H^3M - A Rapidly Deployable Architecture with QOS Provisioning forWireless Networks

Edwin C. Foudriat, Research Professor, foudr_e@cs.odu.edu
Kurt Maly, Head, maly@cs.odu.edu
Stephan Olariu, Professor, olariu@cs.odu.edu
Department of Computer Science, Old Dominion University
Norfolk, VA 23529
Telephone 757-683-3915
Fax 757-683-4900

Abstract

Future wireless networks will have to be based on highly mobile systems that are self-organizing, rapidly deployable, heterogeneous, and will not rely on expensive infrastructure. Since these networks will be called upon to support real-time interactive multimedia traffic, they must be able to provide their users with adequate quality of service (QoS) guarantees. In addition, we want to consider networks with highly dynamic resource requirements, such as QoS, by multiple, dynamically changing groups.

The architecture and protocol is based upon a protocol called DRAMA.[4 - 8]. We have taken the kernel of DRAMA - operating a framed combination of TDMA and CSMA/CD in a metropolitan area with multichannel load balancing - and created a new architecture suitable to highly mobile environments. We refer to it as the Hierarchical Heterogeneous Highly Mobile network (H^3M, for short). As the name suggests, the H^3M network consists of a hierarchy of heterogeneous hosts distributed over a geographical area and linked together in a wireless communication system. At the bottom level of the hierarchy we have a cluster architecture whose connectivity and management activities are assumed by a mobile base station (MBS). In turn, the MBSs are organized into a virtual network, essentially emulating a local area network like structure. The protocol for arbitrating as to who sends what on what frequency at what time is based on a combination of TDMA and CSMA/CD subframes and frequency allocations to a cluster of nodes that allows for dynamic bandwidth allocation and which support multicasting.

Simulation and analysis have shown that H^3M is robust, scales well and provides much higher efficiency throughput than other protocols, while supporting the same degree of host mobility. Importantly, H^3M turns out to be well suited for handling on-demand multimedia communications in a heterogeneous, highly mobile environment. Multicasting is supported with minimal overhead.

1.0 Introduction

In recent years, wireless and mobile communications have seen an explo-

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35522-1_37

H. R. van As (ed.), *Telecommunication Network Intelligence*

sive growth both in terms of the number of services provided and the types of technologies that have become available. Indeed, cellular telephony, radio paging, cellular data, and even rudimentary multimedia services have become commonplace and the demand for enhanced capabilities will continue to grow into the foreseeable future [13, 14, 15, 16, 17, 18, 1, 19, 20, 21, 22]. It is anticipated that in the not-to-distant future mobile users will be able to access, while on the move, data and other services such as electronic mail, video telephony, stock market news, map services, electronic banking [17, 18, 1, 19, 20, 22].

However, the well-known cellular networks are too rigid and inflexible to adapt to situations in which rapid deployment is critical. Such is the case in disaster relief, search and rescue, law-enforcement, collaborative computing, multimedia classroom, miliary operations, interactive mission planning, and similar special purpose applications. To address the need for self-sufficiency and rapid deployment, a number of designs have been suggested including packet radio networks (PRN) [13, 24, 15] and mobile ad-hoc networks (MANET) [17, 1, 23, 24, 19]. While these architectures feature a number of desirable characteristics [1, 17], many assume that all hosts in the network are identical in processing and communication capabilities. This, in turn, imposes numerous limitations of these networks both in terms of robustness and overhead in routing messages end-to-end.

Our main contribution is motivated by the observation that in practical situation it is very unlikely that all the hosts are identical. It is very often the case that some of the hosts have much higher computational and communication capabilities than others and hence they could serve as mobile base stations (MBS). In this capacity they can serve as cluster heads and can handle the management activities inherent to the efficient operation of the network.

Our architecture, which we refer to as the Hierarchical Heterogeneous Highly Mobile network (H^3M), consists of a hierarchy of heterogeneous hosts distributed over a geographical area and linked together in a wireless communication system. At the bottom level, we have a cluster architecture whose connectivity and housekeeping is assumed by a mobile base station (MBS). In turn, the MBSs are organized into a virtual network, essentially emulating a local area network-like structure. This is a virtual structure - the actual implementation depends on the availability of resources in the network.

In section 2, we discuss the basic architecture and operations of H^3M including intra- and inter-cluster communication, framing for multimedia traffic, call setup, bandwidth management and load balancing. In section 3, we handle the cluster organization and operations and both node and cluster mobility. In section 4, we discuss performance studies for both synchronous and asynchronous traffic. Section 5 compares H^3M with other mobile proto-

cols based upon important features such as delay, jitter, routing, bandwidth management, multicasting and load management.

2.0 H³M Architecture and Operational Protocols

Each cluster operates as a quasi-independent LAN whose protocol is patterned upon the concepts first developed in DRAMA - Dynamic Resource Allocation in a Metropolitan Area network [4 - 8]. DRAMA's objectives were:

- to deliver a mix of real-time and datagram traffic to a metropolitan sized network using multiple LANs with high efficiency and robustness; and
- to provide LAN sharing of the multiple bands so that load balancing is effectively achieved over an extremely wide range of load conditions.

For H³M, we retain the basics of these objectives but modify them to support mobility and heterogeneity in the nodes:

- to deliver a mix of real-time and datagram traffic with high efficiency and robustness to metropolitan size networks using multiple clusters ;
- to provide clusters sharing the frequency and bandwidth range of the network so that load balancing is effectively achieved over an extremely wide range of load conditions; and
- to provide QoS provisioning and multicasting in real time over wide ranges of operational parameters such as receiving and transmission capabilities, traffic types, load conditions, and node mobility.

2.1 H³M Architectural Assumptions

We consider a number of nodes distributed over an area linked together in a wireless communication system. To do so we assume neighborhoods of nodes which form local clusters. Each cluster supports both inter- and intra-cluster communication. A typical configuration is shown in Figure 1. We make a number of assumptions about the nodes and clusters.

Power capability: Each node has a transmitter and at least one receiver.

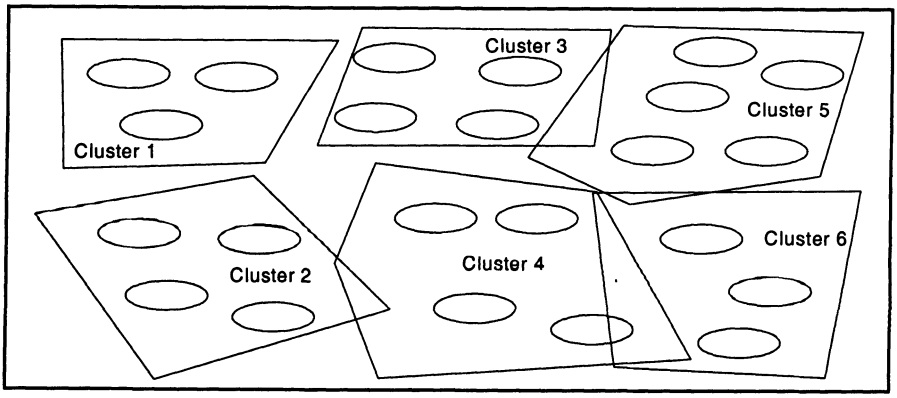

Figure 1 Nodes partitioned into a network of clusters

Transmitters may have different power capabilities. Some nodes, like those which are vehicle based, may have strong transmitter signals which carry across all clusters; others may be hand-held devices which can only reach nodes within their own cluster and possibly nodes in adjacent clusters. Each receiver is capable at any one time of tuning to any cluster's transmissions. Each has good signal-to-noise so that it can receive signals from distant clusters assuming sufficient transmitter power.

Frequency scheme: Each cluster operates on its own CDMA code (non-interfering frequency) in such a way that inter-cluster interferences are kept at an acceptable level.

Cluster basic operations: Each cluster integrates the its nodes into a local area network (LAN). It is assumed that *at least one node* in a cluster has sufficient transmitter power to be heard by all clusters in the network. This node will be designated as the *cluster sender transmitter* (CST) and will most likely be the MBS. Each node has *at least one receiver* tuned to listen to its cluster's transmissions and their exist sufficient *additional receivers* in the cluster so that the cluster is able to receive and translate transmissions from all other clusters. Nodes with additional receivers assigned to monitor outside clusters are *cluster monitor receivers* (CMR).

Mobility: Mobility is the norm rather than the exception. We make no prior assumptions about mobility. In fact, our approach supports a large range of mobility patterns.

2.2 Intra-cluster operational protocol

A frame, shown in Figure 2, has three subframes - sync header, a TDMA and a CSMA/CD subframe - which are shown in Figures 3a, 3b and 3c, respectively. Frame time is assumed to be of the order of 10 msec. The header subframe, Figure 3a, sent by the MBS, provides cluster control information needed by nodes and other clusters and possibly information about subsequent frames if a change is in order. The TDMA subframe is time slotted. It provides support for real-time information using the fact that each call (*circuit*) has a reserved slot. A major feature concerning *circuits* is that much of the wasted call capacity is recovered for those which are silent since a silent call packets contains only 1 or 2 bytes needed to maintain the resource. The subframe boundary between the TDMA and CSMA/CD subframes is dynamic. At the end of the TDMA frame, the remainder of the frame is devoted to asynchronous data using a CSMA/CD protocol similar in operation to the well known Ethernet.

As the core of the H^3M architecture, like DRAMA [4 - 8], the cluster protocol is based upon framing. To work for the entire network that might

span an area of many kilometers, each cluster has a non-interfering frequency and bandwidth. (see details in Section 2.4). Nodes use this frequency for all intra-cluster communications and, as discussed in Section 2.3, to send information over the whole network. The bandwidth at which the cluster sends is allocated dynamically in order to provide load balancing QoS (see Section 2.4). In both its TDMA and CSMA/CD subframes, signal overlap from multiple senders will corrupt the information. During the TDMA subframe, the next node must not start sending until the prior node's signal has passed. During the CSMA/CD subframe, a collision event can occur up to the cluster's round trip transit.

2.3 Inter-cluster operational protocol - Communication between clusters

Based upon the architectural features in section 2.1 and the intra-cluster protocol above, we derive procedures which allow a node to communicate with any other node in any other cluster. Three situations occur.

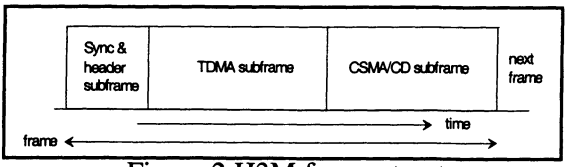

Figure 2 H3M frame structure

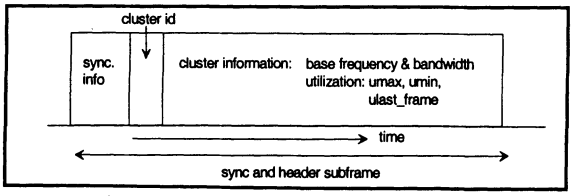

Figure 3a. Sync and header subframe structure - provides cluster operational informa-

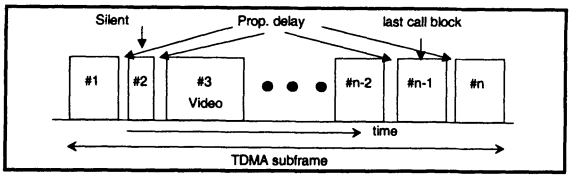

Figure 3b. TDMA subframe structure

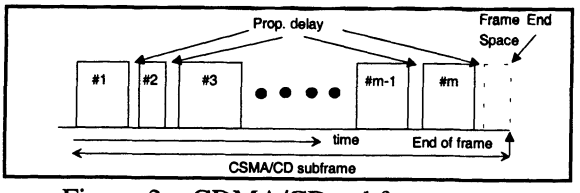

Figure 3c. CDMA/CD subframe structure

Single receiver case: A receiving node does not have sufficient receivers to be able to tune to the sending node's frequency. In this case, the receiving cluster's *cluster monitor receiver* (CMR) which monitors the sender's cluster transmissions must receive the information and retransmit it in a packet in the cluster's frequency band in a succeeding frame so that it is forwarded to the receiver.

Low-power-sender case: A sending node does not have sufficient power to be heard by its

receiving node. In this case, the sending cluster's *cluster sending transmitter* (CST) must copy the sender node's information and retransmit it at the cluster's frequency during a subsequent frame so that it can be heard by the receiver node or receiver cluster's CMR.

Best case: The communicating node has sufficient transmitting power so that information sent can be received directly and the receiving node has sufficient receivers that it can tune to the other's cluster frequency. In this case, no forwarding protocol is needed as is for the first two cases.

Note that the first two cases can be additive, i.e., both situations can exist simultaneously or each can occur separately. In a full duplex circuit for each direction different cases can occur. Note that the data transfers rates between sender and receiver are not seriously delayed by operating under other than the best case scenario. For example in the low-power-sender case using the TDMA subframe, the sending node transmits its information in one packet and the CST then retransmits this information in a subsequent frame. The minimum forwarding delay is the one frame in the best case and 3 frames under the condition where both sender and receiver cases exist. It is also true that the potential information communication rate of a cluster is reduced in situations other than the best case. We will discuss and compare performance features of H^3M in Section 4 of the paper.

Establishing and assigning connections - call setup protocols

With both TDMA and CSMA/CD subframes available, H^3M is able to provide a wide range of services. It is necessary to establish procedures and commit resources so that connections operate efficiently. Three connection types are reasonable. The first are *circuits (calls)* which use the TDMA subframe. They enable time-critical - typically voice, video and control - data transfer. The second are datagrams which consist of multiple packets. They use CSMA/CD subframes but in order to conserve resources may use setup to enable decisions based upon transmitter and receiver cases availability as described above. We designate multiple-packet datagram handling as *datagram circuits.* The final connection type is single-packet *datagrams.* It uses the CSMA/CD subframe also. *Datagrams* are used as the protocol to establish the others.

The protocol steps for *datagrams*:

1. The sender places the packet.
2. It is copied and retransmitted by the sender's leader in a subsequent CSMA/CD subframe.
3. The packet is received by the receiving node's CMR and delivered in a subsequent receiving cluster's CSMA/CD subframe.

Steps 1 - 3 are followed for all *datagrams.*

4.Based upon the cases above, two alternative possibilities exist:
   a. Not low power sender - the sender's transmitter power is sufficient for the monitoring node in receiving cluster to hear. In this case, the CST does not transmit, i.e., skips step 2 above and completes step 3.
   b. Not single receiver - the receiving node may have a receiver tuned to the sending cluster's frequency. In this case it is able to receive the packet directly from the sender's signal as a result of step 1 or 2 so the CMR can ignore in step 3.

While not part of the *datagram* delivery procedure, the receiver is now aware of the situation under which he is capable of receiving additional information from the sender. If he receives the information as a result of steps 1 - 3 and cannot dedicate a receiver to monitor the sender's subsequent transmissions, he knows both sender's CST and 'CMR must participate. If 4a, only the sender's CST need not participate; if 4b, then the CMR need not participate. This information is vital for all participating in the setting up of *circuits* and *datagram-circuits*.

Establishing *datagram circuits* and *circuits* requires the commitment of resources over an extended period of time. Based on the *datagram* delivery procedure above *datagram circuits* and *circuits* that do not have the best case scenario require the CST and/or the CMR to participate for subsequent delivery. In addition, the *circuit* requires the commitment of TDMA subframe block(s) in one or both clusters.

The protocol steps for both circuit types, using the single packet datagram as the handler, are:

1.Steps 1 - 4 above are accomplished. In the call setup, packet information is placed by the CMR whether he was able to read the sender's packet directly or its CST retransmission.
2.The receiver must now decide:
   a. whether to accept or reject the call; and
   b. whether he can commit a receiver.
   c. if he accepts in a and not b when establishing a *circuit*, he must also reserve a TDMA subframe block in his cluster.
3.With these decisions made, he sends a *datagram* packet acknowledge in return. It contains information about the receiver and his cluster's commitment, and the sender's CST requirement.
4.Return packet follows the same procedure as steps 1 above. Note, the return packet may be used to establish the return half of a duplex circuit, thus creating a full-duplex circuit.
5.As in step 2 above, the sender must now reserve and commit his necessary resources based upon how the return packet is to be handled. In addition, based upon information in the return packet, the potential

participants are able to establish how subsequent packets must be handled base upon the commitment of CSTs and CMRs and that the necessary resources have been committed.

The steps provide *datagram circuits* and *circuits* with the optimum commitment of resources and hence provide maximum efficiency under heterogeneous conditions. Note, it is not necessary to use a full duplex connection in either *datagram circuits* or *circuits* and each direction may use different resources.

2.4 Bandwidth management and load balancing

Over an extended period, some clusters are very busy while others may under utilize their capacity. In DRAMA [4 - 8], a mechanism existed where capacity can be reassigned by allocating additional frequency bands as needed. This system required each node to have as many transmitters and receivers as bands available. For heterogeneous mobile wireless-based systems using multiple bands does not appear to be practical. Some nodes may have a wealth of electronic gear connected with their mission; others may have a simple digital-based T/R which provides only simple communication.

A solution whereby a single T/R unit is used while enabling capacity redistribution is suggested here. Both transmitters and receivers may be tuned to any frequency within the overall network frequency range. Further, each transmitter and receiver is built whereby its bandwidth size can be changed at least in steps. For example, a receiver would have an itermediate frequency bandwidth capable of change and a transmitter could modulate over a smaller or larger bandwidth depending upon its band pass filters. With this frequency structure, the system becomes a variable-bandwidth, variable-frequency wireless system (VB/VFWS).

Assuming VB/VFWS units can be built, the major operational problem is to partition the total frequency range so that capacity changes have the minimum impact on other clusters, and to provide a load balancing system able to allocate bandwidth as needed to the cluster as needed and resolve priorities when bandwidth is scarce. To change capacity in increments, a cluster most likely needs a free frequency range either directly above or below its present range or a total reassignment and readjustment of all frequency bands without interference and signal corruption.. This, in turn, requires cooperation between clusters as to what frequency and bandwidth each uses. Methods for the allocation of frequency and bandwidth and their performance they provide will be discussed in a later paper, but the basic load balancing techniques were proven in reference [7-8]. We assume that frequency and bandwidth changes can be made in *O(1 usec)*, so that changes can be made rapidly and successfully between frames.

We pose the question of how each receiver listening to the cluster may be

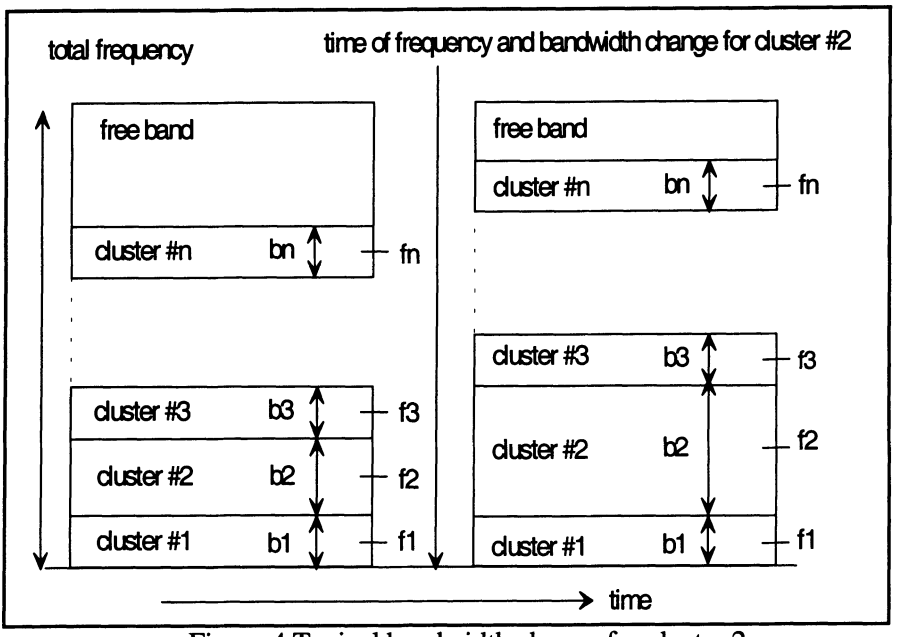

Figure 4 Typical bandwidth change for cluster 2

made aware of change, since the change must be *non-interferingly* implemented for all receivers in the network. One solution appears to be signaling during the synchronizing signal subframe block at the beginning of each frame as noted in Figure 3a. The MBS would indicate changes in frequency and bandwidth for some succeeding frame. Each receiver and transmitter, after detecting a cluster signal change in the sync subframe, would retune at the appropriate frame time. We illustrate this change in bandwidth in Figure 4. Assuming load changes would occur in *O(1 min.),* the VB/VFWS should easily be able to maintain clusters with well balanced performance over a wide range of loading conditions.

3.0 Cluster and network operations

In this section, we discuss how clusters and the network operate to provide high mobility and dynamic load balancing. Although we discuss mechanisms for cluster and network operations, in many cases there may be additional operational requirements based in the actual utilization which may dictate and/or override the procedures we discuss. Therefore, the discussion should be used to understand what features exist, how they are implemented and how they may be used in H^3M when these type activities take place in a real-world environment.

3.1 Cluster organization

The only low-level requirement for cluster organization is based upon the affect of transit time over the distance between and spanned by its nodes and

possibly maximum transmitter signal strength of a node. Cluster dimensions should be *O(1 km)*. The actual decision on cluster organization should be done in-situ based upon distance, frequency, bandwidth, terrain, the number of nodes accommodated, and intra- and inter-node higher level activity, etc. Note that cluster can overlap in span since they operate at non-interfering frequencies. It is assumed that the network will operate over in line-of-sight frequency ranges. Network dimensions are *O(10 km)*.

As described in DRAMA, the TDMA subframe supports circuit-type data. Each member of the cluster is aware that *n* calls exist and know that after the nth call, the CSMA/CD subframe begins. The CSMA/CD subframe is similar to Ethernet operations. Major exceptions are that due to framing, packets arriving during the TDMA frame are delayed until the start of the CSMA/CD subframe and after the mth packet has been transmitted there may be wasted capacity if no packet is able to fit into the space before the frame ends.

3.2 Cluster formation

The operational features needed to form a new cluster are:

1. an MBS node for control and a CST node with sufficient transmitter power to carry across the network;
2. nodes with sufficient additional receivers to act as CMRs to support intercluster reception;
3. cluster dimensions should be *O(1 km)*.
4. an unused frequency and minimal bandwidth within the overall network environment;
5. very probably, a need for some activity unrelated to H^{3}M's architecture or protocol; and
6. a mechanism for interested nodes compare theses activities so as to decide to join.

Because of 5, we provide only a initialization protocol based upon the apriori fact that nodes[1] have used 6 to determine that they need to disconnect from their present cluster and form a new one. To simplify the protocol, we assume that the potential MBS, CST and CMRs, as in 1 and 2 are identified and that they have successfully terminated any *circuits*. We assume nodes satisfy 3. The first step is to select identify the MBS and the resources for 4. The MBS then notifies all existing clusters concerning the new cluster. Upon

[1]Closely spaced nodes do not have to be in the same cluster since clusters can physically overlap without interference.

[2]We do not treat the condition of a new node joining the network because of the diversity of situations under which networks may operate and different solutions exist for different environments. One example is to use low powered beacon signals where clusters advertise their frequencies.

completion of 4 and notice that all clusters are prepared to receive the new cluster, the MBS terminates is residence in his present node and begins frames in the new cluster. Upon receiving this message, all CST and CMR nodes terminate and join the new cluster. At this point, the new cluster is fully operational such that nodes with existing calls can now join using the protocol for node movement (section 3.3). The need for the new MBS, CST and CMRs to terminate all calls should not seriously hamper critical communications since the new cluster can be up and running in very few frames after termination. It begins as a lightly loaded cluster such that resurrecting terminated calls should require minimal delay. All other nodes joining can now do so without any call interruptions.

3.3 Node movement[2]

The situation involving mobile nodes can be very dynamic. Nodes may move to a location where it can no longer work effectively within its present cluster or a group of nodes may form a cluster as above. Note, clusters can overlap physically, since cluster organization is primarily dictated by distance. We have not included reorganization due to cluster size or load since an H^3M cluster can operate under an extremely wide range of load using load balancing technique discussed in section 2.4. These situations must be resolved without serious cluster and network performance degradation. Let us take up the situations seriatim.

Single node movement: Where a node moves, the node needs to join another cluster in which it can operate efficiently. The first step is to determine into which cluster the node should move.[3] The node can then determine its new cluster operational features such as node id and band by using the information available in the cluster's header frame, Figure 3a. Cluster activity unrelated to protocol or architecture is outside H^3M's scope but needs to be done based upon step 7 of Section 3.2. The node should then transfer its active calls as follows:

1. For outgoing *circuits*, once a new cluster has been selected, the node should reserve resources in the new cluster before the switch is completed. Again, the exact protocol, including when, is a matter of further study. However, since any resource reservation in the new cluster can be maintained by silent call block, the capacity cost of reserving resources should be minimal. Further, the cluster to which the node is moving has the option of obtaining additional capacity so it should be able to anticipate future load changes very effectively. At the point of

[3]New cluster selection requires knowledge of the location of clusters in direction the node is headed. Since our protocol does not use routing, alternative source of location information is needed. Since location is needed by many operations we assume that it can be made available without addition to the protocol.

the actual switch, the node can issue "final" call blocks in the TDMA frame for the cluster from which it is departing, retuned it transmitter and receiver and joins the new cluster at its next frame.

2. For incoming and outgoing *datagram circuits*, no resource reservation is necessary. It can inform the new cluster of its present CST and CMR activity; that is which outgoing packets need to be retransmitted and which incoming packets need to be copied, translated and retransmitted because the node does not have receivers to commit.
3. The node then needs to transfer and multicast an "*I am here*". Since the actual transition should be *O(10 msec.)*, the likelihood of missing critical data is minimal. Further, reliability of information transfer should be handled at the transport level so duplicate packets can be send for any information destroyed or lost. If additional time is needed for the node to get established, then for critical messages, the node could use full-duplex circuits and an XON-XOFF protocol to span the transfer period.

Cluster breakup: The breakup of a cooperative-based cluster is accomplished in much the same manner as the movement of a single node. However, the first step for those nodes which are to break off is a decision whether to form a new cluster or to join an existing cluster already covering the area. Once this decision is made, the nodes can use the procedures outlined above to make connections.

4.0 Performance of H^{3}M

In the past, extensive analysis and simulation have demonstrated the feasibility and performance capability of DRAMA [4-8]. As a first step to be assured that H^{3}M can be provide similar performance in an extended environment, it is necessary to extend the analysis of DRAMA into the areas where differences occur. One major difference is that additional loads occur in H^{3}M due to repetition of TDMA and CSMA/CD packets needed to forward information to other clusters. Secondly, H^{3}M uses VB/VFWS to obtain load balancing whereas DRAMA allocated or released channels with fixed bandwidth. Factors affecting capacity and delay for both the TDMA and the CSMA/CD subframes under varying bandwidth conditions were performed. Details of the analysis are presented in Appendix A and B.

TDMA subframe performance for a fixed set of circuit conditions is completely predictable. The time for each *circuit's* information to be delivered from sender to receiver depends only on the repetitive transfers needed, as seen in Figures 2 and 3. Load changes occur due to different mix of circuit

[4] Actually, video circuits should be more efficient than voice especially for larger bandwidths and diameters since the propagation delay is needed only after the *circuit* frame transmission is complete.

Voice circuits in TDMA sub-frame				
Distance 1 km		Number of voice circuits in TDMA		
Bit rate, x 10+6	Round trip prop, bits	40% of frame	60 % of frame	80% of frame
1	10	11	18	24
5	50	59	88	118
10	100	112	169	226
20	200	206	310	413
Distance 2 km		Number of voice circuits in TDMA		
1	20	11	17	24
5	100	56	84	112
10	200	103	154	206
20	400	176	264	352
Distance 5 km		Number of voice circuits in TDMA		
1	50	11	17	23
5	250	49	74	98
10	500	81	123	164
20	1,000	122	183	244

Table 1. Voice circuits for differing frame utilization, different bit rates and cluster distances

types (voice, video, real-time control, etc.) and include factors such as intra- and inter-cluster calls, call silent periods for each type, maximum and minimum bandwidth needed for each type, and the propagation time needed for each call block to pass the next call location. For example, video *circuits* do not have silent periods as do voice *circuits*. The subframe termination boundary occurs after all nodes have had a chance to send their *circuit* calls without overlap. All *circuits* (including repeated) are numbered and are sent in order; when a *circuit* terminates all call numbers above are decremented by 1.

The ability of the system to handle node *circuit* load can be studied using the analysis in Appendix A. Maximum allocated TDMA load is considered to

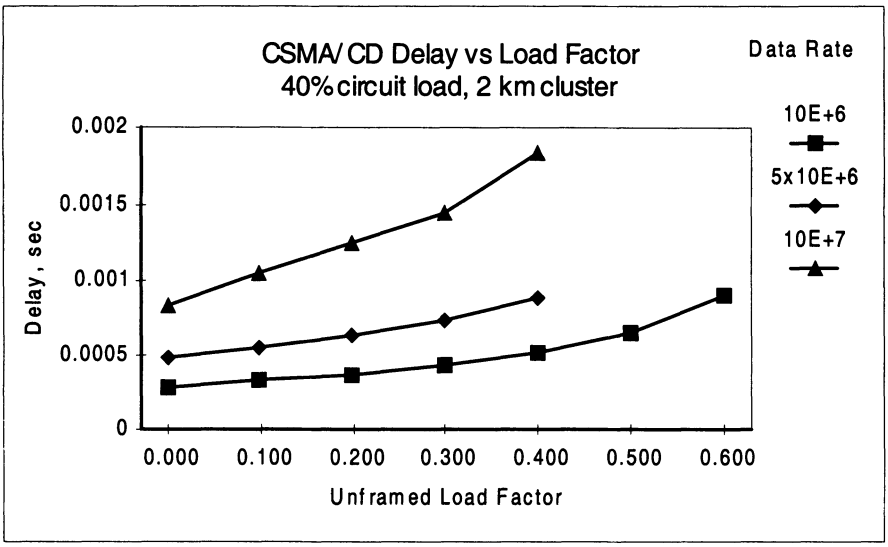

Figure 5 H³M CSMA/CD sub-frame performance for a wide range of bandwidths

be 95% of frame time. Assuming 50% of the calls are silent, this provides an minimum CSMA/CD subframe of approximately 50 - 55% of frame time depending upon the sync subframe size. This still accommodates a reasonable number of single *datagrams* and *datagram-circuits*.

We have used equation A1) to determine the number of voice circuits available under various conditions of bandwidth and cluster diameter. We assume one call per node. These results are shown in Table 1. They demonstrate that H³M can support a wide range of nodes over differing cluster distances.. Further noting that each video circuit is equivalent to ~7 voice circuits for the assumptions made[4], we can provide an estimate of the number of video circuits which can be adequately accommodated in a single cluster. By adjusting bandwidth, H³M can support a node count in a single cluster from very small to extremely large with out operational or performance deterioration. Hence, it scales well with regard to both node size and cluster diameter.

Considerable analytical and simulation work was conducted to document the CSMA/CD performance for DRAMA [4-5]. The early work [4] plus work by others [9, 25] examined the effect of framing on CSMA/CD while the later work [5] showed that when a LAN uses multiple channels, it is more effective to segregate the synchronous and asynchronous data onto separate channels. H³M does not have this option available, so we needed to extend the original work to verify that changing bandwidth and sending frequencies is a feasible alternative method for load balancing. In Appendix B, we de-

scribe a study that extrapolates the original work to H³M.

Figure 5, the result of that study, shows the delay performance for CSMA/CD subframe packets for a typical system with 2 km cluster diameter, 40% TDMA *circuit* load and a bandwidth range factor of 10. Figure 5 uses the unframed load for its load factor, i.e., the load occurring over the 10^{-2} sec. frame length assuming no TDMA *circuit* load. An effective CSMA/CD load, λ_{eff}, can be considered. It is:

$$\lambda_{eff} = (1 - t_{TDMA}/T)\lambda \qquad B3)$$

where the variables are defined using equation A1). For the situation shown in Figure 5, the effective load is 1.667 of the unframed load. Thus, at 40% CSMA/CD load factor, the system is supporting a 68% load on the CSMA/CD subframe.

One additional feature for the CSMA/CD subframe performance should be noted. We suspect that by working with the delay factors to reduce the effect of collisions at the beginning of the CSMA/CD subframe, it should be possible to support larger load factors more effectively. This should widen the range of bandwidth and cluster diameter conditions which H³M provides.

5.0 Comparison of H³M with other Mobile Communication Protocols (MCP)

To put H³M in perspective, we need to compare it to other MCPs [1, 10-12, 17, 19, 22] . We do so by discussing areas where features differ significantly..

Routing - H³M does not need routing tables. It is only necessary that cluster CMRs to be aware of all members residing in that cluster and making the necessary copies, translations and retransmissions. Copy, reformat and retransmit requirements are established at *circuit* setup time. In comparison, for most MCPs, routing tables are required and must be maintained at either global or quasi-global (hierarchically) level and must be periodically updated to enable inter-cluster communications [10]. For example, in reference [10], routing overhead can be as high as 500 kbps for in-cluster updating and 18.3 kbps for out-of-cluster updating. During large node movements, both route updates and data packet loss may occur. This significantly reduces throughput and it may take a significant time to recover.

Delay and jitter - These are hard to quantify in MCPs because of the various ways in which packet access and routing are provided, the effects of hopping and secondary factors relates to transmission lost and retransmission. Further, the importance of delay and jitter depends upon the type of information being transferred.

Delay - in H³M *circuit* hop count ranges from 0 to a maximum of 3 regardless of cluster or distance. Delays are predictable within *O(10 msec.)* at setup time and do not change with mobility. For many MCPs, hop count is

unknown until call setup and may change due to loading. In reference [10], for example, the average hop count is 10 and ranges from 4 to as much as 11. This creates average delays ranging from 100 msec. to 1 sec. depending upon routing. Further conditions such as mobility, varying loads, changed circuit bandwidth needs, etc. over the *circuit's* life span may change, causing extreme different delay in predicting delay in many MCPs.

Jitter - for *circuits* in H^{3}M, this factor occurs mainly due to call active/silent activity since dynamic subframing is used to recover capacity. While this causes more jitter than when the fixed circuit time blocks are used as in many MCPs, the jitter is still within subframe times and its nature is highly regular because it is manly due to call silent/active changes and call admission and terminations.

Multicasting - multicasting causes no difficulty in H^{3}M. Since a cluster's CMR monitors all packets from a sender's cluster, it only needs to recognize that a particular packet is destined for a node in its cluster; then make the necessary copy, translation and retransmission. For multicasting, special addressing is needed but can easily be resolved during protocol setup or dynamically during cast setup. Further, there is no additional capacity costs and routing issues found in other MCPs.

Bandwidth and capacity utilization - H^{3}M has a number of these features which makes it superior to other MCPs. First, with dynamic framing, recovery of silent *circuit* activity is easily accomplished. Second, without routing (see above) there is no need to use capacity to propagate routing tables. Further as noted above, routing problems during high node movement may cause packet loss further diminishing utilization.. Third, hopping in H^{3}M is minimized such that clusters are not required to transfer packets in which they have no interest. Fourth, multicasting does not impact capacity as it does in MCPs where clusters must propagate packets as in hopping. Fifth, node movement is accomplished with only a small increase in reserved but unused resources. Finally, load balancing (see below) provides an excellent and effective method of sharing capacity between clusters so that loss due to loading effects is significantly ameliorated.

Load balancing and fairness - dynamic load changes are probably the most difficult situation to handle in many MCPs. The problem's causes are varied and diverse. In most cases, they must be handled by alternative routing; this complicates call handling, network control and management significantly. In H^{3}M, dynamic load balancing is a fundamental feature of its design and can be readily implemented using VB/VFWS. Further, as investigated in past DRAMA studies, it is very effective and can be implemented in a time frame that alleviates most situations under which load changes normally occur.

6.0 Summary of results

The items in section 5 demonstrate that H³M provides a QoS which is superior to many MCPs. The features provide support for real-time, multimedia traffic in a wireless network that is highly mobile, rapidly deployable and effectively integrates heterogeneous nodes without an expensive infrastructure.

We have adapted the best features developed in DRAMA to wireless networks. We have devised techniques for inter-cluster data transfer which are both efficient and easily implemented and have significantly extended the DRAMA concepts by using variable bandwidth/variable frequency (VB/VFWS) equipment. These techniques significantly reduce the number of receivers which nodes must have and make it feasible for simple nodes with only a single T/R to participate in the network; they support highly predictable QoS features that support multimedia in data transfer environment that is most suited for each media data type.

VB/VFWS enables adoption of multiple channel concepts first developed for DRAMA to be adapted for load balancing in H³M. Load balancing enables fairness for data traffic when vastly different loading and loading types occur in clusters over the network. This feature easily extends QoS to a type of traffic which has been difficult to service in many networks, wireless or not. Finally, we have extended the analysis and simulation work in DRAMA to H³M to show that the latter system can be effective and to reinforce the understanding of its limitations.

References

1. Lin, R.; Gerla, M.: "Adaptive Clustering for Mobile Wireless Networks"; *IEEE Jour. On Selected Areas in Comm.*; Vol 15; No. 7; Sept 1997; pp. 1265 - 1275.
2. Nakamura, H.; Tsuboya, M.N.; Nakajima, A.: "Applying ATM to mobile Infrastructure Networks"; *IEEE Comm. Mag.*; Jan 1998; pp. 66-73.
3. Zhou, S.; Senevirtne, A.; Percival, T.: "An Efficient Location Management Scheme for Hybrid Wireless"; Faculty of Engineering, Un. Of Tech, Sidney.
4. Sharrock, S.M.; Maly, K.J.; Du, H.C.; Ghanta, S.: " A CSMA/CD-Based, Integrated Voice/Data Protocol with Dynamic Channel Allocation," Tech. Rpt.# TR-86-009; Computer Science Dept.; Un. of Minn.; Minneapolis, MN; July 22, 1986
5. Sharrock, S.M.; Maly, K.J.; Ghanta, S.; Du, H.C.: "A Framed, Moveable-Boundary Protocol for Integrated Voice/Data in a LAN," *Proc. Of SIGCOM ' 86*; August 1986, #9.
6. Sharrock, S.M.; Maly, K.J.; Ghanta, S.; Du, H.C.: "A Broadband Integrated Voice/Data/Video Network of Multiple LANS with Dynamic Bandwidth Partitioning," *Proc. INFOCOM '87*; March 1987; pp 417-425.
7. Maly, K.; Overstreet, C.M.; Qui, X.; Tang, D.: "Dynamic Resource Allocation in a Metropolitan Are Network," Tech. Rpt.# TR-88-03; Computer Science Dept.; Old Dominion University; Norfolk, VA 23529-0162; Feb. 16, 1988.

8. Maly, K.; Overstreet, C.M.; Qui, X.; Tang, D.: "Dynamic Bandwidth Allocation in a Network," *Proc. of SIGCOM'88*;
9. Metcalfe, R.M.; Boggs, D.R.: "Ethernet: Distributed Packet Switching of Local Computer Networks," *Communications of the ACM*; vol 19; July 1976.
10. Iwara, A.; Chiang, C.; Pei, G.; Gerla, M.; Chen, T.: "Scalable Routing Strategies for Ad Hoc Wireless Networks," *IEEE Jour. On Selected Areas in Communications*; Vol 17, No. 8; August 1999; pp 1369 - 1379.
11. Jubin, J.; Tornow, J.D.: "The DARPA packet radion network protocols," *Proce. IEEE*; Jam 1987
12. Shacham, N.; Craighill, E.J.; Poggio, A.A.: "Speech transport in packet-radio networks with mobile nodes," *IEEE J. Select. Areas of Commun.*; Dec 1`983; pp 1084-1097.
13. N. Abramson, Ed., *Multiple Access Communications: Foundations for Emerging Technologies,* IEEE Press, New York, 1993.
14 D. J. Baker, "Data/voice communication over a multihop, mobile, high frequency network", *Proc. MILCOM'97*; Monterey, CA, 1997, 339-343.
15. D.Bertzekas and R. Gallager, *Data Networks,* Second Edition, Prentice-Hall, 1992.
16. W. C. Fifer and F. J. Bruno, "Low cost packet radio," *Proceedings of the IEEE,* 75, (1987), 33-42.
17. M. Gerla and T.-C. Tsai, "Multicluster, mobile, multimedia radio network," *Wireless Networks,* 1, (1995), 255-265.
18. E. P. Harris and K. W. Warren, "Low power technologies: a system perspective," *Proc. 3-rd International Workshop on Multimedia Communications,* Princeton, 1996.
19. R. Ramanathan and M. Steenstrup, "Hierarchically-organized, multihop wireless networks for quality-of-service support," *Mobile Networks and Applications,* 3, (1998), 101-119.
20. D. Raychaudhuri and N. D. Wilson, "ATM-based transport architecture for multiservice wireless PCN," *IEEE Journal of Selected Areas in Communications*, 12, (1994), 1401-1414.
21. R. Sanchez, J. Evans, and G. Minden, "Networking on the battlefield: challenges in highly dynamic multihop wireless networks," *Proc. of IEEE MILCOM'99,* Atlantic City, NJ, October 1999.
22. J. E. Wieseithier, G. D. Nguyen, and A. Ephremides, "Multicasting in energy-limited ad-hoc wireless networks," *Proc. MIL COM'98,* 1998.
23. W. Mangione-Smith and P. S. Ghang, "A low power medium access control protocol for portable multimedia devices," Proc. Third International Workshop on Mobile Multimedia Communications, Princeton, NJ, September 1996.
24. M. Joa-Ng and I.-T. Lu, "A peer-to-peer zone-based two-level link state routing for mobile ad-hoc networks," *IEEE Journal of Selected Areas in Communications,* 17, (1999), 1415-1425.
25. Kleinrock, L and Tobagi, F.A.; "Packet Switching in Radio Channels: Part 1," *IEEE Transactions on Communications*, COM-23; 12, Dec. 75; 1400-1416.

Appendix A: H³M TDMA analysis

TDMA subframe performance is predictable. The time for each *circuit's* information to be delivered from sender to receiver depends to a large extent on the repetitive transfers used. The TDMA block for the *circuit* is based its count number which decrements by 1 for each prior count *circuit* that terminates. The *circuit* access time it determined by it count and the number of bytes each prior *circuit* has to send and the propagation delay between each prior *circuit* node. While there are many variables which can be random over time, most will not vary greatly between frames. What is more important than frame delay and jitter is how many *circuits* can be supported. This relates to the number of nodes a cluster can support assuming each node is talking on a single *circuit* at any one time.

The frame structure is shown in Figure 2 and the TDMA subframe structure is shown in Figure 3b. The subframe termination boundary occurs after all nodes have had a chance to send their *circuit* calls without overlap. Let:

η_υ = number of voice circuits including repeated calls

η_τ = number of video circuits including repeated calls

p = percentage of silent voice circuits

d = cluster diameter, km

β = data rate, bps

b_υ = circuit voice bit rate per frame (32K bps) = 320 bpf

b_τ = circuit video bit rate per frame (200K bps) = 2K bpf

b_s = circuit voice silent bit rate = 16 bits

δ = round trip propagation = $2*5d\beta$ - signal travel time is assumed as 5μsec/km

T = frame time = 10 msec.

b_f = frame sync block size = 128 bits

Then for a TDMA sub-frame:

δ_d = prop.delay in bits = $(\eta_\upsilon + \eta_\tau)\delta/3$ – assume average node separation 1/3 d

δ_d = voice circuit delay in bits = $RU[(1-p)\eta_\upsilon]b_\upsilon + RD[p\eta_\upsilon]b_s$

δ_d = video circuit delay in bits = $\eta_\tau b_\tau$

and

τ_{TMDA} = TDMA subframe time = $(b_f + \delta_d + \delta_d + \delta_d)$ A1)

For the analysis, *RU* and *RD* designate round upper and lower respectively since partial voice circuits are not realistic. Repeated calls, both incoming/outgoing and voice/video, use the same load resources as to local cluster calls so they can be treated in an identical manner. Further, it is assumed video circuits do not have silent periods as do voice circuits.

By solving for η_υ assuming $\eta_\tau = 0$ and $p = 0\%$, we can calculate the number of voice calls per unit of bandwidth assuming that the TDMA sub-

frame take up (40% - 80%) for a reasonable percentage of the TDMA load. The results in Table 1 are conservative since allowance has been made for all voice *circuits* to be active. With p = 50% the space allotted for *datagrams* and *datagram-circuits* is 80% to 40% of the frame time. It is just that there is no assurance that this condition would exist at all times. Further noting that each video circuit is equivalent to approximately 7 voice circuits, we can provide an estimate of the number of nodes which can be adequately accommodated in a single cluster. This analysis can result in strategies for acquiring or releasing bandwidth. These results are shown in Table 1

Tables 1 illustrate the wide range of capacity available for TDMA subframe voice circuits. Clusters of widely varying count are feasible so that nodes which frequently communicate with other nodes and are within a few kilometers can be easily accommodated within a single cluster. Further, in many situations a cluster can support a number of video circuits.

Table 1 also illustrates that at higher bit rates and larger diameters, propagation delay creates some reduction in performance. Some ways to regain some of this lost capacity would be to change the *circuit* protocol slightly so that each node would complete all its established *circuits* before the next node starts, increase the size of *circuits* blocks and have them sent in alternative frames and/or other mechanisms which would cut down on the need for inter-nodal propagation delays.

Appendix B H^3M CSMA/CD analysis

With the help of the performance analysis for DRAMA [4 - 5], we are able to predict with good accuracy the performance of H^3M for the CSMA/CD subframe. The CSMA/CD subframe operates similar to Ethernet. However, two effects influenced by framing can alter the performance significantly. They are:

1. packets arriving at the end of the frame which are too large to fit into the space available must be held until the start of the next CSMA/CD frame; and
2. packets arriving during the TDMA subframe are held until the start of the CSMA/CD subframe.

For unframed CSMA/CD, Metcalfe and Boggs [9] have shown that delay is dependent upon the ratio of packet transmission time to propagation delay in the network; i.e., the relative performance measures are the same whether packet size is 2000 bits and the propagation delay is 10 μsec., or the packet size is 50,000 bits and the propagation time is 250 μsec. For framed CSMA/CD, larger packet size leads to increased normalized delays; i.e., for a given frame length, F, and packet data transmission time, P, the delay increases as P/F increases. As shown in reference 4, for ratios ranging between 2% - 5%, framing causes little significant increase for the major portion of the load

range, but for ratios of 20%, increases become significant at loads of 30% and higher. For large packets of information, it is better to schedule them either as *circuit* blocks for the TDMA subframe or to break the packet up into small packets to be transmitted as a *datagram circuit.*

Reference [4] also studies CSMA/CD packet delay due to split framing using an approximate analytical model and simulation to extend the results. Four factors influence the packet delay:

1. those arriving during the voice region;
2. those arriving during the data region but delayed due to transmission of those in 1;
3. those arriving during the remainder of the data region; and
4. those arriving within one packet time of the end of the frame which are delayed until the beginning of the subsequent data region.

Taking all four contributions [4]:

$$\text{Delay} = (V^2/2F)(\lambda^2+\lambda+1)+(P/2F)(P+2V-V \text{ if } (\lambda V/P) > 1 \qquad \text{B1)}$$
$$= (V^2/2F)(\lambda+1)+(P/F)(P/2+V)+P \quad \text{otherwise}$$

where

V = voice length region,

λ = data packet arrival rate over entire frame,

F = frame length, and

P = packet transmission time.

The statistical value of number of voice calls based upon fluctuating silent periods can be estimated as:

$$\underline{V} = N(sp+d)+c \qquad \text{and} \qquad \text{B2)}$$
$$\underline{V^2} = N^2d^2+c^2+2Ndc+s^2Np+s^2p^2N(N-1)+2Nsp(Nd+c)$$

where

N = number of voice calls,

s = voice sample packet transmission time,

d = voice slot overhead time (control bits and propagation time),

p = fraction of time call is talking, and

c = time for framing information.

Reference [4] shows that by substituting equations B2) into B1), it accurately predicts the delay, D, for small packet arrival rates, λ, for a wide range of voice calls, N, and fraction of calls talking, p. However, equation B1) does not include the effect of collision resolution and so, as the arrival rate increases, the delay predicted by equation B2) is too low. Reference [4] simulated DRAMA for a wide range TDMA subframe loads (15% - 75%) and load ranges from 0% - 70%[5] obtained curves which show that actual delay which occurs. It is easy to fit these curves very accurately using poly-

[5]At the higher TDMA sub-frame loads, the higher load ranges tend to overload the system so finite delay, D, does not exist.

nomial regression and to interpolate between curves for other loads. Since data rates, λ, can be considered relative, it is possible to use the curves of reference [4] for a wide range of conditions as long as packet size/frame size condition are similar to those used in the simulation - 5%. We use this technique to demonstrate that H³M is capable of supporting a wide range of bandwidth conditions efficiently. Figure 5 shows[5] the delay performance for CSMA/CD subframe typical system with 2 km cluster diameter, 40% circuit load and a bandwidth range factor of 10. This is important since H³M supports varying loads by changing bandwidth, where as original DRAMA does so by changing bands. By working with the delay factors to reduce the effect of the collision problem at the beginning of the CSMA/CD subframe it should be possible to support a wide range of bandwidth and cluster diameter conditions. One additional feature for the CSMA/CD subframe performance should be noted. Figure 5 shows performance for the unframed load factor, i.e., the load occurring over the 10^{-2} sec. frame length assuming no circuit load. An effective CSMA/CD load, λ_{eff}, considered. It is:

$$\lambda_{eff} = (1-t_{TDMA}/T)^{-1}\lambda \qquad \text{B3)}$$

where the variables are defined for equation A1). For the situation shown in Figure 5, the effective load is 1.667 of the unframed load. Thus, at 40% load factor, the system is supporting a 68% load on the CSMA/CD subframe.

One can conclude, as a result of the above analysis and discussion, that the CSMA/CD subframe performance of H³M should be similar to that expected nominal CSMA/CD and that it should adequately support packet data transport. As we have noted in section 2.4, when cluster bandwidth is inadequate, additional bandwidth can be allocated in order to relieve congestion.