

# Quality of Service Support in Mobile *ad-hoc* IP Networks\*

Derya H. Cansever  
GTE Laboratories  
40 Sylvan Road  
Waltham, MA 02454

Arnold M. Michelson  
GTE Government Systems  
400 John Quincy Adams Road  
Taunton, MA 02780

Allen H. Levesque  
Worcester Polytechnic Institute  
100 Institute Road  
Worcester, MA 01609

## ABSTRACT

Real-time IP applications such as IP telephony, IP video streaming, and transport of time-sensitive information need predictable network resources. Examples of network resources include minimum guaranteed bandwidth, and special processing of certain types of packets at points of congestion, regardless of traffic fluctuations in the network. Support for such applications in mobile ad-hoc networks requires acceptable channel conditions, QoS aware mechanisms for channel access, identification of transit nodes that can sustain the resulting traffic, as well as measures for congestion prevention and management at the nodes. This paper provides a framework for QoS support in Mobile ad-hoc IP networks, and discusses methods to enable QoS aware media access control, and routing IP packets with QoS constraints.

## I. INTRODUCTION

As real-time applications find their way into the Internet, efforts to support QoS in IP networks have intensified. Approaches range from allocating resources to individual flows, to throwing "enough bandwidth" to the problem. Mobile networks in general, mobile ad-hoc networks (MANET) that transport military applications in particular, have very distinct characteristics that require specialized solutions for QoS support. In this paper, we review components of an overall architecture to support QoS, and propose a solution specialized for MANETS.

In this paper, we assume that applications have the ability to request appropriate network resources, and to indicate it by generating specific packets, and/or marking the packets that carry the application content. We also assume that the links are bi-directional, and that the bandwidth available in the direction from one node to another is roughly similar to the bandwidth

available in the reverse direction. Discussion in this paper is limited to unicast transmission of packets, and support for multicast will be addressed in a future paper. MANETS are expected to carry tactical Internet traffic on poor quality radio circuits, *i.e.*, channels with high bit-error rates (BER). In addition to Gaussian background noise, channel errors may also be caused by factors such as impulsive noise, signal fading, unintentional interference from other users of the band, and intentional enemy jamming. Clearly, for tactical communications, performance over a wide range of channel conditions is essential.

One of the characteristics of MANETS is that they are bandwidth limited, and the available bandwidth is subject channel impairments. The latter induces the use of redundant codes, which further reduces the available bandwidth. When channel impairments reach beyond a certain threshold, coding methods alone do not suffice to ensure transmission. Such a threshold is identified as a BER of  $10^{-2}$  in [CANS]. When the BER reaches beyond such a threshold, it would make sense to declare that particular link unreachable, and seek an alternate path.

Another characteristic of MANETS is that there is contention from multiple users in inserting each packet into the shared transmission link. Thus, a solution that involves a large number of control packets is not likely to be desirable for MANETS, in that it would tend to increase the collision rate, and also decrease the probability of obtaining a channel. This, in turn, would deteriorate the overall efficiency of the system.

Nodes in a MANET send, receive and relay packets. We argue that a node is not likely to serve as transit to a very large number of flows, as it would be the case in a backbone node of a high-speed terrestrial network. There are two reasons for this assertion. One is that in the majority of proposed schemes [TORA], [DSR], [AODV], there is not an identifiable set of "backbone nodes" that provide transit to a majority of the packets

---

\* Prepared through collaborative participation in the Advanced Telecommunications & Information Distribution Research Program (ATIRP) Consortium sponsored by the U.S. Army Research Laboratory under the Federated Laboratory Program, Cooperative Agreement DAAL01-96-2-0002. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.

in the network. We expect that in many cases, especially in high mobility environments, nodes will randomly assume transit responsibilities, in such a way that, no node is significantly more likely than others in serving as a relay node. Another reason that the relatively low speed channels in MANETS will be quickly saturated with real time traffic. Such an analysis is provided in [GUPTA], which identifies an upper bound to the number of active nodes in a MANET.

## II. QOS SUPPORT IN MANETS

QoS support roughly means providing the applications with enough network resources, so that they can function within acceptable performance limits. There may be two reasons for not having sufficient network resources. One is due to physical layer impairments. This may be addressed using stronger coding, or increasing the power level, or alternate routing. The other reason for QoS impairment is congestion. Solutions to physical impairment problems generally increase the congestion in the network. This will occur either by reducing the bandwidth due to stronger codes, or by increasing the traffic at other nodes due to alternate routing resulting from impaired channels, or by increasing the probability of packet collisions due to the augmented power. The focus of this paper is to address congestion issues in providing QoS support.

There are several approaches to the problem of congestion management. One approach is to over-engineer the network so that resource contention is avoided. In general, this method is not applicable in MANETS. Another method is to identify packets carrying real-time applications and provide them with special treatment, such as priority in the transit nodes. Of course, an example of such a method is Differentiated Services (DS) [RFC 2474], [RFC 2475]. Roughly speaking, Differentiated Services consists of marking the packets with particular values of the DS field [RFC 2474], and then configuring the nodes appropriately so that the marked packets are treated as intended at the nodes they transit. Differentiated Services does not identify individual flows, but provides a prescribed treatment to an aggregate of flows, as specified in the Per Hop Behavior (PHB) that corresponds to a particular value of the DS field. Proposals for PHB include [AFP] and [EFP].

PHBs are implemented using Packet Scheduling Algorithms such as Weighted Fair Queuing [WFQ], or Start Time Queuing [STQ]. These algorithms ensure that at the minimum, a given portion of the bandwidth in the outgoing link of a queue will be allocated to a certain class of traffic. Since DS does not differentiate the individual flows, the guaranteed bandwidth is allocated to an aggregate of flows, differentiated from

other aggregated flows by their respective DS field value. Each class may correspond to a minimum bandwidth, or a percentage of the total bandwidth in the node, for example. Using this approach, one guarantees that a class of traffic obtains a given portion of the network resources. In the extreme case that the portion allocated to a specific class is 100%, one faces a strict priority scheme. Of course, there may be too many packets from a given class for the allocated resources. Then, packets from this particular class would be subject to congestion within their allocated resources, as if they were in a single class best-effort system. Here, the allocated resources being determined by the packet scheduling algorithm as a function of the overall traffic, and that class' predetermined minimum amount of network resources.

A way to overcome congestion within a class, and to obtain finer granularity in QoS definition, is to allocate end-to-end network resources for a particular flow during the life of the application. This is a challenging problem even in terrestrial networks with fixed nodes. MANETS make it more difficult due to mobility, lack of fixed infrastructure, shared medium and high bit error rates.

## III. END-TO-END RESOURCE ALLOCATION

A QoS support methodology for MANETS that would allocate resources to individual flows needs the following ingredients:

- A QoS metric
- A MAC protocol that supports QoS
- A method to identify flows
- A method to indicate QoS requirements
- A method to identify nodes with sufficient resources (QoS routing)
- A method to reserve resources
- A method to release resources.

We discuss the above ingredients and proposed methods in the following subsections.

### 3.1 A QoS metric

There are several conceivable QoS metrics that are related to the performance requirements of real-time applications, such as delay, jitter and throughput. The chosen metric(s) will be used in the routing algorithm to find a feasible path that satisfies the associated constraints. It is well known that [RFC 2386], finding a route that satisfies multiple constraints is a computationally hard problem. Implementing a multi-dimensional QoS metric for MANETS, where control packet exchange is to be minimized, and the topology is subject to constant changes, is not advisable. Instead, we

propose to use the maximum available bandwidth in a node as a simple, but generally sufficient metric. With this metric, nodes can account for their used and unused resources, and communicate it easily with their peers. Also, note that such a metric would approximate a circuit switched network, and the latter has been supporting real time applications for decades.

The amount of bandwidth available through at Node  $i$  depends not only on the traffic generated at that node and the transit traffic through the node, but also on the traffic generated by the nodes at the neighborhood of the Node  $i$ , as they share a common media. Neighborhood of Node  $i$  is defined as the set of nodes whose transmission interferes with the transmission of Node  $i$  when they both transmit simultaneously. Here, we differentiate between the maximum *unused* bandwidth in a node and the maximum *available* bandwidth in a node in the following manner: Let  $C_i$  denote the maximum bandwidth, or the capacity of the node  $i$  in bits/second. Let also  $l_{ij}$  denote the traffic from node  $i$  to node  $j$  in bits/second. Here,  $l_{ij}$  includes traffic generated at the node  $i$ , as well as transit traffic through that node. Then, the maximum unused bandwidth in node  $i$  is defined as

$$MUB_i = C_i - \sum_j l_{ij}, \forall j \in \text{Neighborhood of } i \quad (1)$$

When we take into account traffic from neighboring nodes of the Node  $i$ , maximum available bandwidth at that node becomes

$$MAB_i = MUB_i - \sum_{j \in N_i} \sum_{k \in N_j} l_{jk} \quad (2)$$

Thus, assuming that the nodes are aware of each other's presence and their respective capacities, they can infer each other's Maximum Available Bandwidths by keeping track of their Maximum Unused Bandwidth.

### 3.2 A MAC Protocol that supports QoS

Since the transmission occurs over a shared media, it is necessary that terminals be able to access it such that:

- They fulfill their QoS requirements, and
- They do not prevent their neighbors from fulfilling their own QoS requirements.

This is a difficult problem, especially when nodes in a neighborhood have access to dissimilar information. That is, suppose Node  $i$  is a neighbor of Nodes  $j$  and  $k$ , respectively. Furthermore, assume that Nodes  $j$  and  $k$  cannot communicate directly. Then, Nodes  $j$  and  $k$  affect each other indirectly through the actions of Node  $i$ , even though Nodes  $k$  and  $j$  do not know about each

others QoS requirements. In the rest of this section, we assume that all the nodes have access to similar information. That is, if Node  $i$  exchanges QoS information with Nodes  $j$  and  $k$ , then Nodes  $j$  and  $k$  exchange QoS information with each other. We will address the problem of dissimilar QoS information among members of a neighborhood in a future paper.

The proposed MAC protocol will support regulated access, as well as random access to the media. We assume that the random access will occur at given time intervals, and that a minimum amount of bandwidth is allocated for this function, such as 10% of the time. The random access algorithm will follow a standard mechanism, such as CDMA-CD, as defined in the IEEE 802.11 Protocol. Using the random access mechanism, all the nodes broadcast their MUBs, thus their total bandwidth requests. Now that all the nodes are aware of their neighbors' traffic demands, it is possible to implement a distributed algorithm that allocates the "air time" among the nodes in its neighborhood. To this goal, we define a Cycle, which consists of a maximum number of time slots. At the beginning of a Cycle, each node is aware of all the neighbors' traffic demands, and runs a simple algorithm that allocates the time slots among the neighbors, in proportion to their demand. It is assumed that all the nodes are running the same algorithm with identical information, thus they are synchronized in the sense that they all have the same understanding on who will be sending at which slot. An example of such an algorithm is the one that allocates the slots in proportion to the declared demands among the nodes. The order at which the nodes utilize their time slots is determined by their IP addresses, which were broadcast earlier along with the traffic demand. For fairness, the order may rotate at the end of each cycle. At the end of each allocated slot, there will be a period of time to allow random access. This period will be used to transmit best effort traffic, as well as changes in the traffic demands. Also, new nodes in the neighborhood will broadcast their traffic demands during that period. Nodes will broadcast new traffic demand information only when there is a change in their traffic profile, or when they realize that the established path cannot support their traffic any longer. When a transit node realizes that it cannot support a flow with QoS requirements, it will inform the originating node using the path information that it already cached. If the originating node cannot be informed, higher layer protocols will be responsible for that task. The updated traffic information will be used for the following Cycle. A Cycle has a maximum number of slots, and that number may decrease in proportion to the total traffic in the neighborhood. Now, each node runs a Packet Scheduling Algorithm such as the one discussed in [CANS], or [STQ]. Using such a Packet Scheduling Algorithm, nodes allocate the slots among the flows that they support. These flows can be generated,

relayed, or terminated by the node. Once they know their share of the slots and the total traffic, nodes can implement a Connection Admission Control (CAC) algorithm for accepting, or rejecting new traffic demands that they receive, or relay to other destinations. This is further discussed in Subsection 3.5.

### *3.3 A method to identify flows*

A possible way to identify a flow is to use the triplet (Source IP Address, Destination IP Address, Differentiated Services Field). Here, a specific value of the Differentiated Services Field would indicate that this packet is to receive a pre-specified treatment.

### *3.4 A method to indicate QoS requirements*

Packets that need to receive special treatment will have a specific value in their DS field. The exact nature of the special treatment will be indicated in the resource reservation process, which will be described in subsection 3.6

### *3.5 A method to identify nodes with sufficient resources (QoS routing)*

QoS routing in MANETS can be considerably simplified by using the Minimum Available Bandwidth metric and the on-demand route creation feature of some of the existing proposals [TORA], [DSR], [AODV]. In the Query [TORA], or in the Route Request [DSR, AODV] packet is normally used to identify a path to the destination node. Here, we also let the sender indicate the requested minimum bandwidth for the corresponding flow. A node that receives this Query, or Route Request packet, compares the requested minimum bandwidth to its Maximum Available Bandwidth (MAB). If the requested bandwidth is less or equal than its MAB, it delivers the Route Request packet to the next hop nodes, as specified in the routing algorithm. Otherwise, Since the IP addresses of all the transit nodes are added to the Route Request packet, the receiving end node has one, or multiple sequence of routers that form an end-to-end path and that can also fulfill the QoS requirement as indicated in the Route Request packet. Among this list of router paths, the node at the receiving end will identify the shortest one to be used, and send it back to the node that issued the route request. If the issuer of the Route Request packet does not receive a Route Reply packet within a given period of time, it will re-issue the Route Request packet, possibly with a different QoS requirement. Of course, Route Requests that correspond to Best Effort traffic will not be rejected due to insufficient available bandwidth.

### *3.6 A method to reserve resources*

When the Route Reply packet traverses the sequence of nodes derived from the Route Request packet, it will reserve the bandwidth indicated in the original request packet in both directions, unless indicated otherwise. If the requested bandwidth will be different in each direction, it will be indicated in the request packet. In this case, nodes in the transit path of the Route Request packet will consider the larger of the requested bandwidth, and the reservation will be made according to the directional requests. The tasks of maintaining the routes, and adjusting to topology changes are to be handled by the underlying MANET routing algorithm, such as [DSR]. When the path becomes unusable due to mobility, or some other reason, a new path will be searched and established per DSR specifications. For our purposes, a link between two nodes will be considered functional only if both nodes have enough resources to fulfill the requested minimum bandwidth. Otherwise, a link whose available bandwidth is smaller than the requested QoS will be ignored by the routing algorithm.

### *3.7 A method to release resources*

Since minimizing the control packet flow is a goal, a possible way to accomplish this is to make use of local timers. That is, a node will release resources reserved for a flow if it does not receive any packet from that flow within a period of T seconds, where T is an adjustable parameter.

## VI CONCLUSIONS

In this paper, we presented a methodology for providing end-to-end QoS support in MANETS. In particular, we proposed a distributed QoS aware MAC protocol, and a QoS routing scheme specialized for mobile ad-hoc networks. The proposed method requires a minimal number of control packets, and avoids complexities of constraint-based routing normally encountered in packet networks\*.

## REFERENCES

- [CANS] D. Cansever, A. Michelson and A. Levesque "Error Control and Resource Management in Mobile Ad-hoc Networks", Proceedings of the PIMRC 1999.

---

\* The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied of the Army Research Laboratory or the U.S. Government.

- [TORA] V. Pak and S. Corson, "Temporally Ordered Routing Algorithm (TORA) Version 1 Functional Specification". IETF Draft, work in progress.
- [DSR] J. Broch, D. Johnson, and D. Maltz, "The Dynamic Source Routing Protocol for Mobile Ad-hoc Networks", IETF Draft, work in progress.
- [AODV] C. Perkins and E. Royer, "Ad Hoc On Demand Distance Vector Routing", IETF Draft, work in progress.
- [CEDAR] R. Sivakumar P. Sinha and V. Bharghavan, "Core Extraction Distributed Ad hoc Routing (CEDAR) Specification", IETF Draft, work in progress.
- [GUPTA] P. Gupta and P.R. Kumar, "The Capacity of Wireless Networks", submitted for publication, available in <http://black.csl.uiuc.edu/~prkumar>.
- [AFP] J. Heinanen et al, "Assured Forwarding PHB Group", IETF Draft, work in progress.
- [EFP] V. Jacobson et al, "An Expedited Forwarding PHB", IETF Draft, work in progress.
- [WFQ] A. Parekh and R. Gallager, "A generalized Processor Sharing Approach to Flow Control - the Single Node Case", ACM/IEEE Transactions on Networking, 1(3) 344-357, June 1993.
- [STQ] P. Goyal et al., "Start-time Fair Queuing: A Scheduling Algorithm for Integrated Services", in the Proceedings of the ACM-SIGCOMM 96, p.p. 157-168, Paulo Alto CA August 1996.