into a *hierarchy* of ALOHA levels, with only a small population contending at the top level, can improve centralized systems but does not improve networks.

I. Introduction

CONSIDER a large number of terminals, physically distributed over a large geographic region. If all terminals wish to communicate with one destination, we shall call the system *centralized* and the common destination the *station*. Assuming the communication resource available is a radio channel of a given bandwidth, how should this common channel be shared among the terminals? If the terminals were colocated in the same place, the best way to use the channel is to form a queue of busy terminals (i.e., those having anything to transmit) and to let them use the full bandwidth available one after the other. Forming one queue is much better than giving each terminal a fraction of the bandwidth, and letting each terminal queue its own messages [7].

It is no trivial matter to have all terminals form one queue when the terminals are numerous and distributed over large distances. Of special interest, then, is the ALOHA approach, which invests no resources in coordination and control of terminals. When using the (unslotted) ALOHA scheme, each terminal transmits whenever it has a message ready. If more than one terminal is transmitting at the same time, a conflict will occur in the use of the radio channel, and we shall assume at first that all messages involved in such a collision will be destroyed. When the destruction of its message becomes known to the terminal it will, after a somewhat randomized delay, retransmit the message. We shall not specify how the failure of its message becomes known to the terminal, but assume that this knowledge is free.

Schemes based on the ALOHA idea have been extensively treated [1], [9], [12]. ALOHA is obviously good when the system is lightly utilized and destructive interference is not very likely. When the load is heavy, a significant fraction of the transmissions will fail as a result of collisions. The wasteful effect of collisions can be reduced if all transmissions are of the same length [5]. This is usually achieved by breaking long messages into packets of a fixed maximum size. We assume that this is always done and, despite the fact that one message may result in several packets, we assume that arrival of separate packets into our system is independent, and that the total arrival process is Poisson. The wasteful effect of collisions can be further reduced if time is slotted (where each slot has a duration which is equal to a packet transmission time) and if terminals are constrained to start transmitting only at the beginning of a slot. The resulting access scheme is called slotted ALOHA, and the maximum fraction of the time slots it can use for successful transmissions is known to be $1/e$ [16].

Let us choose the data unit so that the average length of a message is equal to 1. This is simply a convenient normalization, which is equivalent to measuring the capacity of the communication channel in messages (of an average length) per second, instead of measuring in bits per second. The throughput-delay performance of the ALOHA schemes is not described by a simple analytic expression [12]. For simplicity we shall use the following ad hoc expression to describe the performance of the ALOHA schemes:

$$T = \frac{1}{C - eS}. \tag{1}$$

Here T is the average response time of the system, C is the capacity (bandwidth) of the communication channel, and S is the system throughput (messages per slot). We shall assume that this expression describes the optimum envelope of slotted ALOHA and unslotted ALOHA performance curves. (For $S \to 0$ it describes unslotted ALOHA; for $S/C \to 1/e$ it describes slotted ALOHA.) Equation (1) is a simple two-parameter approximation that reproduces the known behavior when $S = 0$ and when $S/C = 1/e$. For a similar three-parameter approximation see [10].

Assume that the throughput S and the acceptable delay T are specified, and that we seek an access scheme that will minimize the necessary system capacity C. For most purposes it is sufficient to specify the communication needs by the dimensionless product ST, whose inverse we shall call burstiness [2], [6], [13]. We shall define the *quality* [2] of an arbitrary access scheme as the inverse ratio between the capacity necessary when using this scheme and the capacity necessary when using the best possible scheme, in which messages form one queue and share one channel.

Paper approved by the Editor for Computer Communications of the IEEE Communications Society for publication without oral presentation. Manuscript received August 31, 1979; revised November 9, 1982. This work was supported in part by the Advanced Research Project Agency of the Department of Defense under Contract MDA 903-77-C-0272.

L. Kleinrock is with the Department of Computer Science, University of California, Los Angeles, CA 90024.

G. Y. Akavia is with the Ministry of Defense, Haifa, Israel.

0090-6778/84/0100-0040$01.00

Inverting (1) we see that the capacity necessary when using ALOHA is given by $C = eS + 1/T$. When messages arrive independently and their lengths are exponentially distributed, the best scheme is the $M/M/1$ queue, where the necessary capacity is $C = S + 1/T$. The quality of the ALOHA scheme is therefore simply $(ST + 1)/(eST + 1)$. We see that the ALOHA scheme has a quality of 1 when the traffic is very bursty ($ST \ll 1$), i.e., it then needs no more capacity than the $M/M/1$ scheme, and a quality $1/e$ when the traffic is very steady ($ST \gg 1$).

In the centralized system described above, all messages have one common destination, even though their sources are distributed. When the traffic to be carried is between many terminal pairs we have a different problem, which we shall call the *network* problem. That is, in a network, both the sources of messages and their destinations are distributed. In describing the centralized system we have implicitly assumed that all terminals can transmit with enough range to reach the station (i.e., we are not power limited), and that transmitting directly to the station is the best policy. If the transmission range is not enough to span the distance from source to destination, the message will have to be received by some intermediate node and relayed towards its destination. That is, a message may need more than one hop in order to reach its destination. The intermediate node is often called a *repeater*.

We have assumed that the centralized system is a one-hop system, but we shall explicitly treat the question of transmission range in networks, since it introduces an important tradeoff: a short transmission range makes more hops necessary, but reduces the interfering traffic. We shall see that choosing an appropriate range, as a function of traffic characteristics, will lead to the self-adjusting capability referred to in our title.

In Section II we give a model for the per-hop delay in networks, assuming we have a model for the delay in centralized systems and that we can calculate the total contending traffic at any point. In Section III we use this one-hop delay model to analyze the choice of transmissison range, and demonstrate the self-adjusting capability of random access networks. In Section IV we introduce two ideas that help random access centralized systems when they are really bad, i.e., when they are very steady. These ideas contribute much less to random access networks, because when these can adjust they will rarely be very steady. General conclusions are given in Section V.

II. One-Hop Delay in Networks

Explicit and simple models for delay in random access communication systems are rare. Even if we had such models for centralized systems, they are not directly applicable to networks. In this section we present a model for the one-hop delay in networks. We assume that the transmission policy of all terminals is chosen to optimize the overall network performance. We also assume that our network covers a region of space that is large enough to make edge effects negligible, that terminals are placed everywhere with the same density, and that the terminal density is very high, so we may make all calculations as if we had a continuum of terminals. Other assumptions we adopt are as follows.

1) The rate of traffic exchanged between any two small geographic areas depends only on the size of the areas and the distance between them. The rate does not depend on the identity (i.e., location) of the areas or the direction from one to the other. That is, our network is homogeneous and isotropic in its statistical properties.

2) The terminal's antenna is simple, and the signal propagates equally in all directions.

3) A transmission will not be bothered by other transmissions that are not within range of its (possibly intermediate) destination, but will be destroyed by any simultaneous transmission that takes place within range of its destination. A transmission will be successful whenever it is the only one within range of its destination. That is, we assume a definite range, beyond which no interference is felt. This is, of course, an abstraction of the real world, in which both successful reception and destructive interference are probabilistic events.

Consider, for example, a network using slotted ALOHA. For simplicity we shall ignore the fact that the synchronization necessary for slotted ALOHA is hard to achieve in a network with long-range transmissions and partially overlapping ranges. Consider a given terminal with a rate of s messages per slot destined to another terminal. A transmission will be successful only if there is no other transmission with enough range to interfere with it. Our terminal will, therefore, have to offer a total traffic of g messages per slot in order to succeed at a rate s, where g includes retransmissions of previously unsuccessful messages. Let G be the total offered traffic per slot heard at the destination. Assume that G is created by an infinite population of terminals, and that the amount contributed to it by every source–destination pair is a Bernoulli process independent of the traffic offered by any other source–destination pair. Returning to our given terminal, whose contribution to G is minute, we must have $s = ge^{-G}$, where e^{-G} is simply the probability that no other message is transmitted in the slots used by our terminal. Summing over all transmissions heard at our destination we get

$$S_c = Ge^{-G} \tag{2}$$

where S_c denotes the total rate of successful traffic heard at our destination. This total traffic consists of messages with many different destinations, and the success of each message depends on what happens at its destination. But all these messages contend with our transmission for the use of the channel around our destination.

Equation (2) looks exactly like the equation describing a centralized slotted ALOHA system [16]. G and S_c do not, of course, depend on the transmission in question, and we can therefore say that any transmission sees an ALOHA system at its destination with a throughput equal to S_c, where the subscript on S_c stands for contending. If we unnormalize S_c and measure it in messages per unit time, we may use (1) and write the average delay per hop suffered by any message as follows:

$$T = \frac{1}{C - eS_c}. \tag{3}$$

In the centralized case, interference always destroys both messages involved. In the network case analyzed here, this is not necessarily true. Since the ranges of the transmission involved and their destinations may be very different, a collision of two messages at the first's destination will destroy the first, but may not bother the second at its destination. We shall use (3) for the delay in ALOHA networks, even though what happens at each destination is not equivalent to a closed, centralized ALOHA system; this is supported by [17] where the optimal transmission policy for ALOHA networks, given the hearing matrix, is shown to be identical to the optimal policy in centralized ALOHA systems. However, our goal here is to choose the optimum hearing matrix by choosing the transmission range.

Equation (3), as a model for the per-hop delay in ALOHA networks, rests on two procedures, both of which can be applied to random access networks in general. The first procedure is to use the contending traffic S_c and the available capacity C in the expression for delay in the *centralized* system to get the per-hop delay in the *network*. This procedure was first used in [2] to model large networks. It was presented

in [14] and evaluated by comparing with simulation results for networks with 10 and 20 nodes. Evaluating this procedure for really large networks is much harder—there is nothing feasible to compare with. The second procedure is to approximate the delay in a random access system by a simple two-parameter approximation like (1). If we substitute $1/e$ by the maximum utilization of any centralized random access scheme, (3) will then model the per-hop delay in the corresponding random access network.

The discussion so far applies to any network which is homogeneous and isotropic in a statistical sense. We shall now calculate S_c assuming every message is transmitted *with exactly the range necessary to reach its destination*. Let S be the total traffic coming out of a unit area, and let $f(r)$ be the traffic density. That is, the traffic going from one small (source) area dA_s to another small (destination) area dA_d is given by $f(r)dA_s dA_s$, where r is the distance between the two small areas. We obviously have $S = \int_{r=0}^{\infty} f(r)2\pi r\, dr$ and $f(r)2\pi r/S$ is therefore the probability density function for the distance traveled by a message. N, the average distance traveled by messages, is given by $NS = \int_{r=0}^{\infty} rf(r)2\pi r\, dr$. Let dS_c be the contribution to S_c of messages whose range is between r and $r + dr$. Such a message will be heard at a given destination if it starts anywhere within a circle with radius r around that destination. We can then write $dS_c = \pi r^2 f(r)2\pi r\, dr$, where πr^2 is the source area, $2\pi r\, dr$ the destination area, and $f(r)$ the traffic density. Integrating we get

$$S_c = \int_{r=0}^{\infty} \pi r^2 f(r)2\pi r\, dr = \pi S\overline{N^2} \tag{4}$$

where $\overline{N^2}$ is the second moment of the distance traveled. Substituting (4) in (3) we see that an ALOHA network in which every message reaches its destination exactly in one hop has the same delay-capacity relationship as a centralized ALOHA system carrying a total traffic $\pi S\overline{N^2}$.

Equation (4) can also be obtained directly. By symmetry S_c must be equal to S times the average area in which messages are heard, and this area is $\pi\overline{N^2}$.

III. Choosing the Range

The simplicity of (4) is a result of the assumption that power can be adjusted exactly to reach the destination. But even if we can adjust the range so as to exactly reach the destination in one hop, is this a good policy? In [8] the question was posed thus: should we take giant steps, assuming we can? It was shown there that if, for a given C and traffic requirement, the delay per hop grows without bound as a function of the step size R, then there is an optimal step size, and steps should not be giant. We wish to find the optimal range policy as a function of traffic requirements, and for this we need the following.

Theorem 1: If a message has to travel a distance X in k hops it should, in order to make the best use of the communication resources, do so in k equal hops, each of length X/k.

Proof: Whether we want to minimize T when S and C are given, or to minimize the necessary C when S and T are given, we must, in order to get the best system, minimize the total contending traffic at each destination. But this is equivalent to minimizing the total area at which any given message is heard. Let X_i be the length of the ith hop, where $\Sigma X_i = X$. The area in which our message is heard is proportional to the ΣX_i^2. Minimizing the area at which our message is heard is therefore the following convex quadratic programming problem:

$$\text{Minimize} \sum X_i^2$$

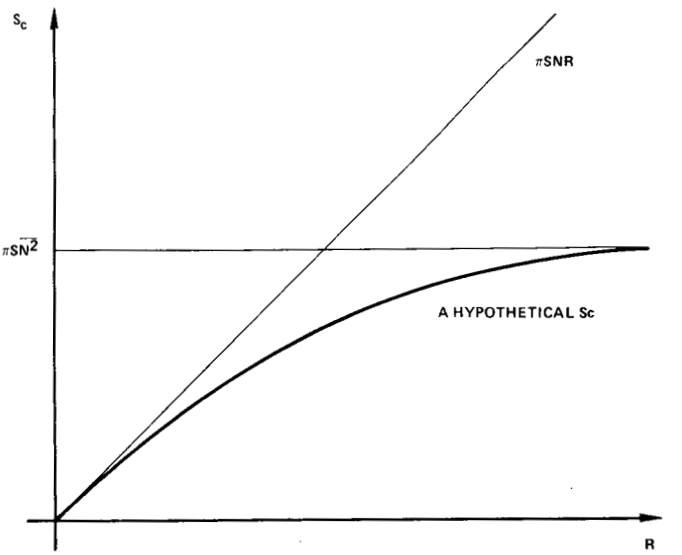


Fig. 1. Two bounds on S_c, the total traffic contending at each point.

$$\text{subject to} \sum X_i \geq X, \qquad X_i \geq 0.$$

The solution of this minimization problem gives the equal step result stated in the theorem.

Let us now consider the following family of policies which use a perfectly adjustable but limited transmission range. Given the maximum range R, the path of every message will be divided into the minimum number of equal hops. Which R will give the best overall system performance? Should we try to make R as large as possible? To answer these questions we must determine how S_c depends on R.

Writing S_c as a function of R and the distribution of the distances traveled is a straightforward but cumbersome operation. However, the following bounds are simple to obtain. Since $S_c(R)$ is a monotonic increasing function of R, an obvious bound is $S_c(R) \leq S_c(\infty) = \pi S\overline{N^2}$. When R is very large, all messages will reach their destination in one hop, so the equality here follows from (4). Another bound, especially useful when R is small, can be obtained as follows. The total area covered by the several transmissions of a message that has to travel a distance r can be bounded from above by $(r/R)\pi R^2$. In analogy to (4), $S_c(R)$ can therefore be bounded by

$$S_c(R) \leq \int_{r=0}^{\infty} \frac{r}{R} \pi R^2 f(r)2\pi r\, dr = \pi RNS. \tag{5}$$

Fig. 1 shows the two bounds and a hypothetical $S_c(R)$.

We shall assume that the traffic to be carried is specified, that an acceptable delay is specified, and that the goal of a good design is to make the necessary bandwidth as small as possible. The specification can be summarized by the dimensionless quantity N^2ST. When $N^2ST \ll 1$ we call the network and the traffic *bursty*, and when $N^2ST \gg 1$ we call the network *steady*.

For small R we can use the bound of (5) as an approximation for $S_c(R)$, and we will combine it with N/R as an approximation for the average number of hops per message, to get the following approximate expression for the delay

$$T = \frac{N/R}{C - e\pi SNR}.$$

Inverting we get

$$C = e\pi SNR + \frac{1}{T}\frac{N}{R} \tag{6}$$

and from this approximate expression for C we get that R^*, the optimal R (i.e., the R that minimizes the necessary C for given N, S, and T), is given by

$$\frac{R^*}{N} = \frac{1}{\sqrt{\pi e N^2 ST}}. \tag{7}$$

While we use the term optimal R, (7) actually determines the optimal value for the maximum transmission range. Given the distance a specific message has to travel, R^* determines the necessary number of hops, and the transmission range of all hops is then chosen according to Theorem 1. The capacity necessary when using the optimal R can be obtained from (6) with the use of (7); it is given by the following relation between CT and N^2ST, both of which are dimensionless quantities:

$$CT = 2\sqrt{e\pi N^2 ST}. \tag{8}$$

When the traffic is very steady (i.e., when $N^2ST \gg 1$), (7) says that R^* will be much smaller than N. The approximations made when writing (6) are consistent with this result, which is also quite intuitive. Consider a steady system with a given S and a large T. When we are willing to tolerate a large T the number of hops can be large, and we can therefore choose a small R. Each message will then be heard only in a narrow strip along its path, so S_c will be small, and the necessary bandwidth will therefore also be small. When the traffic is very bursty, we get from (7) that R^* is much larger than N. This is again very intuitive—when the traffic is bursty there is little contention, and therefore, almost nothing is gained by forcing a message to undergo more than one hop. But the exact value given by (7) is not meaningful when the traffic is bursty, because the approximations used when writing (6) are not valid when R is large.

A general conclusion that emerges is that in a random access network it is better to limit the transmission range, even if our terminals can adjust their range exactly and have no power limitation. This voluntary limiting is especially important when the traffic is very steady, and the optimal range limit R for ALOHA networks is then given by (7).

How shall we define the *quality* of networks? Clearly one *should not* compare a network to one huge centralized $M/M/1$ system that carries all messages to one common destination because practical networks have an advantage over centralized systems: the same capacity can be used in different regions of the network to successfully transmit different messages at the same time. That is, network capacity can be spatially reused.

A common measure used to characterize access schemes is the maximum utilization they can make of the given communication resources. This maximum utilization is sometimes called capacity, especially by authors whose variables are normalized by the slot size, and who therefore do not explicitly mention the channel bandwidth. We use the word capacity to describe an amount of communication resources (i.e., the number of bits or messages that can be transmitted per second), and utilization to denote the useful fraction of that capacity.

The quality of a very steady centralized system, as defined by us [2], is equal to its maximum utilization. But utilization is not a good measure for networks with a continuum of terminals since utilization can be abritrarily increased by spatial reuse, i.e., by limiting the transmission range.

It seems that every network organization must address the question of how to coordinate every transmission with at least all the traffic that is heard at its destination. Since the best possible system will coordinate this traffic perfectly, we shall compare all networks to the network that uses the same technology (i.e., omnidirectional antennas) but that somehow achieves perfect coordination between the traffic contending at every point, and in which transmission ranges are chosen optimally. We shall call this "best possible" network with perfect coordination the $M/M/1$ network, and shall define the quality Q of any network to be the inverse ratio between the capacity necessary for it when S and T are given and the capacity necessary in the $M/M/1$ network for the same S and T. In general $Q \leqslant 1$, and equality holds only for the $M/M/1$ network itself. The capacity necessary for this *best possible* $M/M/1$ network scheme is in general a function of S, T, and the distribution of distances traveled. For very steady traffic we get, in analogy to (7), that the optimal R is given by

$$\frac{R^*}{N} = \frac{1}{\sqrt{\pi N^2 ST}} \tag{9}$$

and when using this R^*, the capacity necessary is

$$CT = 2\sqrt{\pi N^2 ST}. \tag{10}$$

Dividing (10) by (8) we get that the quality of a heavily loaded ALOHA network with the optimal step size is $1/\sqrt{e} = 0.607$! How did we get this dramatic improvement over the heavily loaded centralized ALOHA system, whose quality is $1/e = 0.367$?

We may say that every message sees at its destination an ALOHA system whose utilization, which we shall call local utilization, is S_c/C. When the traffic is very steady and when the optimal R is used, we get by substituting (7) in (5) that the centralized local utilization is $1/2e$, i.e, half the maximum possible utilization of a centralized ALOHA system. The quality of a centralized ALOHA system with utilization $1/2e$ is 0.68. It is only at much higher utilizations (closer to $1/e$) that the quality of a centralized ALOHA system goes down to $1/e$. The need for several hops will bring the quality of the ALOHA network down, from 0.68 to 0.607. We see, therefore, that by choosing the optimal R as a function of burstiness, our ALOHA network has gained a self-adjusting capability, and it will not allow itself to be pushed to higher loads, where it is really bad.

Results analogous to (8) can be obtained for any random access network, and the self-adjusting capability is common to all of them: the quality of the very steady random access *network* with the optimal R is the square root of the quality of the corresponding very steady *centralized* system. From (8) we also see that random access networks with the optimal R show an economy of scale when very steady: for a given T, the necessary C grows only like $\sqrt{S}$.

Comparing (7) and (9) we see that the optimal transmission radius R in a steady ALOHA network is smaller than the optimal R in an $M/M/1$ network by a factor $1/\sqrt{e}$. The optimal R in both networks goes to zero as the traffic becomes very steady. We have implicitly assumed that there always is a terminal at the end of the hop that can receive our message and forward it. But if R becomes too small, there may not be a terminal so conveniently situated. If R becomes even smaller, our terminal may not be able to communicate with any other terminal, and the network may become disconnected. Kleinrock and Silvester [11] treat this issue explicitly, while calculating the optimum transmission range with a different objective: obtaining the maximum throughput from the given channel, assuming infinite delay is acceptable. We shall not treat this issue here, but our assertion about the self-adjusting capability of networks must be qualified.

Consider once again an ALOHA network and an $M/M/1$ network, both carrying the same very steady traffic. If it is practical for the ALOHA network to choose the optimal

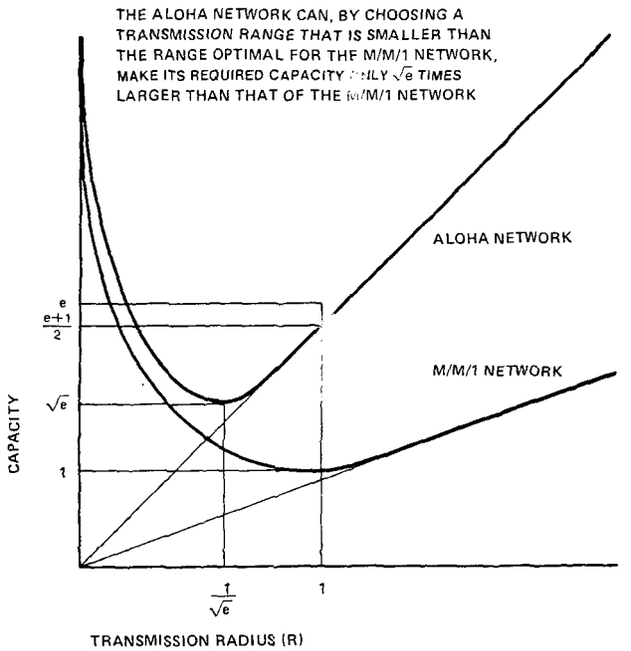


Fig. 2. Capacity necessary for very steady two-dimensional networks.

R according to (7), then it will need only $\sqrt{e}$ times more capacity than the optimal $M/M/1$ network, i.e., its quality will be $1/\sqrt{e}$. But if R cannot be made so small, the quality of the ALOHA network will go down. If the ALOHA network is constrained to use the same R as the optimal $M/M/1$ network, then its local utilization will be $1/(e+1) = 0.269$ and its quality will be $2/(e+1) = 0.538$. If both the ALOHA and the $M/M/1$ networks carry a very steady traffic but are constrained to use an R that is much larger than the one given by (9), then the local utilization of the ALOHA network and its quality will be $1/e$.

Fig. 2 sketches the dependence of the necessary capacity on the transmission range, in the ALOHA and $M/M/1$ networks.

Our treatment of random access networks can be summarized as follows.

Theorem 2: Consider a network carrying a very steady traffic and using a random access scheme whose maximum utilization, when used in a centralized communication system, is u.

Assume that the range of every transmission can be perfectly adjusted, but only up to a maximum range R. If R can be optimized freely (i.e., made as small as necessary), then each transmission will see a system whose *local utilization* is $u/2$ and the *network quality* will be $\sqrt{u}$.

Proof: Follows trivially from the preceding discussion.

Theorem 2 can be immediately generalized to the situation in which the antenna carried by terminals is somewhat directional. Assume the antenna radiates into a cone, which takes a fraction α of the sphere. This is, of course, a gross simplification of the real radiation pattern, but is consistent with our simple modeling of transmission range. If we compare the case of an omnidirectional antenna to this case of an α-directional antenna we find that, with any transmission policy, the total interfering traffic at any point is smaller by a factor α. The optimal R for steady traffic, given by (13), will become larger by $1/\sqrt{\alpha}$ (we shall not have to push so much towards small R), and the necessary capacity of (14) will become smaller by $\sqrt{\alpha}$. But when we compare an α-directional random access network to an α-directional $M/M/1$ network we find that the local utilization and the network quality in the optimized structure will remain as stated in Theorem 2. An improved technology (i.e., directionality) will help both the random access network and the $M/M/1$ network. But whenever they use the same technology, a comparison between them will show the inherent cost due to the random access aspect of the network, and this inherent cost is $1/\sqrt{u}$.

Until now we have assumed the networks consist of many terminals distributed uniformly in the plane. Theorem 2 can be easily generalized [3] to networks consisting of terminals distributed in more than two dimensions.

Somewhat surprisingly, Theorem 2 is not valid for one-dimensional networks whose terminals are distributed in one dimension, for example, along a coastline. In that case we get the following.

Theorem 3: In a one-dimensional network S_c is equal to $2NS$, and is independent both of the need to break message paths into several hops and of the policy of implementing such a break, as long as the policy is applied everywhere in the same way, that is, as long as a message path of a given length will be broken in the same way, wherever it originates.

Proof: Given in [3].

In one-dimensional networks, if range can be perfectly adjusted we should, therefore, take a giant step whenever possible. Even when the traffic is very steady there is no reason to limit the step size, since no decrease in S_c will follow. One-dimensional ALOHA networks have a local utilization and a network quality both of which are equal to $1/e$. In the rest of this paper we shall consider only two-dimensional networks.

Theorem 2 answers the question of the optimal transmission range when the traffic is very steady. This is satisfying because random access schemes have an efficiency problem exactly when the traffic is steady. When the traffic is bursty, there is little need for improving random access networks. When range is perfectly adjusted, the range limit R grows when the traffic becomes bursty, and when the traffic is very bursty, giant stepping is the best. That is, each message should be transmitted with enough range to reach its destination directly (in one hop). These general conclusions change, once we consider networks in which range cannot be perfectly adjusted.

Assume now that terminals cannot adjust the range of their transmissions, and that all transmissions, by all terminals, must have a fixed range R. Since the range of all transmissions is fixed and constant, some messages will overshoot their destinations. The amount of traffic contending at every point will therefore be larger now than it was when range was perfectly adjusted. When the traffic is very steady, R will be very small, and the overshoot will not contribute anything significant to S_c. Theorem 2 will therefore be valid even if all transmissions must use the predetermined range [3]. When the traffic is bursty, S_c will grow significantly when all transmissions have a fixed range, and R will then have to be limited.

To summarize this section: when considering centralized systems we can say that random access schemes are good when the traffic is bursty and bad when the traffic is steady. This statement is true in general for networks, too. But networks have a self-adjusting property—by controlling the maximum transmission range and reducing it when the traffic is steady, we can make random access networks suffer less from destructive interference than centralized systems.

IV. Improving ALOHA Networks

In this section we shall consider two ideas that can improve random access communication systems by making them less random, in a sense. These ideas—dividing terminals into power groups and creating a multilevel hierarchical organization—are in principle applicable to every random access scheme, but we shall analyze only their effect on ALOHA.

In the models of ALOHA systems presented so far, we assumed that in the case of interference, both messages will be destroyed. But if the colliding messages vary greatly in received power, the receiver may be able to receive the stronger one correctly even in the presence of the other, weaker, signal. The receiver is then said to *capture* the stronger signal. The capability to capture some messages will obviously improve every ALOHA system. Let us first see the resulting improvement in a centralized ALOHA system, where all messages have one common destination. Roberts [6] proposed and analyzed a capture model in which the power differences resulted from different distances to the common destinations. Our approach is different. We shall assume that the terminal population is split into two groups, that one group is transmitting with more power than the other, and that this splitting is purposely done in order to improve system performance. In order to abstract the geometric details out of the model, we shall adopt the following assumption [15]. The power of the two groups is significantly different. When two transmissions from the same group occur simultaneously, they will always destroy each other. When one strong transmission and any number of weak transmissions compete for the ear of the common station, the strong one will always be captured successfully. This separation into groups introduces, therefore, a partial coordination into the random world of ALOHA.

It may be possible to achieve such a coordination between groups by techniques that do not rely on a power difference between them. A distinctive preamble, for example, may allow a terminal to successfully receive a transmission from one group, which we shall call strong, even in the presence of transmissions from the weak group. In a system which is not perfectly slotted, the first of two interfering signals of equal strength to arrive at a receiver may survive the collision and be successfully received. From now on, strong and weak should not therefore be taken literally—they do not necessarily refer to transmission power, but simply characterize the group of transmissions likely to win or lose when competing with the other group.

What will be the resulting improvement if we introduce groups into a heavily loaded ALOHA centralized system? If the strong group is selfish it can ignore the weak group, and use the channel as much as possible. The strong group will then successfully utilize $1/e = 0.367$ of the slots, and will leave $1/e$ of the slots free. (In addition, 0.276 of the slots will be wasted on collisions.) The weak group can utilize at most $1/e$ of what is left free for it, i.e., it can utilize $1/e^2 = 0.135$ of the slots, and the total rate of success by both groups will be 0.503.

The channel can be better utilized if the strong group will not be so selfish. To see this, let us now consider the division into groups as a design parameter. Assume that we have an infinite population of terminals, and that each terminal contributes only a minute fraction of the total traffic. While we have spoken of strong and weak terminals, the important design question is not the identity of terminals in each group but the portion of the traffic in each group. If we have an extremely heavy load, our goal is to find the division into groups that will allow our system to utilize the greatest portion of the communication resource available. Let G_1 and S_1 be the total offered traffic and the rate of success of the strong group, G_2 and S_2 the corresponding values for the weak group. For simplicity we shall assume in this section that S and G are measured per slot size. Using our standard assumption, that the total traffic offered by a terminal is a Bernoulli process, independent of the traffic offered by all other terminals, we can write

$$S_1 = G_1 e^{-G_1}$$
$$S_2 = G_2 e^{-G_2} e^{-G_1}. \tag{11}$$

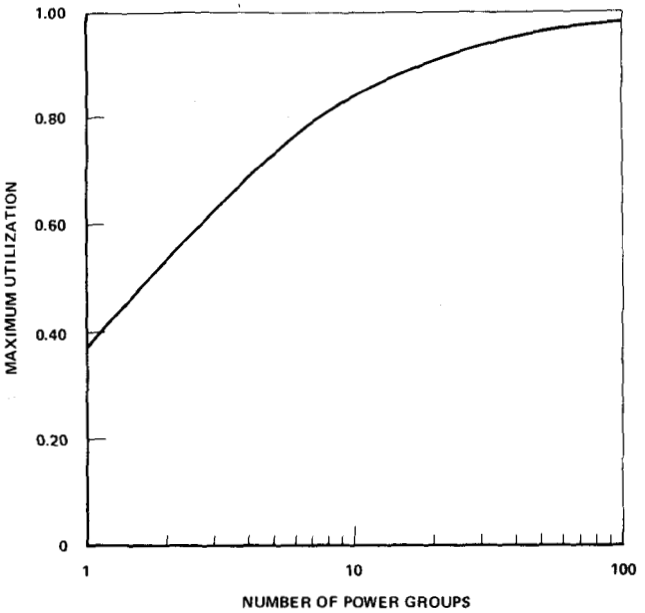


Fig. 3. Maximum utilization of ALOHA with power-groups.

Choosing G_1 and G_2 in order to maximize $S = S_1 + S_2$ we find that the best values are $G_1 = 1 - 1/e$ and $G_2 = 1$, that the maximum utilization of a system with two groups is $e^{-(1-1/e)} = 0.531$, and that this utilization is achieved when $S_1/S_2 = e - 1$. The above treatment can be generalized to many groups.

Theorem 4: Consider a slotted ALOHA system whose infinite population of terminals is optimally divided into r progressively weaker groups, such that a message will never be bothered by transmissions from weaker groups, and will always be destroyed by any transmission from its own group, or from a stronger group. Then V_r, the maximum utilization of this r-group ALOHA system, satisfies the following recursion relation:

$$V_{r+1} = e^{-(1-V_r)}.$$

Proof: Follows directly from the generalization of (11). See [3].

The sequence V_r, whose first portion is shown in Fig. 3, is a monotonic increasing sequence converging (slowly!) to 1. This is not surprising, since when we have a large number of groups, most collisions will be between messages from different groups, and one of the messages will be successful.

Until now we have applied the idea of partially coordinated groups (i.e., power groups) to centralized ALOHA systems. How can it be applied to networks? In our analysis of ALOHA networks we have used the transmission power to control range. We shall now assume that the division into groups is done by means which are independent of power so that transmission range can still be freely chosen. We shall also assume that the policy of assigning transmission power is independent of position, and that the density of both strong and weak sources is high and uniform.

One simple way to improve ALOHA networks by using groups is the following. The same transmission range will be chosen for both strong and weak transmissions, and the partial coordination between them will simply improve the local ALOHA system. We saw that the maximum local utilization of a two-group ALOHA system is 0.531. Substituting this in (8) we see that by using two groups with the same range, the quality of ALOHA networks can be improved from $\sqrt{0.367} = 0.607$ to $\sqrt{0.531} = 0.729$. We see that since networks are less sensitive than centralized systems to the limited

utilization of the ALOHA scheme, it is harder to improve them by introducing a better scheme.

The capability to divide terminals into two partially coordinated groups can lead to a greater improvement of ALOHA networks (in two or more dimensions) if transmission range is chosen independently for the two groups. Let N_1 and N_2 be the average distance traveled by messages from the strong and weak group, respectively. Let S_1 and S_2 be the traffic density of the strong and weak group, and let T_1 and T_2 be the average delay suffered by messages from the strong and weak group, respectively. In a heavily loaded system, if the strong group is absolutely selfish it will utilize the full channel in the way best for it, and we then get from (8) that T_1 and S_1 satisfy

$$T_1 = 4e\pi \frac{N_1^2 S_1}{C^2}.$$

The local utilization of the strong group, when optimized for heavy traffic, is $1/2e$. It is easy to calculate that the strong group leaves then a fraction $b = 0.793$ of the time slots unused, and these slots are available for the weak group. That is, the capacity available to the weak group is bC. Using (8) we get that

$$T_2 = 4e\pi \frac{N_2^2 S_2}{b^2 C^2}.$$

T, the message delay averaged over all messages, from both groups, is given by $TS = T_1S_1 + T_2S_2$, and our goal is to minimize T by choosing N_1, N_2, S_1, and S_2 subject to $S_1 + S_2 = S$ and subject to $N_1S_1 + N_2S_2 = NS$. It is simple to see that T is minimized when $N_1S_1/N_2S_2 = 1/b^2 = 1.59$ and is then given by

$$T = 4\pi \frac{e}{1+b^2} \frac{N^2}{C^2} S. \tag{12}$$

The quality of this two-group network is therefore $\sqrt{(1+b^2)}/e = 0.774$.

It is interesting to note that $(T_1/T_2) = (N_1/N_2)$ but that $(S_1T_1/S_2T_2) = (1/b^2) = 1.59$. That is, we can choose the ratio between T_1 and T_2 at will (by adjusting N_1/N_2) but the contribution of the strong and weak group to the average delay and to the average number of messages in the network will always, in an optimized system, be in a fixed ratio.

In deriving (12) we assumed the strong group is selfish. In [3] we show that the very steady two-group ALOHA network will be slightly better if the strong group is not absolutely selfish. For a summary of the optimal range and the necessary capacity in various two-dimensional networks, see Table I.

A random access communication system carrying a given amount of traffic generated by few terminals will perform better than a system carrying the same traffic generated by many terminals. The reason is that two messages generated by the same terminal will never collide. A system with fewer terminals will therefore have to suffer less contention. For example, Abramson [1] showed that while the maximum utilization of "infinite population" ALOHA is $1/e$, the maximum utilization of an ALOHA system consisting of a station and two terminals is $1/2$.

Since random access systems with a small population have better utilization and smaller delay than systems with a large population, one is led to the following hierarchical scheme for a centralized communication system. Divide the very large terminal population into a small number of groups. Assign a repeater to each terminal group. Each group will communicate with its repeater, and the repeaters will communicate with the station. All communications will use the full capacity of the channel. Repeaters may sometimes be necessary in order to extend the range of transmission, but we shall assume this is not a problem, and shall only be interested in introducing repeaters in order to improve system performance, that is, to lessen the delay when S and C are given, or lessen the capacity necessary when S and T are given. In [3] we show that such a two-level centralized system based on ALOHA will be better than the one-level ALOHA when heavily loaded. When very heavily loaded three levels will be even better, but more than three ALOHA levels are never necessary.

TABLE I
BEST TRANSMISSION RANGE AND NEEDED CAPACITY FOR NETWORKS

Organization		Range	Capacity
$M/M/1$		R_0	C_0
ALOHA (one group)		$0.607R_0$	$1.647C_0$
ALOHA (two groups, same range)		$0.729R_0$	$1.372C_0$
ALOHA (two groups, separate ranges)	selfish	$0.774R_0$	$1.292C_0$
	considerate	$0.782R_0$	$1.279C_0$

$$R_0 = \frac{1}{\sqrt{\pi ST}} \qquad C_0 = 2\sqrt{\pi N^2 S/T}$$

Multilevel ALOHA centralized systems can be better than one-level ALOHA when the traffic is heavy, because in the top level we can have a contention system with a small population, which can better utilize its communication resources. In [3] we show that such a multilevel organization will never improve ALOHA networks. In heavily loaded ALOHA networks the optimal transmission radius is small. That is, even without repeaters, whenever the traffic is steady we should make our contending terminal system as small and as finite as we dare! Repeaters are not necessary for improving the utilization of heavily loaded networks, and the extra level they introduce is wasteful. Repeaters can be very useful, for networks of intermediate burstiness, if ALOHA is used for terminal-repeater communication and dedicated channels are used for repeater-repeater communication. For a treatment of such mixed-mode networks see [4].

In this section we have considered two ideas that can improve random access schemes: power groups and multilevel hierarchical organization. We have treated ALOHA in detail, but the conclusions are general and intuitive. Heavily loaded centralized systems must carry all traffic to the common destination. But networks can adjust, by choosing the transmission range, and make sure the channel is not very heavily loaded at any point. Schemes, like power groups, that improve centralized systems by a certain factor will improve networks by the square root of that factor, a less significant improvement. Schemes, like a hierarchy of levels, that improve centralized systems only when they are heavily loaded will not improve networks at all, because networks will never be that heavily loaded.

V. Conclusions

Using a random access scheme for a communication system consisting of a large number of distributed terminals is extremely simple and therefore appealing. But for example, a heavily loaded *centralized* ALOHA system, in which all messages must reach one common destination, will need e times more bandwidth than the theoretical best (and impossible!) $M/M/1$.

Random access *networks* are in a better position. Since messages have various distributed destinations the channel can

be spatially reused: i.e., various transmissions can successfully use the channel at the same time if they are separated spatially and do not interfere at their destinations. The contention between messages is not directly determined by the given traffic, and it can be adjusted by choosing the transmission range.

By modeling a homogeneous and isotropic network by a continuum of terminals, we calculated the optimal transmission range. An ALOHA network need be only $\sqrt{e}$ times worse than the corresponding $M/M/1$ network, even when very heavily loaded, as long as the calculated optimal range is not too small to be practical. The calculated range becomes too small when only a few terminals are within range of each other. But the problem of organizing and coordinating a system with a large number of terminals, which was the original motivation for using random access, has disappeared, and other access modes can then be used to advantage, although we have not considered any in this paper.

Since networks pay a smaller price for contention than do the centralized systems, it is harder to improve them by reducing contention. Splitting terminals into power groups can improve any random access system, especially when the traffic is split between groups in a good way, but the resulting improvement in centralized systems is much more significant than the resulting improvement in networks.

In a centralized system all messages must reach the station, and must therefore contend for its ear. A multilevel organization using ALOHA at all levels can improve heavily loaded single-destination systems by having only a small number of intermediate nodes communicate directly with the station. Multilevel ALOHA organizations do not help networks, because choosing the transmission range is a much more effective means for controlling the amount of contention.

REFERENCES

[1] N. Abramson, "Packet switching with satellites," in *AFIPS Conf. Proc., Nat. Comput. Conf.*, 1973, vol. 42, pp 695–702.

[2] G. Y. Akavia, "Hierarchical organization of distributed packet-switching communication systems," Ph.D. dissertation, Dep. Comput. Sci., Univ. California, Los Angeles, Mar. 1978.

[3] G. Y. Akavia and L. Kleinrock, "Performance tradeoffs and hierarchical designs of distributed packet-switched communication networks," Comput. Syst. Modeling and Anal. Group, School Eng. Appl. Sci., Univ. California, Los Angeles, Rep. UCLA-ENG-7952, Sept. 1979.

[4] ——, "On the advantage of mixing ALOHA and dedicated channels," submitted for publication.

[5] M. J. Ferguson, "A study of unslotted ALOHA with arbitrary message lengths," Univ. Hawaii, Honolulu, Tech. Rep. B75-13, Feb. 1975.

[6] ——, "A bound and approximation of delay distribution for fixed-length packets in an unslotted ALOHA channel and a comparison with time division multiplexing (TDM)," *IEEE Trans. Commun.*, vol. COM-25, pp. 136–139, Jan. 1977.

[7] L. Kleinrock, "Resource allocation in computer systems and computer-communication networks," in *Information Processing 1974, Proc. IFIP Cong.*, Stockholm, Sweden, Aug. 1974. Amsterdam, The Netherlands: North-Holland, 1974, pp. 11–18.

[8] ——, "On giant stepping in packet radio networks," Univ. California, Los Angeles, Internal Note, Mar. 1975.

[9] ——, *Queueing Systems, Vol. II: Computer Applications*. New York: Wiley-Interscience, 1976.

[10] ——, "Performance of distributed multi-access computer communication systems," in *Information Processing 1977, Proc. IFIP Cong.*, Toronto, Ont., Canada, Aug. 1977. Amsterdam, The Netherlands: North-Holland, 1977, pp. 547–552.

[11] L. Kleinrock and J. Silvester, "Optimum transmission radii for packet radio networks or why six is a magic number," in *Conf. Rec., IEEE Nat. Telecommun. Conf.*, Birmingham, AL, Dec. 3–6, 1978, pp. 4.3.1–4.3.5.

[12] S. Lam, "Packet switching in a multi-access broadcast channel with application to satellite communication in a computer network," Comput. Syst. Modeling and Anal. Group, School Eng. Appl. Sci., Univ. California, Los Angeles, Rep. UCLA-ENG-7429, Apr. 1974 (also Ph.D. dissertation, Dep. Comput. Sci., U.C.L.A.).

[13] ——, "A new measure for characterizing data traffic," *IEEE Trans. Commun.*, vol. COM-26, pp. 137–140, Jan. 1978.

[14] B. M. Leiner, "A simple model for computation of packet radio network communication performance," *IEEE Trans. Commun.*, vol. COM-28, pp. 2020–2023, Dec. 1980.

[15] J. J. Metzner, "On improving utilization in ALOHA networks," *IEEE Trans. Commun.*, vol. COM-24, pp. 447–448, Apr. 1976.

[16] L. G. Roberts, "ALOHA packets system with and without slots and capture," *Comput. Commun. Rev.*, vol. 5, pp. 28–42, Apr. 1975.

[17] Y. Yemini and L. Kleinrock, "On a general rule for access control or, silence is golden," in *Proc. Int. Symp. Flow Contr. Comput. Networks*, Versailles, France, 1979, J.-L. Grange and M. Gien, Eds. Amsterdam, The Netherlands: North-Holland, 1979, pp. 335–347.

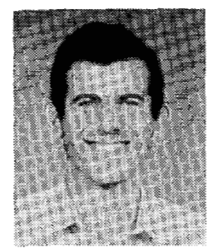
★

Leonard Kleinrock (S'55–M'64–SM'71–F'73) received the B.S. degree in electrical engineering from the City College of New York, New York, NY, in 1957 and the M.S.E.E. and Ph.D.E.E. degrees from the Massachusetts Institute of Technology, Cambridge, in 1959 and 1963, respectively.

While at M.I.T., he worked at the Research Laboratory for Electronics as well as with the computer research group of Lincoln Laboratory in advanced technology. He joined the faculty at the University of California, Los Angeles, in 1963. His research interests focus on computer networks, packet radio systems, and local area networks. He has had over 120 papers published and is the author of three books, *Communication Nets: Stochastic Message Flow and Delay* (New York: McGraw-Hill, 1964), *Queueing Systems, Vol. I: Theory* (New York: Wiley, 1975), and *Queueing Systems, Vol. II: Computer Applications* (Wiley, 1976), as well as the *Solutions Manual for Queueing Systems, Vol. I* (Wiley, 1982). He served as the Head of the U.C.L.A. Department of Computer Science Research Laboratory and is a well-known lecturer in the computer industry. He is Principal Investigator for the DARPA Advanced Teleprocessing Systems contract at U.C.L.A. and Co-Principal Investigator for the NSF Advanced Network Environment for Distributed Systems Research Project.

Dr. Kleinrock was recently elected to the National Academy of Engineering, is a Guggenheim Fellow, and serves on the Boards of Governors of various advisory councils in the computer field. He is a member of the Science Advisory Committee for IBM. He has received numerous best paper and teaching awards, including the ICC '78 Prize Winning Paper Award, the 1976 Lanchester Prize for outstanding work in operations research, and the IEEE Communications Society 1975 Leonard G. Abraham Prize Paper Award. In 1982, as well as having been selected to receive the C.C.N.Y. Townsend Harris Medal, he was co-winner of the L. M. Ericsson Prize, presented by His Majesty, King Carl Gustaf of Sweden, for his outstanding contributions in packet switching technology.

★



Gideon Y. Akavia (S'76–M'78) was born in Haifa, Israel, in 1946. He received the B.Sc. degree in mathematics and physics from the Hebrew University of Jerusalem, Jersalem, Israel, in 1967, the M.Sc. degree in physics from the Weizmann Institute of Science, Israel, in 1969, and the Ph.D. degree in computer science from the University of California, Los Angeles, in 1978.

From 1968 to 1970 he was an Experimental Physicist at the Stanford Linear Accelerator Center, Stanford, CA. From 1970 to 1973 he served in the Israel Defense Force. From 1973 to 1975 he was a Software Analyst with NATAM, Tel Aviv, Israel. From 1978 to 1980 he was on the Faculty of Computer Science, Technion—Israel Institute of Technology, Haifa. Since 1980 he has been with the Center for Military Analyses, Ministry of Defense, Haifa. His research interests include computer-communication networks, office automation, military history, and the impact of advanced technology on the structure and function of large organization and systems.