ON SOME CHALLENGES AND DESIGN CHOICES IN AD-HOC COMMUNICATIONS

Zygmunt J. Haas
School of Electrical Engineering
Cornell University,
Ithaca, NY, 14853
e-mail: *haas@ee.cornell.edu*

Siamak Tabrizi
Air Force Materiel Command
Air Force Research Laboratory
Rome, NY 13441
e-mail: *tabrizis@rl.af.mil*

ABSTRACT

Ad-hoc *networks are network architectures that can be rapidly (ideally immediately) deployed and that do not need to rely on pre-existing infrastructure. The salient feature of this breed of networks is that they can operate in different and differing propagation and network operational conditions, which cannot be predicted during the network design stage. In this position paper, we discuss some of the challenges and choices that need to be made while designing an ad-hoc network. In particular, we address the following three issues: hierarchical vs. flat network architectures, proactive vs. reactive (on demand) routing protocols vs. a hybrid approach, and sensing-based vs. dialog-based medium access control.*

THE DESIGN CHALLENGES OF AD-HOC NETs

The topic of ad-hoc networking has received recently increased attention. This interest comes from two different directions – from the military and from the Internet community. Of course, as the communication and networking environment of these two "markets" is quite different, the requirements, and more important the expectations, of what this technology can accomplish are quite different as well. In this paper, we address only the military applications. Nevertheless, we emphasize that much of what we discuss here is either directly or with minor changes applicable to the commercial/educational environment as well.

The three main challenges in the design and operation of the ad-hoc networks stem from:

- the lack of a centralized entity,
- the possibility of rapid platform movements, and
- the fact that all the communication is carried over the wireless medium.

In "regular" cellular wireless networks, there are a number of centralized entities; e.g., the base-stations, the Mobile Switching Centers (MSC-s), and the Home Location Registry. In ad-hoc networks, since there is no preexisting infrastructure, these centralized entities do not exist. The centralized entities in the cellular networks perform the function of coordination. Thus, lack of these entities in the ad-hoc networks requires more sophisticated distributed algorithms to perform these functions. In particular, the traditional algorithms for mobility management, which rely on the HLR/VLR and the medium access control schemes, which rely on the base-station/MSC support, cannot be used here.

All communications between all network entities are carried in ad-hoc networks over the wireless medium. Of course, due to the radio communications being extremely vulnerable to propagation impairments, connectivity between network node is not guaranteed. In fact, intermittent and sporadic connectivity may be quite common. Additionally, as the wireless bandwidth is limited, its use should be minimized. Finally, as some of the mobile devices are expected to be hand-held with limited battery power, the required transmission power should be minimized as well. The last two attributes, conservation of wireless spectrum and reduction in transmission power, lead naturally to an architecture in which the transmission radius of each mobile is limited and channels assigned to mobiles are spatially reused. Consequently, since the transmission radius is much smaller than the network span, communication between two terminals may need to be relayed through intermediate nodes; i.e., multi-hop routing.

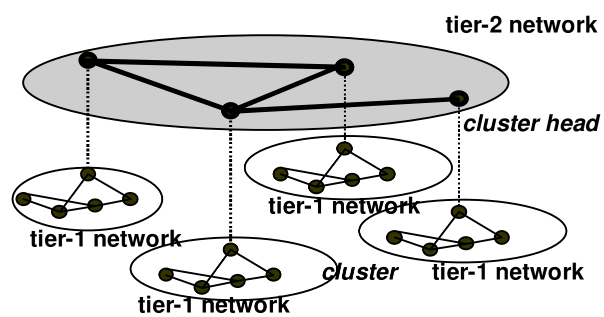

Figure 1: A Two-Tier Hierarchical Ad-Hoc

Because of the possibly rapid movement of the nodes and fast changing propagation conditions, network information, such as routing, for example, becomes quickly obsolete. This leads to frequent network reconfigurations and frequent exchanges of control information over the wireless medium. Of course, as the wireless spectrum is at premium, frequent exchanges of large amount of data over the air should to be avoided. Finally, in spite of these attributes, the design of the ad hoc networks still needs to allow for a high degree of reliability, survivability, availability, and manageability of the network.

We focus here on three fundamental choices in the design of ad-hoc networks:

- the network architecture
- the routing protocol
- the medium access control

FLAT vs. HIERARCHICAL NETWORKS

The configuration of an ad-hoc network can be either *hierarchical* or *flat.* In a hierarchical network, the network nodes are partitioned into groups called *cluster*s. Within each cluster, one node is chosen to perform the function of a *cluster head.* Routing traffic between two nodes that are in two different clusters is always through the cluster heads of the source and destination clusters. Depending on the number of hierarchies, the depth of the network can vary from a single tier to multiple tiers. A two-tier example is shown in Figure 1. In a flat ad-hoc network, all nodes are equal. Connections are established between nodes that are in close enough proximity to allow sufficient radio propagation conditions to establish connectivity. Routing between any two nodes is constrained only by the connectivity conditions and,

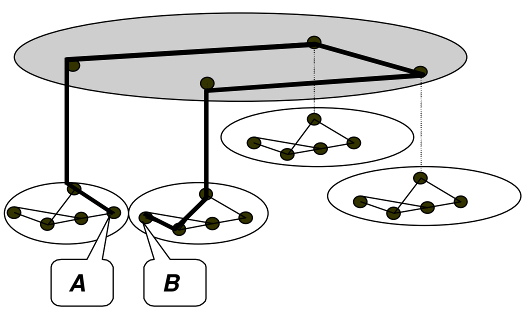

Figure 3: An example of sub-optimal routing

possibly, by security limitations. An example of a flat network is shown in Figure 2. Note that flat networks could be perceived as a special case of "zero-tier" hierarchical networks.

One of the biggest advantage of the flat networks is in the existence of multiple paths between a source and a destination. This allows to "spread" the traffic among multiple routes, reducing congestion and eliminating possible traffic "bottlenecks" in the network. Routes can be chosen to better match the specific requirements of a traffic stream. For example, low delay and low capacity paths could be used for voice traffic, while file transfer could be done over high capacity and longer delay routes. In other words, QOS-based routing is possible.

Routing in hierarchical networks is often sub-optimal. A good example is routing between two nodes that, although are assigned to two different clusters, are close enough to maintain direct connection. Such an example is depicted in Figure 3, where the traffic from node A to node B is detoured through a complicated path due to the lack of direct connectivity between the cluster heads of nodes A and B.

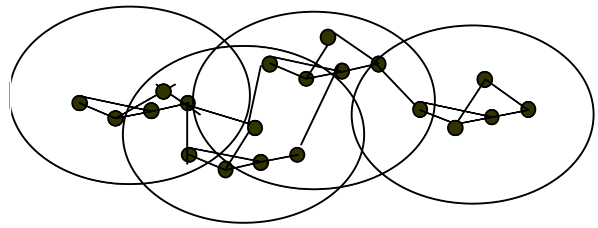

Figure 2: A Flat Ad-Hoc Network

Nodes in flat networks transmit at a significantly lower power than the transmission power of the cluster heads in the hierarchical networks. This has several implications: First, the battery power of the nodes in ad-hoc networks is preserved. Second, the wireless spectrum can be better reused, leading to more network capacity. Third, and possibly most importantly, larger degree of Low Probability of Interception / Low Probability of Detection can be achieved, resulting in a more secure network operation.

Finally, only one "type" of equipment is necessary, as all the nodes has to perform the same operation and there is no "single point of failure," created by the cluster heads. Additionally, the hierarchical networks require complex algorithms to maintain the tiers; e.g., creation and reconfiguration of the backbone network.

The main advantage of the hierarchical ad-hoc network is the ease of the mobility management process. Cluster heads can act as a databases that contain the "location" of the nodes in their own clusters. To determine the existence and the "location" of a mobile node, a query is broadcasted to all the cluster heads. The cluster, under which the node resides, responds to the query originator. This relatively simple mobility management scheme can be mimicked in the flat networks by a routing algorithm, as shown in the next section.

Many see the hierarchical networks as matching the underlying hierarchical structure of the nodes and their organization. This is especially true in the hierarchical organization of the military. However, one needs to separate the *physical* structure from the *logical* structure. In spite of the fact that in many cases the underlying *logical* structure is indeed hierarchical, there is no reason why this logical structure cannot be implemented on top of a physical flat architecture.

Our comparison above suggests that flat networks have a number of crucial advantages over the hierarchical network architectures. We see the biggest benefit of the flat networks in the high degree of LPI/LPD that they can achieved and in their high degree of survivability to failures of subsets of network components. Furthermore, the relative shortcoming of the flat networks, as compared with the hierarchical network architecture, can be overcome with a simple routing protocol, for example, the protocol described in the next section.

PROACTIVE vs. REACTIVE ROUTING

The challenge in designing a routing protocol for the ad-hoc networks communication environment stems from the fact that, on one hand, to determine a packet route, at least the reachability information of the source's neighbors needs to be known to the source node. On the other hand, in an ad-hoc network, this topology may change quite often. Furthermore, as the number of network nodes can be large, the potential number of destinations is also large, requiring large and frequent exchange of data (e.g., routes, routes updates, or routing tables) among the network nodes. Thus, the amount of update traffic is quite high. This is in contradiction with the fact that all updates in the wireless communication environment travel over the air and are costly in resources. However, an even more disappointing fact is that as the network size increases and as the nodal mobility increases, smaller and smaller fraction of this total amount of control traffic will be even used. This is so, since as the nodes become more mobile, the lifetime of a link decreases. Thus, the period in which the routing information remains valid decreases as well. In fact, it is easy to show that for any given network capacity, there exists a network size and nodal mobility for which all the network capacity will be wasted on control traffic only!

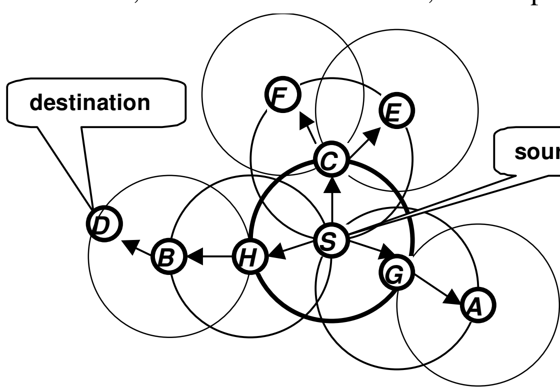

Figure 3: An example of the ZRP protocol

In general, the existing routing protocols can be classified either as *proactive* or as *reactive.* Proactive protocols attempt to continuously evaluate the routes within the network, so that when a packet needs to be forwarded, the route is already known and can be immediately used. The family of Distance-Vector protocols is an example of a proactive scheme. Reactive protocols, on the other hand, invoke a route determination procedure on demand only. Thus, when a route is needed, some sort of global search procedure is employed. The classical flood search algorithms are reactive protocols.

The advantage of the proactive schemes is that, once a route is needed, there is little delay until the route is determined. In reactive protocols, because route information may not be available at the time a route request is received, the delay to determine a route can be quite significant. Furthermore, the global search procedure of the reactive protocols requires significant control traffic. Because of this long delay and excessive control traffic, pure reactive routing protocols may not be applicable to real-time communication. However, pure proactive schemes are likewise not appropriate for the ad-hoc network environment, as they continuously use a large portion of the network capacity to keep the routing information current. And as already pointed out before, since in an ad-hoc network nodes move quite fast, most of this routing information is never used. This results again in an excessive waste of the network capacity.

A related issue is that of updates in the network topology. For a routing protocol to be efficient, changes in the network topology have to have local effect only. In other words, creation of a new link at one end of the network is an important local event but, most probably, not a significant piece of information at the other end of the network. Proactive protocols tend to distribute such topological changes widely in the network, incurring large costs.

What is needed is a protocol that, on one hand, initiates the route-determination procedure on-demand, but at limited search cost. An example of such a protocol is the *Zone Routing Protocol (ZRP)* [Haas97], which is a hybrid reactive/proactive routing protocol. On one hand, it limits the scope of the proactive procedure only to the node's local neighborhood. On the other hand, the search throughout the network, although it is global, is done by efficiently querying only selected nodes in the network, as opposed to querying all the network nodes.

ZRP operates on the premises that each node defines for itself an area around it, which is termed the node's *zone,* and which topology the node learns through a limited-scope proactive procedure. To discover routes that are further away from the node than its zone, the node sends a query to other selected nodes on the boundary of its zone. This search process continues until the destination is discovered.

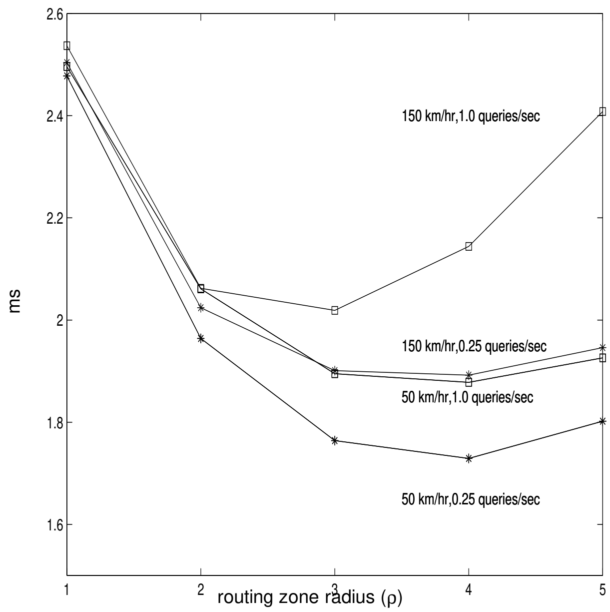

Figure 4: ZRP Route Discovery Delay

On one hand, due to the limited scope of propagation of topological changes the total amount of ZRP's control traffic is severely reduced, as compared with full proactive protocols. On the other hand, due to the destination search proceeding in quanta of the zone size, the destination discovery process is much faster than the pure reactive protocols. In most cases, an "optimal" size of the zones can be defined, which leads to minimum volume of control traffic or minimum latency of the route discovery process. (The size of a zone is referred to here as the *zone radius.)* This is depicted in Figure 4 for a number of the mobility and the frequency of route query parameters.

TO SENSE OR NOT TO SENSE …

MAC protocols have been extensively studied in the early 70's, mainly for a single shared channel applications; e.g., ALOHA, CSMA-family, and CSMA/CD. However, because of the possible large size of the ad-hoc networks, much larger than the transmission range of a single transmitter, single shared channels will not perform well in this environment. Furthermore, the "hidden terminal problem" needs to be addressed by the MAC protocol.

The ALOHA protocol is the earliest version of wireless Medium Access Control (MAC) protocol. In ALOHA, every user is allowed to transmit its outgoing data packet whenever it is ready. A collision occurs when more than two users are transmitting at the same time. The channel utilization of ALOHA is very low -- 18% -- due to the high probability of collisions, especially in the case of high network loads.

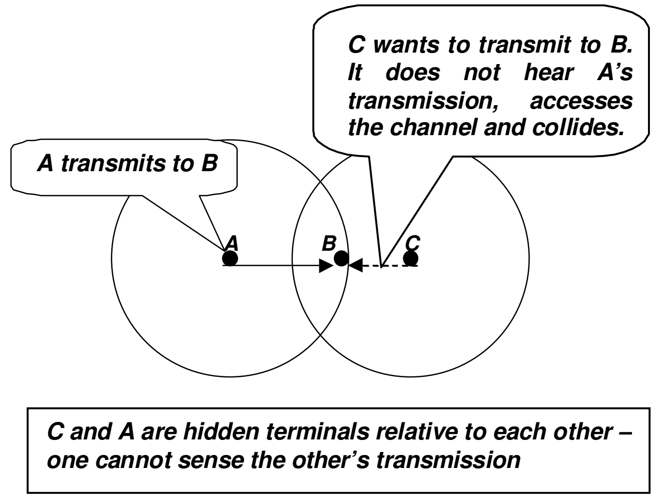

Figure 5: The *hidden terminal* problem

The Slotted ALOHA protocol achieves a channel utilization of 36% by dividing the channel into time slots. Transmission is allowed to start at the beginning of a time slot only. Slotted ALOHA maintains a higher channel utilization by avoiding collisions in the middle of the data transmission, but it introduces somewhat longer access delay and more complexity due to the need for slot synchronization. ALOHA and slotted ALOHA are particularly applicable in network environment with long propagation delay, such as satellite communication. In network with shorter end-to-end propagation delay, much better channel utilization is possible.

In the Carrier Sense Multiple Access (CSMA) protocol, every user senses the channel for carrier before it transmits. In principle, systems that use CSMA protocol have lower probability of collision, since whenever a node transmits, other users will sense the transmission and defer accessing the shared channel. Unfortunately, in the ad-hoc networks communication environment, not all users in the network can hear other users. This problem is commonly referred to as the *hidden terminal problem*, example of which is depicted in Figure 5. Thus, situations in which users' transmissions collide may be quite common, despite the use of the CSMA protocol. Furthermore, because in the CSMA protocol users sense the channel at the transmitter but collisions occur at the receiver, stations may unnecessarily defer from transmission. This is commonly referred to as the *exposed terminal problem* and is shown in Figure 6. Both, the hidden and the exposed terminal problems render the pure CSMA scheme inefficient for the ad-hoc communication environment.

Recently, two other protocols that attempt to eliminate the shortcomings of the CSMA protocol were proposed - the Multi-hop Access Collision Avoidance (MACA) [Karn90] protocol and the Media Access Protocol for Wireless LANs (MACAW) protocol [Bharghavan94]. Both of the protocols use, what is referred to as the *RTS/CTS (Request-To-Send/Clear-To-Send) Dialog*, totally abandoning the CSMA mechanism. The RTS/CTS dialog precedes the actual transmission of the data by the stations and allows to reserve the channel, so that collisions with other stations are avoided. MACAW is an improvement of MACA in the sense that it solves some of the *fairness* problems in MACA; i.e., in MACA some nodes can be denied access for a long time.

FAMA [Fullmer95] further improves the MACAW performance by including a non-persistent CDMA mechanism at the beginning of every free slot. The IEEE 802.11 wireless LAN standard employs the idea of collision avoidance mechanism through the CSMA/CA (Carrier Sense Multiple Access with

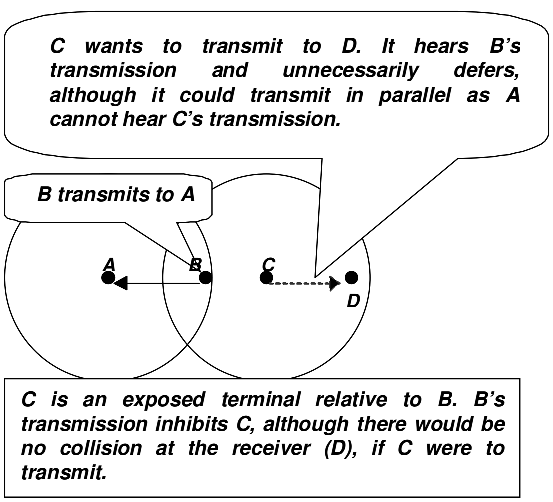

Figure 6: The *exposed terminal* problem

Collision Avoidance) protocol, which relies on immediate positive acknowledgments to the sender and possible retransmission of unacknowledged packets. The recently proposed MACA/PR (Multiple Access Collision Avoidance with Piggyback Reservations) [Lin97], uses the CSMA/CA protocol to support multimedia traffic reservations. Most of these wireless LAN protocols do not address, or insufficiently address, the central issue of a large-scale mobile network, namely that the connectivity between node may be highly unstable. Thus, for example,

most of these studies assume that interfering nodes can hear the all RTS/CTS dialogs.

However, this assumption is often invalid in a highly mobile environment. In fact, due to the inability to hear the RTS/CTS dialogs, a large portion of the network capacity is lost to collisions in data packets.

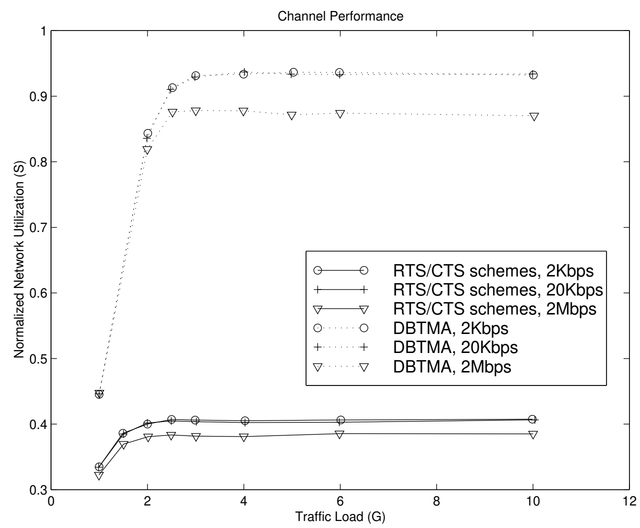

Fig. 7: Comparison of DBTMA and RTS/CTS schemes

Another incarnation of the same problem stems from the mobility aspect of the ad-hoc networks. In other words, these protocols assume that the RTS/CTS dialog is always received by all possible interfering stations. However, in a mobile network, node will migrate close to and away from communicating nodes all the time. Thus, the fact that a mobile node did not hear a RTS/CTS dialog does not indicate that the channel is, indeed, free for use. Our simulations of the RTS/CTS-based protocols clearly show that a large number of collisions occur, especially when the speed of the network nodes is substantial.

A multi-hop network architecture based on distributed cluster formation has been presented in [Gerla95]. The MAC protocol proposed for this type of networks is based on synchronous access. The network was shown to perform well for multimedia traffic. We postulate that, although theoretically possible, even limited-scope synchronization is difficult to achieve in a highly mobile RWN.

What is required is a protocol that, on one hand, resolves the collision based on the state of the channel at the receiver but provides a constant indication of the status of the channel, so that when a mobile migrates within range of the transmitting/receiving node, the mobile's transmission does not interfere with the transmission in progress. The

Dual Busy Tone Multiple Access (DBTMA) protocol [Deng98] support exactly this type of operation.

Without going into too much details, the protocol's operation is as follows. The single common channel is split into two sub-channels: a data channel and a control channel. Data packets are transmitted on the control channel, while control packets (RTS, CTS, etc) are transmitted on the control channel. Additionally, two busy tone are assigned to the control channel: BT_r (the receive busy tone, which shows that a node is receiving on the data channel) and BT_t (the transmit busy tone, which shows that a node is transmitting on the data channel).

When a node becomes ready, it senses the channel and if it does not hear the BT_r tone, it can be sure that its transmission will not interfere with reception of any other node in its vicinity. It, then transmits the RTS message and waits for the CTS message from the receiver. The receiver, upon receipt of the RTS message, checks the BT_t tone and if it is absent, the receiver knows that the transmitter's transmission will not interfere with the transmission of any other node in its vicinity. The receiver then issues the CTS and raises the BT_r to prevent any other nodes in the area to transmit. The receiver, upon receipt of the CTS message, raises the BT_t to indicate that it will transmit and to prevent any other node in its vicinity to accept a connection request. When the connection is terminated, the busy signals are turned off. As the busy tones are maintained continuously during the connection, a node migrating into the vicinity of the communicating nodes will still be able to learn the status of the channel and refrain from accessing it. As shown in the Figure 7, the DBTMA scheme significantly improves the performance of the basic RTS/CTS-based schemes by more than doubling the network capacity.

SUMMARY AND CONCLUDING REMARKS

In this position paper, we have discussed three fundamental issues in the design of ad-hoc networks: the architecture, the routing protocol, and the medium access control scheme.

We believe that because of the many of its advantages, the flat network architecture is preferred for military-type of ad-hoc networks, especially due to its high resilience to failures and its LPI/LPD features.

Due to the highly versatile communication environment, the ad-hoc networks have to be able to adapt their routing operation to the particular network operational conditions. More specifically, the degree of nodal mobility and the activity of the nodes (reflected by the frequency of the routing query process) drive the degree of proactivity vs. reactivity of the routing protocol. The *Zone Routing Protocol* is an example of such a protocol that allows to adjust its behavior dynamically by sizing the dimension of the nodes' zones.

The traditional CSMA MAC scheme were shown to be inadequate for the ad-hoc communication environment. This is due to the fact that the collisions occur at the receiver and not at the transmitting end, while the CSMA schemes test the status of the channel at the transmitter. However, it has been also claimed that the RTS/CTS-based MAC schemes do not totally solve the problems due to the large number of collisions caused by the hidden terminal problem. The hybrid scheme, termed the *Dual Busy Tone Multiple Access (DBTMA)*, provides improved channel capacity, while avoiding most of the collisions.

While the interest in the ad-hoc networks continues to grow both in the military and the commercial markets, we expect that the three basic questions raised in this paper will provide some guidance in the design choices needed to be addressed in an implementation of this type of networks. It is usually true that "one size fits all" network design strategies is inappropriate even in the "stationary" world; it is so much more true in the dynamic communication environment as the ad-hoc networks is. Thus, adaptivity in the operation of the schemes and protocols, as exemplified by the schemes discussed here, is nearly a prerequisite to any successful design of the ad-hoc communication.

REFERENCES

[Haas97] Z.J. Haas and M.R. Pearlman, "Providing Ad-Hoc Connectivity with the reconfigurable Wireless Networks," submitted for publication. See also Internet draft, *draft-haas-zone-routing-protocol-00.txt.*

[Karn90] P. Karn, "MACA - a New Channel Access Method for Packet Radio," in *ARRL/CRRL Amateur Radio 9th Computer Networking Conference,* pp.134-140, ARRL, 1990

[Bharghavan94] V. Bharghavan, A. Demers, S. Shenker, and L. Zhang, "MACAW: A Media Access Protocol for Wireless LANs," in Proc. of *ACM SIGCOMM'94.*

[Fullmer95] C.L. Fullmer and J.J. Garcia-Luna-Aceves, "Floor Acquisition Multiple Access (FAMA) for Packet-Radio Networks," in Proc. of *ACM SIGCOMM'95.*

[Lin97] C.R. Lin and M. Gerla, "Asynchronous Multimedia Multihop Wireless Networks," *IEEE INFOCOM'97,* Kobe, Japan, 7-11 April, 1997.

[Gerla95] M. Gerla and J. T-C. Tsai, "Multicluster, Mobile, Multimedia Radio Network," *ACM/Baltzer Wireless Networks Journal,* vol.1, no.3, (1995).

[Deng98] J. Deng and Z.J.Haas, "Dual Busy Tone Multiple Access (DBTMA): A New Medium Access Control for Packet Radio Networks," submitted for publication.