Chapter 3
Multiple Access Protocols

3.1. INTRODUCTION

Consider the sharing of a single broadcast channel by a population of \( N \) distributed users. The set of rules and algorithms that define the method of sharing is referred to as the multiple access protocol. When we compare different multiple access protocols, we should bear in mind that the performance of a multiple access protocol depends on various parameter values that can differ greatly in different networks. Also, certain protocols assume specific hardware capabilities that are available in one network but that cannot be easily implemented in another network.

Let us begin by assuming only that the transmission of any user in the broadcast channel (if it is the only transmission present) is observable and can be received by all other network users after a small propagation delay. If the transmissions of two different users overlap in time in the channel, the "collision" is observable to all network users after a small propagation delay, but none of the transmissions can be received correctly (i.e., no spread spectrum capability). Several tutorial articles on multiple access techniques are available in the literature [LAM 79, TOBA 80a, LAM 83]. One of them [LAM 83] is reprinted below.

Traditional techniques for solving the multiple access problem are based upon dividing the broadcast channel into a pool of subchannels, using either frequency division or time division. These subchannels may be fixed-assigned to the population of users; in this case, \( N \) subchannels are needed. The subchannels may also be demand-assigned to the population of users; in this case, fewer than \( N \) subchannels are needed. Demand assignment may be achieved with either a central controller or a distributed control algorithm. These techniques are said to be channel-oriented since each unit of allocation is one subchannel for a relatively long time duration.

It is well known in queueing theory that a single, high-speed server is often preferable to multiple servers whose aggregate service capacity is the same as that of the single server. This is particularly true in data communications where traffic sources are typically very bursty [LAM 78]. Multiple access protocols of recent interest are primarily packet-oriented, i.e., packets from different users are scheduled in some fashion to use the entire broadcast channel.

The multiple access problem is thus simply the problem of forming a queue among a population of distributed users. To form such a distributed queue, there are two subproblems. (As shown below, these subproblems can be solved with either a central controller or a distributed control algorithm.)

1. Identify those users with data to send and who desire channel access (the ready users).

2. Assign the channel to a ready user according to some scheduling discipline.

Instead of identifying all ready users present in the population and employing an explicit scheduling discipline, the following variation may be employed. If the channel is idle, it is next assigned to the first ready user identified. In this case, the scheduling discipline is implicit in the algorithm for finding a ready user.

In Section 3.2, we give a taxonomy of multiple access protocols. We assert that most multiple access protocols that have been proposed are based upon one or a hybrid of these techniques. (Note: This taxonomy is different from the one in [LAM 83].) In Section 3.3, we provide two simple performance criteria for comparing different
protocols. We also discuss techniques for enhancing the throughput-delay performance characteristics of the basic protocols described in Section 3.2. Section 3.4 gives an overview of protocols based on the availability of some advanced capabilities in the broadcast channel.

3.2. A TAXONOMY

Multiple access protocols employ one of three basic techniques for finding ready users: linear search, tree search, and contention.

Linear search

A prime example of a multiple access protocol based upon linear search is roll-call polling. A central controller is employed, and the population of users is queried according to some sequential ordering of the users.

Distributed control may also be used. In hub polling, a special message, called a token, is passed among the users according to some sequential order. When the token reaches a user who is ready, he is assigned access to the channel.

An alternative method to implement distributed control is as follows: Suppose that the users are time synchronized so that the channel is divided into time slots following the end of a transmission (Figure 3-1). A linear search of the users can be performed by assigning one user to each time slot according to a predefined sequential order. A user can indicate in his time slot whether or not he is ready. The time slot duration must be long enough so that at the end of a time slot the outcome of the slot is globally known.

Suppose that following the end of a transmission, the channel is next assigned to the first ready user found. The scheduling discipline can be implicitly defined by specifying the user sequence following a user n transmission (Figure 3-1). If the sequence is always 1, 2, 3, ..., then we have a priority discipline with user 1 having the highest priority and user N, the lowest priority. If the sequence is n+1, n+2, n+3, ..., (modulo N) then we have a round-robin discipline.

\[
\text{Figure 3-1. A time-slotted channel.}
\]

For example, both MSAP [KLEI 77] and BRAM [CHLA 79] are linear search techniques based on the use of time slots as described above.

Tree search

Protocols that find ready users by a tree search can also be implemented with either a central controller or by means of time slots and distributed control similar to what has been described earlier. Suppose that the network users correspond to the leaves of a tree. Without loss of generality, consider a binary tree. The tree is traversed starting from the root node, and the ready users are identified using a divide-and-conquer type of algorithm.

A central controller, if used, can "probe" a group of users simultaneously instead of polling them one after another. When probed, all ready users in the group will respond. The central controller can distinguish between two outcomes: (1) no response, or (2) some response (at least one ready user in the group). In the tree traversal, when a node is visited, the central controller probes the entire group of users that are the descendants of the visited node. If there is no response, the entire subtree is "pruned." On the other hand, if there is a response, then each of the two sons of the node will have to be visited subsequently in the tree traversal [HAYE 78].

The tree traversal can also be carried out using time slots (see Figure 3-1). If all users exercise the same tree traversal algorithm and observe the same outcomes in the time slots, they will arrive at the same conclusion about the status of each user. In this case, the node to be visited
next in the tree traversal is assigned the next time slot. It is assumed that when a node is visited, all ready users who are descendants of that node will transmit into the assigned time slot. Suppose that three possible outcomes can be distinguished in a time slot: (1) no one ready, (2) a single user ready with his identity indicated, and (3) a collision (two or more users ready). When a node is visited, the occurrence of either of the first two outcomes will provide information on the status of all its descendants [CAPE 77, CAPE 79]. Hence, the subtree may be pruned. It should be clear that the ability to distinguish three outcomes in a time slot instead of the two outcomes from a probe will give rise to a more efficient search. In practice, the ability to recognize these and, perhaps, other outcomes will depend upon the communication technology being used. As before, if the channel is assigned to the first ready user found in each tree traversal, then the scheduling discipline is implicitly specified by the tree traversal order. We reprint [CAPE 79] below.

We note that the MLMA protocol in [ROTH 77] and the binary elimination protocol in [MOK 79] are also based on a tree search as described above.

As discussed below, tree search protocols have poor performance whenever the population of users contains many ready users. Both Hayes [HAYE 78] and Capetanakis [CAPE 77] investigated traffic conditions and techniques for adaptively switching their protocols between tree search and linear search. (See also Section 4.2.2 of [LAM 83] reprinted below.)

**Contestion**

The third basic technique for finding ready users is based solely upon distributed control. In its purest form, called the ALOHA protocol, each user simply transmits a newly generated packet into the channel immediately, hoping that it will be successfully received without colliding with someone else’s transmission [ABRA 70, ABRA 73]. A packet that has unfortunately collided with another packet is retransmitted after a random delay. A random delay is necessary to minimize the probability of collision with the same packet again [KLEI 73, KLEI 75]. This random delay determines the throughput-delay performance of the system and affects its stability behavior. It was found that the user population size is another important parameter that affects system stability. The stability, throughput, and delay of a slotted ALOHA system can be evaluated using a Markov chain model [LAM 74, KLEI 74, KLEI 75a]. Protocols based on contention should always be implemented in conjunction with some adaptive control algorithm to avoid unstable behavior. A variety of techniques is described in [LAM 75a, LAM 75b, GERL 77, LAM 80a]. The URN protocol described in [KLEI 78] is based on the use of asymmetric control policies instead of symmetric control policies considered in the above references. Two articles, one on the stability behavior and delay performance of ALOHA channels [KLEI 75a] and one on adaptive control schemes [LAM 75b], are reprinted below. The Exponential Binary Backoff algorithm employed in Ethernet [METC 76] is a specific instance of the class of Heuristic Retransmission Control procedures studied and described in [LAM 75b]. The article by Kleinrock and Yemini on the URN protocol [KLEI 78] is also reprinted below.

Contention protocols, in general, give rise to a random order of service. However, it is possible to favor some users over others by specifying different retransmission delays for them [LAM 80b].

### 3.3. PERFORMANCE CONSIDERATIONS

#### 3.3.1. Two Simple Criteria for Comparing Protocols

Let us compare some multiple access techniques based on the three methods described in Section 3.2 for identifying ready users.

Consider two extreme traffic conditions: (1) exactly one of the $N$ users is ready (light traffic condition) and (2) all of the $N$ users are ready (heavy traffic condition). In Table 3-1, adapted
Table 3-1. Mean delays to find a ready user under two extreme traffic conditions.

<table>
<thead>
<tr>
<th>Light Traffic Condition</th>
<th>Heavy Traffic Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>One out of $N$ users ready</td>
<td>All $N$ users ready</td>
</tr>
</tbody>
</table>

### Linear search via
- **polling**
  - $(N+1)/2$ queries (mean value)
- **time slots**
  - $(N+1)/2$ time slots (mean value)
  - 1 query
  - 1 time slot

### Binary tree search via
- **polling/probing**
  - $1 + 3/2 \log_2 N$ queries (mean value)
  - 1 time slot
  - $1 + \log_2 N$ queries
  - 1 time slot
  - $1 + \log_2 N$ time slots

### Contention in time slots
- 1 time slot
  - 2.72 time slots (mean value given optimal symmetric controls)

From [LAM 80a], we show the mean assignment delay (amount of channel time wasted) for each of the techniques to identify a ready user under the two extreme traffic conditions. Under the light traffic condition, the ready user is assumed to be chosen randomly.

In Table 3-1, the linear search techniques are self-explanatory. The tree search techniques assume a balanced binary tree with users at all leaf nodes. For the contention technique we assume a time-slotted channel. At the end of a time slot, it is globally known whether it is empty, contains a single user's transmission, or contains a collision. It is assumed that a user who has a new packet will transmit it in the next time slot with a probability of 1. A packet that has had a collision is retransmitted in the next time slot with a probability ($<1$) that is adaptively controlled. The mean delay of 2.72 time slots shown in Table 3-1 corresponds to a slotted ALOHA protocol that is optimally controlled (with a maximum channel throughput of 0.368). With nonoptimal controls, the mean assignment delay under the heavy traffic condition would be somewhat larger than 2.72 time slots (3-4 time slots would be quite realistic).

Given a large $N$, Table 3-1 shows that linear search performs well under the heavy traffic condition but performs poorly under the light traffic condition. The opposite is true for tree search and contention techniques.

Note that the actual performance of each technique in a specific system depends upon system parameters which determine the time slot size or the time duration to conduct a query. Whether one technique is better than another (over a wide range of channel utilization) depends on the specific parameter values of a particular network. Adaptive control is necessary for the contention technique to handle heavy traffic (no matter how infrequently it occurs). It is also desirable for a protocol based on tree search to...
adaptively switch to linear search whenever the channel traffic becomes heavy.

In Figure 3-2, the general behavior of the mean delay versus throughput characteristic of a multiple access protocol is shown. Note that the maximum throughput $S_{\text{max}}$ and the minimum delay $D_{\text{min}}$ (at zero throughput) in Figure 3-2 can be predicted from the entries in Table 3-1 for each multiple access protocol. Specifically, the mean assignment delay $d_1$ of a multiple access technique under the light traffic condition determines $D_{\text{min}}$ as follows

$$D_{\text{min}} = P + d_1$$

where $P$ is the mean transmission time of a user data unit. The mean assignment delay $d_2$ under the heavy traffic condition determines $S_{\text{max}}$ as follows

$$S_{\text{max}} = \frac{P}{P + d_2}$$

(Note: The above formulas may need to be adjusted slightly to account for channel propagation and carrier detection times.)

The $D_{\text{min}}$ and $S_{\text{max}}$ values that can be calculated from the entries in Table 3-1 indicate that generally the delay-throughput curves of a linear search technique and a contention technique cross each other as shown in Figure 3-3. (The comparison between linear search and tree search is similar.) The exact crossover point which determines which technique is better over a wider range of channel utilization depends on specific network parameters. A protocol that switches between contention and linear search (or between tree search and linear search) at the crossover point will get close to optimum performance.

To obtain the complete throughput-delay performance curve of a multiple access protocol requires a detailed mathematical analysis using queueing-theoretic or Markovian models. However, the $D_{\text{min}}$ and $S_{\text{max}}$ values of a protocol as calculated above are often adequate as an indication of a protocol's performance for many practical purposes.

### 3.3.2. Techniques to Enhance Performance

Suppose that data units offered by users for transmission in the broadcast channel are much longer than the duration of a time slot or a query used in the methods for identifying ready users. In this case, the throughput of the broadcast channel can be substantially increased with either the incorporation of a reservation protocol [ROBE 73, TOBA 76, CHU 78, JACO 78] or a deference protocol [CROW 73, KLEI 75b, METC 76, LAM 80c].

### Reservation

Roberts [ROBE 73] described the first reservation protocol. In his protocol, the broadcast channel is time-multiplexed into two subchannels: a subchannel for reservation requests and a subchannel for data units. The request subchannel is accessed using a contention protocol. Data units belonging to successfully broadcasted requests join a (logical) queue to await transmission in the data subchannel. The maximum throughput of this protocol is high because the maximum throughput of the data subchannel is unity. Robert's protocol inspired the class of PODA protocols for satellite packet switching described in [CHU 78, JACO 78, CHU 79].
Deference

In deference protocols, the broadcast channel alternates between search periods and transmission periods. A search period corresponds to the time taken by a multiple access protocol to find a ready user. It is terminated as soon as a ready user is identified; transmission of the ready user’s data unit(s) immediately follows. During this transmission period, all other users are required to keep quiet, i.e., to defer to the ongoing transmission. It is clear that the channel throughput is given by the ratio

\[
\frac{\text{mean transmission period}}{\text{mean search period} + \text{mean transmission period}}
\]

and can be made arbitrarily close to unity by having very long transmission periods (i.e., very large data units or maintaining very long queues at the users).

Deference can be easily implemented in a local area network environment with short propagation delays. Users are required to sense the presence of any ongoing transmission in the channel (carrier sensing) and to keep quiet until the transmission terminates. Any of the multiple access techniques can be used in search periods. However the CSMA protocol [TOBA 74, KLEI 75b, HANS 79] and the CSMA/CD protocol [METC 76], both based upon the use of contention in search periods, are the most well known. The use of collision detection in CSMA/CD improves channel throughput relative to CSMA by shortening the mean search period. It is interesting to note that if every transmission period in a CSMA/CD channel is shrunk to zero, then the CSMA/CD channel becomes an ALOHA channel.

The idea of deference actually first appeared in the R-ALOHA protocol proposed by Crowther et al. [CROW 73]. R-ALOHA was intended for a satellite channel with long propagation delays. A description of it can also be found in [LAM 83].

Adaptive hybrids

We have already encountered one other method to improve the delay-throughput performance of a broadcast channel, and that is to use a multiple access protocol that adaptively changes from one basic method to another for identifying the ready users. We refer to these protocols as adaptive hybrids. We have seen tree-search protocols that adaptively change to the use of linear search when traffic becomes heavy. The URN protocol changes from contention to linear search when traffic becomes heavy. On the other hand, the access protocol of Hyperchannel [THOR 75] uses a linear search initially (upon the termination of a transmission period). If a linear search of all users did not turn up a ready user, then the protocol switches to the use of contention (i.e., it switches to contention when traffic becomes light).

A local network access protocol described in [LAM 80b] also uses a combination of linear search and contention (but not adaptively). In this protocol, users belong to different priority classes. Upon the termination of a transmission period, the protocol performs a linear search of the classes (from high to low priority) to determine the highest priority class with ready users. If there are more than one ready user in this priority class, they compete for channel access by contention.

3.3.3. Performance Analyses

The delay-throughput performance of an
FDMA channel (fixed assigned) can be calculated using classical queueing theory results (e.g., the P-K formula for an M/G/1 queue). For a fixed-assigned TDMA channel, some slight modifications to the classical results are necessary owing to the need to observe time-slot boundaries. (See [LAM 77].)

The model of Konheim and Meister [KOHN 74] was developed for a polling system. Conceptually, their model can be used for any protocol employing a linear-search technique and round-robin scheduling. Equal traffic rates are assumed for all users.

For contention protocols, the Markov chain model developed in [LAM 74, KLEI 75a] for a slotted ALOHA system can predict not only the system's throughput-delay performance but also its stability behavior. A performance analysis of R-ALOHA was given by this author in [LAM 80c]. Performance analyses of CSMA/CD were given by Tobagi and Hunt [TOBA 80b] and this author [LAM 80a] using different analytic methods. Bux [BUX 81] presented a performance comparison of local area networks utilizing the multiple access methods described above. For their throughput-delay performance, he employed the model in [KOHN 74] for linear search, the model in [LAM 80a] for CSMA/CD, and a model of his own for the MLMA tree search technique. Both [BUX 81] and [LAM 80a] are reprinted below.

3.4. PROTOCOLS BASED ON ADVANCED NETWORKING CAPABILITIES

In recent years, many multiple access protocols have been proposed for local area networks based on the implementation of various advanced capabilities in broadcast channels.

In [MOK 79, MARK 80, ESWA 81, FRAT 83], the interconnection bus is made up of separate control wires in addition to a data channel. These control wires provide the capability to resolve access conflicts by a priority mechanism. User priorities are determined either by their addresses [MOK 79, MARK 80] or by their positions on the interconnection bus [ESWA 81, FRAT 83].

In [LI 81, LI 82], an improved protocol is proposed for token-passing on a bus. Each data unit contains a "token direction code" which enables users to detect the direction of the current token scan in the bus. This capability reduces the token-passing delays.

In the EXPRESS-NET proposal [TOBA 80c, FRAT 81], users are connected to directional inbound and outbound channels for receiving and transmitting, respectively. The channels are connected so that all signals in the outbound channel are duplicated in the inbound channel to achieve broadcast communication. The protocol provides a conflict-free, round-robin scheduling. It can guarantee users minimum bandwidth requirements and maximum packet delay constraints (which are essential for voice traffic).

References


[KLEI 75b] Kleinrock, L., and F. A. Tobagi, "Packet Switching in Radio Chan-


(* article reprinted below.)