QoS in Mobile Ad Hoc Networks*

Prasant Mohapatra†, Jian Li, and Chao Gui
Department of Computer Science
University of California
Davis, CA 95616
{prasant, lijian, guic}@cs.ucdavis.edu

December 13, 2002

Abstract

The widespread use of mobile and handheld devices is likely to popularize ad hoc networks, which do not require any wired infrastructure for intercommunication. The nodes of mobile ad hoc networks (MANETs) operate as end hosts as well as routers. They intercommunicate through single-hop and multi-hop paths in a peer-to-peer fashion. With the expanding range of applications of MANETs, the need for supporting Quality of Service (QoS) in these networks is becoming essential. This paper provides a survey of issues in supporting QoS in MANETs. We have considered a layered view of QoS provisioning in MANETs. In addition to the basic issues in QoS, the report describes the efforts on QoS support at each of the layers, starting from physical and going up to the application layer. A few proposals on the inter-layer approaches for QoS provisioning have also been addressed. The paper concludes with a discussion on the future directions and challenges in the areas of QoS support in MANETs.

Keywords: Inter-layer QoS, Layered QoS, Mobile ad hoc networks, Quality of Service, QoS Routing.

*This work was supported in part by the National Science Foundation under the grants CCR-0296070 and ANI-0296034.

†Corresponding author.

1 Introduction

The wireless mobile networks and devices are becoming increasingly popular as they provide users access to information and communication anytime and anywhere. The conventional wireless mobile communication is usually supported by a wired fixed infrastructure (like ATM or Internet). The mobile devices use single-hop wireless radio communication to access a base station that connects it to the wired infrastructure. In contrast, the class of mobile ad hoc networks (MANETs) does not use any fixed infrastructure. The nodes of MANETs intercommunicate through single-hop and multi-hop paths in a peer-to-peer fashion. Intermediate nodes between two pairs of communicating nodes act as routers. Thus the nodes operate both as hosts as well as routers. The nodes are mobile, and so the creation of routing paths is affected by the addition and deletion of nodes. The topology of the network may change rapidly and unexpectedly. Figure 1 shows an example of a mobile ad hoc network.

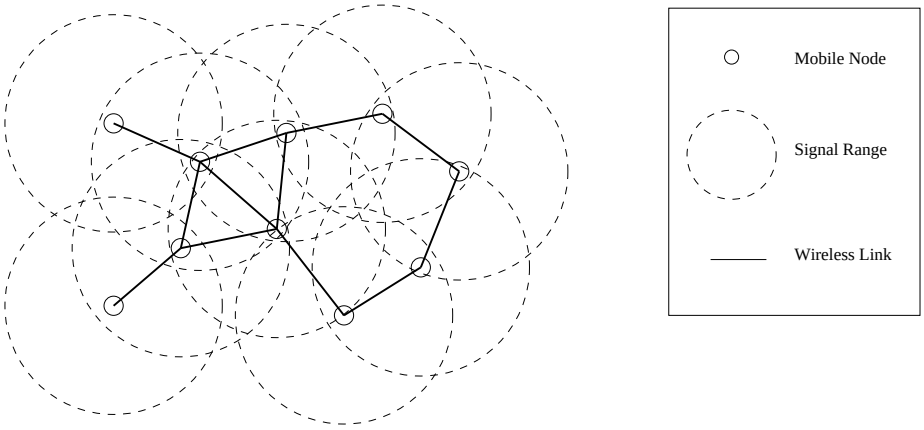

Figure 1: A mobile ad hoc network.

MANETs are useful in many application environments and do not need any infrastructure support. Collaborative computing and communications in smaller areas (building organizations, conferences, etc.) can be set up using MANETs. Communications in battlefields and disaster recovery areas are other examples of application environments. Similarly communications using a network of sensors or using floats over water are other potential applications of MANETs. The increasing use of collaborative applications and wireless devices may further add to the needs and usage of ad hoc networks.

With the increase in quality of service (QoS) needs in evolving applications, it is also desirable to support these services in the MANETs. The resource limitations and variability further add to the need for QoS provisioning in such networks. However, the characteristics of these networks make the QoS support a very complex process. QoS support in MANETs encompasses issues at the application layer, transport layer, network layer, media access layer, and the physical layer of the network infrastructure. This paper provides a detailed survey of the issues involved in supporting QoS across all the protocol layers in MANETs. We have classified different approaches, discussed

various techniques, and outlined the future issues and challenges related to the QoS provisioning in MANETs.

The rest of the paper is organized as follows. In Section 2, we define the QoS metrics and review the basics of QoS support in MANETs. The QoS issues at all the layers of the Internet protocol are discussed in Section 3. The inter-layer design approaches are described in Section 4, followed by the outline of future challenges in Section 5.

2 Issues in QoS-Aware MANETs

In this section, we first define QoS and its metrics, followed by an outline of the generic issues and difficulties in QoS-aware MANETs.

2.1 Quality of Service Metrics

QoS is usually defined as a set of service requirements that needs to be met by the network while transporting a packet stream from a source to its destination. The network needs are governed by the service requirements specified by the end user applications. The network is expected to guarantee a set of measurable prespecified service attributes to the users in terms of end-to-end performance, such as delay, bandwidth, probability of packet loss, delay variance (jitter), etc. Power consumption and service coverage area are two other QoS attributes that are more specific to MANETs. The QoS metrics could be defined in terms of one of the parameters or a set of parameters in varied proportions.

The QoS metrics could be concave or additive. Bandwidth is concave in the sense that end-to-end bandwidth is the minimum of all the links along the path. Delay and delay jitter are additive. The end-to-end delay (jitter) is the accumulation of all delays (jitters) of the links along the path. It has been proven that if QoS contains at least two additive metrics then the QoS routing is an NP-complete problem [20]. Thus heuristic algorithms are usually developed for multi-constraints QoS routing.

2.2 QoS Support in MANETs: Issues and Difficulties

Mobile multi-hop wireless networks differ from the traditional wired Internet infrastructures. The differences introduce unique issues and difficulties for supporting QoS in the MANET environments, which are itemized as follows.

- **Unpredictable Link Properties:** Wireless media is very unpredictable. Packet collision is intrinsic to wireless network. Signal propagation faces difficulties such as signal fading, interference, and multi-path cancellation. All these properties make the measures, such as bandwidth and delay of a wireless link, unpredictable.

- **Hidden Terminal Problem:** Multi-hop packet relaying introduces the hidden terminal problem. This problem happens when signals of two nodes, which are out of the transmission range of each other, collide at a common receiver.

- **Node Mobility:** Mobility of the nodes create a dynamic network topology. Links will be dynamically formed when two nodes come into the transmission range of each other and be torn down when they move out of range.

- **Route Maintenance:** The dynamic nature of the network topology and the changing behavior of the communication medium make the precise maintenance of network state information very difficult. Thus the routing algorithms in ad hoc networks have to operate with inherently imprecise information. Furthermore, in ad hoc networking environments, nodes can join or leave anytime. The established routing paths may be broken even during the process of data transfer. Thus arises the need for maintenance and reconstruction of routing paths with minimal overhead and delay.

  QoS-aware routing would require reservation of resources at the routers (intermediate nodes). However, with the changes in topology the intermediate nodes also change and new paths are created. Thus the reservation maintenance with the updates in the routing path becomes cumbersome.

- **Limited Battery Life:** Mobile devices generally are dependent on finite battery sources. The resource allocation for QoS provisioning must consider the residual battery power and the rate of battery consumption corresponding to the resource utilization. Thus all the techniques for QoS provisioning should be power-aware and power-efficient.

- **Security:** Security can be considered as a QoS attribute. Without adequate security, unauthorized accesses and usages may violate the QoS negotiations. The nature of broadcasts in wireless networks potentially results in more security exposures. The physical medium of communication is inherently insecure. So we need to design security-aware routing algorithms for ad hoc networks.

2.3 Compromising Principles

The dynamic nature of MANETs is attributed to multiple sources, variable link characteristics, node movements, changing network topology, and variable application demands. Providing QoS in such a dynamic environment is very difficult. Two compromising principles for QoS provisioning in the MANETs are: soft QoS and QoS adaptations.

Because of the special properties of mobile wireless networks, some researchers have proposed the notion of soft QoS[19]. Soft QoS means that after the connection setup, there may exist transient periods of time when the QoS specification is not honored. However, we can quantify the level of QoS satisfaction by the fraction of total disruption time over the total connection time. This ratio should not be higher than a threshold.

In a fixed-level QoS approach, a reservation is represented by a point in an n-dimensional space with coordinates defining the characteristics of the service. In a dynamic QoS approach [18], we can allow a reservation to specify a range of values, rather than a single point. With

such an approach, as available resources change, the network can readjust allocations within the reservation range. Similarly, it is desirable for the applications to be able to adapt to this kind of re-allocations. A good example of this case is the layered real-time video, which requires a minimum bandwidth assurance and allows for enhanced level QoS when additional resources are available. The QoS adaptation can be also done at various layers. The physical layer should take care of changes in transmission quality, for example, by adaptively increasing or decreasing the transmission power. Similarly, the link layer should react to the changes in link error rate, including the use of automatic repeat-request (ARQ) technique. A more sophisticated technique involves adaptive error correction mechanism which will increase or decrease the amount of error correction coding in response to changes in transmission quality. As the link layer takes care of the variable bit error rate, the main effect observed by the network layer will be a change in effective throughput(bandwidth) and delay.

3 QoS from a Layered Perspective

In this section, we examine the QoS provisioning issues in MANETS with a layered prospective, starting from the physical layer and going up to the application layer.

3.1 QoS Support in Physical Channels

Wireless channel in a MANET is time-varying, which means that the channel model fluctuates in time. Thus adaptive modulation which can tune many possible parameters according to current channel state (e.g. instantaneous singal-to-noise ratio) is necessary to derive better performance from wireless channels. So one of the major challenges in supporting QoS communication over wireless media is channel estimation. It involves accurate channel estimation at the receiver and then the reliable feedback of the estimation to the transmitter so that the transmitter and receiver can be properly synchronized. The time-varying fading channel also makes these coding schemes designed for a fixed channel model unsuitable for use in MANETs. Wireless channel coding needs to address the problems introduced by channel fading, multipath fading and mobility.

Communications over wireless channels are subject to noise and collision. Increasing demand for image and real-time audio/video transmission in wireless networks just makes this problem more complicated. It has been realized that supporting QoS in wireless communications should rely not only on improvement in channel techniques but also its tight integration with upper layers, like source compression algorithms at the application layer. Using higher source coding rate(less data compression) can decrease the final end-to-end distortion. Like wise, using more channel protection (longer code words) can reduce possible channel errors, which implies less end-to-end distortion. Since the wireless channel capacity is limited, we have to consider a tradeoff between these two rates. Joint source-channel coding takes into consideration both source characteristics and current channel situation [15].

3.2 QoS Provisioning at the MAC Layer

Recently, many MAC schemes have been proposed for wireless networks, aimed at providing QoS guarantee for real-time traffic support. However, these MAC protocols in general rely on centralized control, which is only viable for single-hop wireless networks. In multihop wireless networks, a fully distributed scheme is needed that should first solve the hidden terminal problem. MACA [8] (Multihop Access Collision Avoidance) is proposed to solve this by RTS/CTS dialogs. It does not solve all of the hidden terminal problem. MACAW[2] was proposed as an extension to MACA to provide faster recovery from the hidden terminal collisions. A few synchronous methods are proposed for multihop wireless networks: cluster TDMA[5], the virtual network and SWAN[1]. These protocols support real-time traffic since slots can be reserved and QoS routing is used to find the route with sufficient bandwidth. The downside is that the strict time framing and global synchronization introduce much implementation complexity and cost.

IEEE 802.11 includes the collision avoidance feature of MACA and MACAW by its distributed control function (DCF). Its fundamental access method is CSMA/CA, which solves hidden terminal problem completely. However, it does not provide real-time traffic support. Veres et al [19] have analyzed the delay incurred by IEEE 802.11 DCF based on which a modified DCF is proposed to support relative service differentiation. In this section, we have surveyed the MAC layer QoS issues proposed for MANETs.

IEEE 802.11 Distributed Control Function (DCF) And Its Extension

IEEE 802.11 is a carrier sense multiple access with collision avoidance (CSMA/CA) protocol. In the DCF mode, after the node has sensed the medium to be idle for a time period longer than distributed inter-frame space (DIFS), it begins transmitting. Otherwise, the node defers the transmission and starts to backoff. Each node holds a value called contention window (CW), the low and high ends of which are represented as CW_{min} and CW_{max}, respectively. The duration of the backoff is decided by a backoff timer which is set to be a random value between 0 and CW. Whenever the medium becomes idle for periods longer than DIFS, the backoff timer is decremented periodically. As soon as the timer expires, the node starts transmission. To improve performance by reducing packet collisions, the sender will first send a short packet called request-to-send (RTS) if the data packet is longer than a threshold value. If the intended receiver grants the request, it will return another short packet called clear-to-send (CTS). Upon receiving the CTS, the sender will start sending the data packet, while other nodes will try to avoid collision with the upcoming data packet.

As we can see, IEEE 802.11 DCF is a good example of best-effort type control algorithm. It has no notion of service differentiation and no support for real-time traffic. Veres et al [19] have proposed a scheme to extend IEEE 802.11 DCF with ability to support at least two service classes, high-priority (i.e. premium service) and best-effort. Traffic of premium service class is given lower values for congestion window $\{CW_{min}, CW_{max}\}$ than those of best-effort traffic. If packets of both types collide, the packet with smaller CW_{min} value is more likely to occupy the medium earlier.

Black Burst Contention Scheme

The Black burst (BB) contention scheme proposed in [17] avoids packet collision in a very novel manner, and solves the packet starvation problem as well. Packets from two or more flows of the same service class are scheduled in a distributed manner with fairness guarantee. Nodes contend for the medium after it has been idle for a period longer than the inter-frame space. Nodes with best-effort traffic and nodes with real-time traffic use different inter-frame space values. This makes real-time traffic as a group to have higher priority over data nodes. BB contention scheme is added to any CSMA/CA type protocol in the following manner. Right before sending their packets when the medium remains idle long enough, real-time nodes first contend for transmission right by jamming the media with pulses of energy, which are called BB's. The novelty of this scheme is that each contending node is using a BB with different length. The length of each BB is an integral number of black slots, each slot is of length t_{bslot}. The number of slots that forms a BB is an increasing function of the contention delay experienced by the node, measured from the instant when an attempt to access the channel has been scheduled until the node starts the transmission of its BB. Following each BB transmission, a node senses the channel for an observation interval. Since distinct nodes contend with BB's of different length, each node can determine without ambiguity whether its BB is of greatest length. Thus only one winner is produced after this contention, who will transmit its real-time packets successfully. BB contention ensures that real-time packets are transmitted without collisions and with priority over best-effort packets.

MACA/PR

MACA/PR (Multihop Access Collision Avoidance with Piggyback Reservation) [13] provides guaranteed bandwidth support (via reservation) for real-time traffic. It establishes real time connections over a single hop only. However, it should work with QoS routing algorithm and a fast reservation setup mechanism. The first data packet in the real-time stream makes reservations along the path. A RTS/CTS dialog is used on each link for this first packet in order to make sure that it is transmitted successfully. Both RTS and CTS specify how long the data packet will be. Any station near the sender which hears the RTS will defer long enough so that the sender can receive the returning CTS. Any node near the receiver which hears the CTS will avoid colliding with the following data packet. The RTS/CTS dialog is used only for the first packet to setup reservations. The subsequent packets do not require this dialog.

When a sender sends a data packet, the sender schedules the next transmission time after the current data transmission and piggybacks the reservation in the current data packet. Upon receiving the data packet correctly, the intended receiver enters the reservation into its reservation table and returns an ACK. The neighbors which hear the data packet can learn about the next packet transmission time. Likewise, neighbors at receiver side which hears the ACK will avoid to send at the time when the receiver is scheduled to receive next packet. Notice that the ACK serves as renewing of reservation rather than for recovering from packet loss. In fact, if the ACK is not received, the packet is not retransmitted. Instead, if the sender consecutively fails to receive

ACK N times, it assumes that the link is not satisfying the bandwidth requirement, and notifies the upper layer, i.e., the QoS routing protocol. So this "reservation" ACK serves as "protector" for the given time window, and a mechanism to inform the sender if something is wrong on the link.

3.3 QoS-Aware Routing at the Network Layer

Several routing protocols have been proposed for the MANETs, which can be classified into three broad categories: (a) precomputed table-based routing schemes, (b) on-demand source-based routing schemes, and (c) constraint-based routing schemes.

The precomputed table-based routing schemes require each of the nodes in the network to maintain tables to store the routing information, which is used to determine the next hop for the packet transmission to reach the destination. The protocol attempts to maintain the table information consistent by transmitting periodical updates throughout the network. These routing scheme may be flat or hierarchical in nature. Examples of flat table-based routing schemes include destination-sequenced distance vector (DSDV) routing and wireless routing protocol (WRP) [16]. The flat routing schemes require the maintenance of the state of the entire network at all the nodes, which limits its scalability. In the hierarchical approach, the state of only a subset of the network is maintained at all the nodes, and the routing is facilitated through another level of state information, which is stored in a fewer number of nodes. An example of the hierarchical table-based routing scheme is the Clusterhead Gateway Switch Routing (CGSR) [16].

In the case of on-demand source-based routing schemes, the routes are created as and when necessary based on a query-reply approach. When a node needs to communicate with another node, it initiates a route discovery process. Once a route is found, it is maintained by a route maintenance procedure until the route is no longer needed. Examples of on-demand source-based routing schemes include Ad hoc On-demand Distance Vector (AODV) routing protocol, Dynamic Source Routing (DSR), and Temporary Ordered Routing Algorithm (TORA) [16]. These algorithms focus on finding the shortest path between the source and destination nodes by considering the node status and network configuration at the time when a route is desired. Zone Routing Protocol (ZRP) is a hybrid of on-demand routing and table-based routing [17].

The constraint-based routing protocols use metrics other than shortest-path for finding a suitable and feasible route. Associativity-Based Routing (ABR) and Signal Stability Routing(SSR) [16] take into account the node's signal strength and location stability so that the path chosen is more likely to be long-lived. Dynamic Load-Aware Routing (DLAR) [10] considers the load of intermediate nodes as the primary route selection metric, whereas the distributed dynamic load-balancing algorithm [7], constructs a load-balanced backbone tree, which simplifies routing and avoids per-destination state maintenance for routing and per-flow state maintenance for resource reservations.

The routing schemes discussed earlier in this section were proposed for routing messages on the shortest available path or within some system-level constraints. Routing messages in such

paths may not be adequate for applications that require QoS support. In this section, we review the routing schemes that can support QoS in MANETs.

Figure 2 shows the wireless topology derived from Figure 1. The mobile nodes are labeled as A,B,C,...,K. The numbers beside each edge represent the available bandwidths of the wireless links. Suppose we want to find a route from a source node A to a destination node G. For conventional routing using shortest path (in terms of the number of hops) as metric, the route A-B-H-G will be chosen. It is quite different in QoS route selection. Suppose we consider bandwidth as the QoS metric and desire to find a route from A to G with a minimum bandwidth of 4. Now the feasible route will be A-B-C-D-E-F-G. The shortest path route A-B-H-G will not be adequate for providing the required bandwidth.

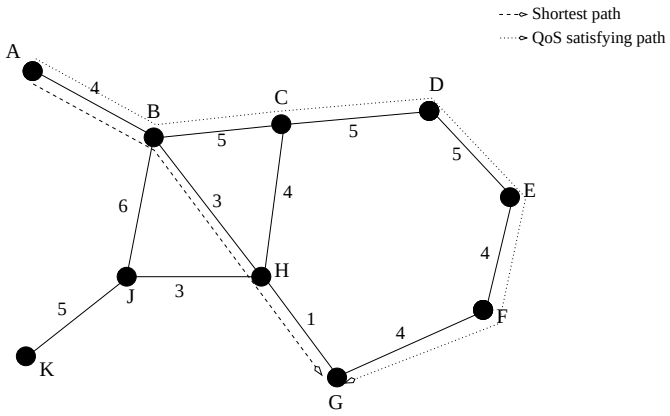

Figure 2: An example of QoS routing in ad hoc networks.

The primary goal of the QoS-aware routing protocols is to determine a path from a source to the destination that satisfies the needs of the desired QoS. The QoS-aware path is determined within the constraints of minimal search, distance, and traffic conditions. Since the path selection is based on the desired QoS, the routing protocol can be termed as QoS-aware. Only a few QoS-aware routing protocols have been proposed yet for MANETs, most of which are outlined in this section.

CEDAR

CEDAR, a Core Extraction Distributed Ad hoc Routing algorithm is proposed as a QoS routing scheme for small to medium size ad-hoc networks consisting of tens to hundreds of nodes [17]. It dynamically establishes the core of the network, and then incrementally propagates the link states of stable high-bandwidth links to the core nodes. The route computation is on-demand basis, and is performed by the core nodes using only local state. CEDAR has three key components: (a) *Core Extraction:* A set of nodes is elected to form the core that maintains the local topology of the nodes in its domain, and also perform route computations. The core nodes are elected by approximating a minimum dominating set[1] of the ad hoc network. (b) *Link*

[1]A dominating set is a subset of the network in which every node not in the set is adjacent to at least one node in the set. A minimum dominating set is one such set with minimum cardinality.

State Propagation: QoS routing in CEDAR is achieved by propagating the bandwidth availability information of stable links to all core nodes. The basic idea is that the information about stable high-bandwidth links can be made known to nodes far away in the network, while information about the dynamic or low bandwidth links remains within the local area. (c) *Route Computation:* Route computation first establishes a core path from the domain of the source to the domain of the destination. Using the directional information provided by the core path, CEDAR iteratively tries to find a partial route from the source to the domain of the furthest possible node in the core path satisfying the requested bandwidth. This node then becomes the source of the next iteration.

In the CEDAR approach, the core provides an efficient and low-overhead infrastructure to perform routing, while the state propagation mechanism ensures the availability of link-state information at the core nodes without incurring high overheads.

Integrating QoS in Flooding-Based Route Discovery

A ticket-based probing algorithm with imprecise state model was proposed by Chen and Nahrstedt [17]. While discovering a QoS-aware routing path, this algorithm tries to limit the amount of flooding (routing) messages by issuing a certain amount of logical tickets. Each probing message must contain at least one ticket. When a probing message arrives at a node, it may be split into multiple probes and forwarded to different next-hops. Each child probe will contain a subset of tickets from their parent. Obviously, a probe with a single ticket cannot be split any more. When one or more probe(s) arrive(s) at the destination, the hop-by-hop path is known and delay/bandwidth information can be used to perform resource reservation for the QoS-satisfying path.

In wired networks, a probability distribution can be calculated for a path, based on the delay and bandwidth information. In an ad hoc network, however, building such a probability distribution is not suitable, because wireless links are subject to breakage and state information is imprecise in nature. Hence a simple imprecise model was proposed for the ticket-based probing algorithm. It uses history and current(estimated) delay variations and a smoothing formula to calculate the current delay, which is represented as a range of $[delay - \delta,\ delay + \delta]$. To adapt to the dynamic topology of ad hoc networks, this algorithm allows different level of route redundancy. It also uses re-routing and path-repairing techniques for route maintenance. When a node detects a broken path, it will notify the source node, which will reroute the connection to a new feasible path, and notify the intermediate nodes along the old path to release the corresponding resources. Unlike the re-routing technique, path-repairing technique does not find a completely new path. Instead, it tries to repair the path using local reconstructions.

Another approach for integrating QoS in the flooding-based route discovery process is proposed in [11]. The proposed positional attribute-based next-hop determination approach (PANDA) discriminates the next hop nodes based on their location or capabilities. When a route-request is broadcasted, instead of using a random rebroadcast delay, the receivers opt for a delay proportional to their abilities in meeting the QoS requirements of the path. The decisions at the receiver side

are made on the basis of a predefined set of rules. Thus the end-to-end path will be able to satisfy the QoS constraints as long as it is intact. A broken path will initiate the QoS-aware route discovery process.

QoS Support using Bandwidth Calculations

Lin et al have proposed an available bandwidth calculation algorithm for ad hoc networks with time division multiple access (TDMA) for communications [17]. This algorithm involves end-to-end bandwidth calculation and bandwidth allocation. Using this algorithm, the source node can determine the resource availability for supporting the required QoS to any destination in the ad hoc networks. This approach is particularly useful in call admission control.

In wired networks, the path bandwidth is the minimum available bandwidth of the links along the path. In time-slotted ad hoc networks, however, bandwidth calculation is much harder. In general, we not only need to know the free slots on the links along the path, but also need to determine how to assign the free slots for each hop. A simple example is illustrated in Figure 3. Time slots 1,2,3 are free between A and B, and slots 2,3,4 are free between B and C. Suppose A wants to send some data to C. Note that there will be collisions at B if A tries to use all three slots 1,2,3 to send data to B while B is using one or both of slots 2, 3 to send data to C. So, we have to somehow divide the common free slots 2, 3 between the two links, namely, from A to B, and from B to C.

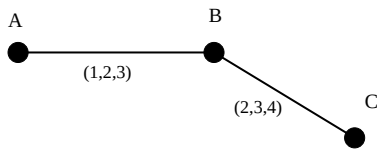

Figure 3: An example of bandwidth calculation in ad hoc networks.

In TDMA systems, time is divided into slots, which in turn are grouped into frames. Each frame contains two phases: control phase and data phase. During the control phase, each node takes turns to broadcast its information to all of its neighbors in a predefined slot. So at the end of control phase, each node has learned the free slots between itself and its neighbors. Based on this information, bandwidth calculation and assignment can be performed distributedly. Deciding slot assignments at the same time as available bandwidth is searched along the path is an NP-complete problem. So Lin et al have proposed a heuristic approach to resolve this issue [17].

An on-demand QoS routing protocol based on AODV is developed for TDMA-based MANETs in [21]. In this approach a QoS-aware route reserves bandwidth from source to destination. In the route discovery process of AODV, a distributed algorithm is used to calculate the available bandwidth on a hop-by-hop basis. Route-request messages with inadequate bandwidth will be dropped by intermediate nodes. Only the destination node can reply to a route-request message that has come along a path with sufficient bandwidth. The protocol can handle limited mobility by restoring broken paths. This approach is applicable for small-sized networks or for short routes.

Multi-path QoS Routing

Liao et al have proposed a multi-path QoS routing protocol [12]. Unlike other existing protocols for ad hoc networks, which try to find a single path between source and destination, this algorithm searches for multiple paths for the QoS route, where the multiple paths refer to a network with a source and a sink satisfying certain bandwidth requirement. The multiple paths collectively satisfy the required QoS. This protocol also adopts the idea of ticket-based probing scheme discussed earlier. The multi-path QoS routing algorithm is suitable for ad hoc networks with very limited bandwidth where a single path satisfying the QoS requirements is unlikely to exist.

3.4 Transport Layer Issues for QoS Provisioning

Transport layer also takes an important role in delivering QoS communications, which mainly involves UDP and TCP protocols. Some real-time applications, like interactive audio/video streams, may be preferably built on top of UDP, which assumes only minimun network functionality and provides much flexibility, while other applications may choose to use TCP, which embodies reliable end-to-end packet delivery and guaranteed in-order packet delivery to applications. In Internet, TCP assumes that most packet losses are due to network congestion. This assumption is not true in the context of wireless networks, where packet losses are mostly due to wireless channel noise and route changes. Whenever a TCP sender detects any packet loss, it will activate its congestion control and avoidance algorithms, which makes TCP perform poorly in term of end-to-end throughput.

TCP performance improvement in mobile wireless networks has been addressed by utilizing a variety of techniques, such as local retransmissions, split-TCP connections, and forward error correction. These schemes attempt either to hide non-congestion-losses from the TCP sender, or to make the TCP sender aware of the existence of wireless hops such that the sender can avoid invoking congestion control algorithms when non-congestion-losses occur. Most of these protocols are, however, designed for infrastructured wireless networks (e.g. cellular networks) and attempt to take advantage of base stations' capabilities in dealing with packet losses that are caused by the high bit error rate of wireless channels caused by hand-off and mobility. These protocols are not suitable for use in infrastructureless environments such as MANETs.

More recently, some work has been done to improve TCP performance over wireless links in MANETs[3, 6, 14], which are dependent on explicit feedback mechanisms to distinguish error losses from congestion losses such that appropriate actions can be taken when packet losses occur. Chandran et al [3] proposed a feedback-based scheme to improve TCP performance over ad hoc networks. When a link breakage is detected by an intermediate node, it will send a Route Failure Notification(RFN) to the sender. Upon receiving a RFN message, the sender will enter snooze state(freeze its retransmission timers and stop sending packets) until a Route Re-establishment Notification(RRN) message comes. A similar approach was proposed by Holland et al[6] on the impact of link breakages on TCP performance. They proposed to use Explicit Link Failure

Notification(ELFN) techniques to improve TCP performance. Upon receiving an ELFN message, a TCP sender will disable its retransmission timers and enter a "stand-by" mode. The sender in "stand-by" mode will send periodic probing packet to check if the route is reestablished or not. Liu et al[14] attempted to address TCP performance problems due to route errors as well as high bit error rate. By inserting a thin layer called ATCP in between TCP and IP layers, this scheme maintains the compatibility with the standard TCP/IP suite. ATCP listens to the network state information provided by Explicit Congestion Notification(ECN) messages and ICMP "destination unreachable" messages, and puts the sender TCP into appropriate state accordingly.

3.5 Application Layer Issues

As mentioned in Section 2, adaptive strategies play very important roles in supporting QoS in MANETs. Application level QoS adaptation is among these adaptive strategies, which includes issues such as a flexible and simple user interface, dynamic QoS ranges, adaptive compression algorithms, joint source-channel coding and joint source-network coding schemes.

A flexible user interface can help achieve easy use of QoS-aware services. Considering the heterogeneous networking environments and user demands in MANETs, it is desirable for the interface to allow users to specify their QoS requirements and able to efficiently map user perceptual parameters into system QoS parameters. Noting its advantages at accommodating imprecision and ambiguity, we believe fuzzy set theory will find its application in achieving the goal of flexible and adaptive QoS services in MANETs.

As proposed in [18], a dynamic QoS range instead of a fixed point of QoS parameters can be used for resource reservation in order to address the dynamic nature of MANETs. This dynamic QoS strategy has implications on the application layer. First, the application must have some notion of the QoS range within which it can operate. These QoS ranges can be programmed in or configured by the user according to her intended use of the application. Second, at the runtime, the application should be able to adapt their behavior based on current feedback from lower layers.

Several approaches have been proposed using application layer techniques for adaptive real-time audio/video streaming over the Internet. These techniques include methods based on compression algorithm features, layered encoding, rate shaping, adaptive error control, and bandwidth smoothing. Most of these techniques were investigated in the context of Internet. Considering the unique characteristic of MANETs, it is conceivable that some modification and improvement must be made to these techniques for use in MANETs. Other techniques are also under investigation, such as joint source-channel coding and joint source-network coding. These joint coding approaches attempt to consider both source characteristics and current channel/network states to achieve better overall performance in transmitting image and real-time audio/video over MANETs.

4 Inter-Layer Design Approaches

In addition to the works on QoS support in individual layers, a few efforts have been directed to the design and implementation of inter-layer QoS frameworks for MANETs. In this section we describe two noteworthy attempts in this direction: INSIGNIA [9] and iMAQ framework [4].

INSIGNIA

The primary design goal of INSIGNIA QoS framework is to support adaptive services which can provide base QoS (i.e. minimum bandwidth) assurances to real-time voice and video flows and data, allowing for enhanced levels (i.e., maximum bandwidth) of service to be delivered when resources become available. INSIGNIA QoS framework is designed to adapt user sessions to the available level of service without explicit signaling between source-destination pairs.

In some QoS routing protocols like CEDAR, the routing protocols interact with resource management to discover and establish end-to-end QoS paths. In such cases, the route discovery and resource reservation are integrated in the QoS routing protocols. Noting that the time scales over which session setup and routing (i.e., computing new routes) operate are distinct and functionally independent tasks, the INSIGNIA designers consider that MANET routing protocols should not be burdened with the integration of QoS functionality that may be tailored toward specific QoS models. Their approach is to develop a QoS framework that can "plug-in" with a wide variety of routing protocols.

The term "in-band signaling" refers to the fact that the control information is carried along with data packets. The term "out-of-band signaling" refers to fact that the control information is typically carried in separate control packets and on channels that may be distinct from the data path. In general, out-of-band signaling system are not good at responding to fast time scale dynamics, because they need to maintain source route information and respond to topology changes by directly notifying the affected nodes to allocate/de-allocate resources. On the contrary, using an in-band signaling approach, the INSIGNIA system can restore the flow-state (i.e., a reservation) in response to topology changes within the interval of a few consecutive IP packets when a standby route is available in cache.

In hard-state connection-oriented communications like virtual circuit, quality of services is guaranteed for the duration of the session. However, these techniques are not suitable in mobile ad hoc networks, where the route discovery and resource reservation need to adapt to topology changes in a timely manner. In MANETs, a soft-state approach for state management at intermediate routing nodes is more flexible for the management of reservations. Soft-state relies on the fact that a source sends data packets along an existing path. When an intermediate mobile router receives a new data packet and no reservation exists, admission control and resource reservation attempt to establish soft state. When a data packet arrives at a mobile router and there exists an associated reservation, the reception of this data packet will refresh the existing soft-state reservation over the next interval. If the soft-state timer times out before a new packet arrives, the associated resources are released. This style of communications is called a "soft connection" when considered on an end-to-end basis and in comparison to the virtual circuit hard-state model.

Figure 4 shows the architectural components of INSIGNIA framework. "In-band signaling" module controls the establishment, restoration, adaptation, and destruction of adaptive QoS-aware paths between source-destination pairs. "Admission control" allocates resources to flows based on base/enhanced QoS request. "Packet forwarding" classifies incoming packets as signaling or data packets, and forwards them to appropriate module. "Routing" adapts to the dynamics of the network and provides routing table to the "packet forwarding" module. "Packet scheduling" responds to location-dependent channel conditions when scheduling packets in a MANET. "Medium access control (MAC)" attempts to hide the underlying media and link layer techniques from the upper IP-based INSIGNIA framework. As a whole, INSIGNIA framework can provide assured adaptive QoS levels to real-time applications, based on the QoS requested by applications and the resource availability in the MANET.

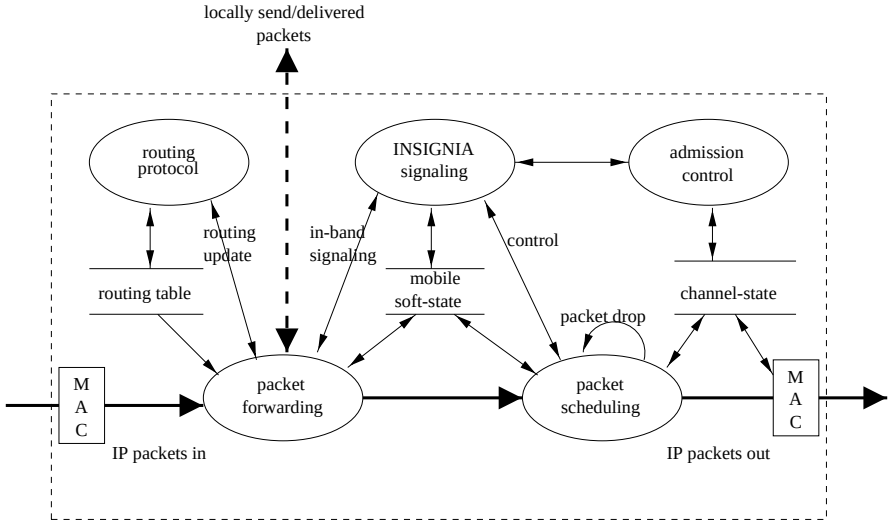

Figure 4: The INSIGNIA framework.

iMAQ Framework

The integrated Mobile Ad-hoc QoS framework (iMAQ) is a cross-layer architecture to support the transmission of multimedia data over a MANET. A model of the framework is shown in Figure 5. The framework involves an ad hoc routing layer and a middleware service layer. At each mobile node, these two layers share information and collaborate to provide QoS assurances to multimedia traffic. The network layer is facilitated with a predictive location-based QoS routing protocol. The middleware layer communicates with the network layer and applications to provide QoS support and maximize overall system QoS satisfaction. The middleware layer also uses location information from the lower network layer and tries to predict network partitioning. In order to provide better data accessibility, it replicates data between different network groups before partitioning occurs. The predictive location-based QoS routing scheme and the data accessibility services are discussed next.

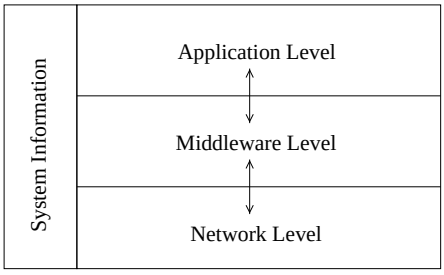

Figure 5: The iMAQ framework model.

In a MANET where mobile nodes may be moving relatively fast and changing direction frequently, the update information may be obsolete when it reaches the correspondent node (in a table-based routing protocol). Even in the case of on-demand routing scheme like Dynamic Source Routing (DSR), the established route is subject to breakage due to intermediate node movement. If a standby route does not exist, there will be a delay before the route is repaired or a new route is computed. To address these problems, a *predictive location-based QoS routing* protocol is proposed, which tries to predict future location of nodes based on their location/resource updates. A mobile node will generate its update message periodically, or when its moving pattern or resource availability has changed considerably. Based on previous updates, the location prediction mechanism will try to predict the time required for a packet to reach its destination (i.e., end-to-end delay), and then based on this delay estimation and destination's location updates, it will try to predict the destination's location at the moment the packet is expected to arrive. When establishing a path, we can choose a best next hop based on our prediction of its future location. This procedure is performed iteratively until the destination is reached. During the course of session, if it is predicted that the route is about to break up due to node movement or resource availability, we can repair the route or compute a new route. Meanwhile, the middleware may re-negotiate QoS with applications when the resource availability degrades.

Based on the location and moving pattern information, the middleware can predict group partitioning in a MANET. Assume that all nodes within a group cooperate to host a set of data that is accessible to each group member. It is a natural idea to improve data accessibility over the network by replicating data into other groups before the predicted partitioning takes place. The middleware data accessibility services include a data lookup service and a data replication service. On each node, the data lookup service maintains a data availability table. Messages advertising about data availability are exchanged between group members periodically. With a soft-state approach, a table entry is refreshed by reception of associated advertising messages. When network partitioning is predicted to happen, certain nodes in different groups are chosen intelligently and data replication is performed in advance.

5 Future Challenges

MANETs are likely to expand their presence in the future communication environments. The support for QoS will thus be an important and desirable component of MANETs. Although

difficult, it is quite interesting and challenging to design and develop QoS provisioning techniques for MANETs. This report provides a survey of the state of the art in this area.

Several important research issues and open questions need to be addressed to facilitate QoS support in MANETs. Use of location, mobility, power consumption, probability of resource and route availability are some of the issues that are currently being examined and need further exploration. It is generally assumed that all nodes in a MANET are "equal" both in capacity and functionality. In capacity, they all support the same wireless communication interfaces. In functionality, they all act as mobile hosts and routers. An interesting question has been raised: "Whether users should be allowed to refuse to be routers, even if this leads to an effectively disconnected network?" Another question arises when we consider some misbehaving nodes in a MANET. A node may misbehave by agreeing to forward packets and then failing to do so, because it is overloaded, selfish, malicious, or broken. Other challenges and open issues include robustness and security, and support for multiple levels of services in QoS routing schemes. Many more similar and other issues will certainly come up as the study and use of MANETs expand. Effective and efficient solutions to these issues will facilitate the design and development of QoS support in MANETs.

References

[1] P. Agrawal, E. Hyden, P. Krzyzanowski, P. Mishra, M.B. Srivastava, and J.A. Trotter, "SWAN: a mobile multimedia multimedia wireless network, " IEEE Personal Communications Volume: 3 Issue: 2 , April 1996.

[2] V. Bharghavan, A. Demers, S. Shenker, and L. Zhang, "MACAW: A Media Access Protocol for Wireless LAN's," Proc of ACM SIGCOMM 1994.

[3] K. Chandran, S. Raghunathan, S. Venkatesan, and R. Prakash, "A feedback-based scheme for improving TCP performance in ad hoc wireless networks," IEEE Personal Communications Magazine, 8(1):34-39, Feb. 2001.

[4] K. Chen, S. H. Shah, and K. Nahrstedt, "Cross Layer Design for Data Accessibility in Mobile Ad Hoc Networks," Journal of Wireless Communications, vol. 21, pp. 49-75, 2002.

[5] M. Gerla and J.T.-C Tsai, "Multicluster, Mobile, Multimedia Radio Network," ACM-Baltzer Journal of Wireless Networks, Vol. 1, No. 3, 1995.

[6] G. Holland and N. Vaidya, "Analysis of TCP performance over mobile ad hoc networks," MobiCom '99, Seattle, Washington, Aug. 1999.

[7] P. H. Hsiao, A. Hwang, H. T. Kung, and D. Vlah, "Load-Balancing Routing for Wireless Access Networks," Proc. of IEEE INFOCOM, pp. 986-995, 2001.

[8] 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.

[9] S. B. Lee, A. Gahng-Seop, X. Zhang, and A. T. Campbell, "INSIGNIA: An IP-Based Quality of Service Framework for Mobile Ad Hoc Networks," Journal of Parallel and Distributed Computing, Special issue on Wireless and Mobile Computing and Communications, Vol. 60 No. 4 pp. 374-406, April 2000.

[10] S.-J. Lee and M. Gerla, "Dynamic Load-Aware Routing in Ad hoc Networks", Proceedings of ICC, Helsinki, Finland, June 2001.

[11] J. Li and P. Mohapatra, "PANDA: A Positional Attribute-Based Next-hop Determination Approach for Mobile Ad Hoc Networks," Technical Report, Department of Computer Science, University of California, Davis, 2002.

[12] W. H. Liao, Y. C. Tseng, S. L. Wang, and J. P. Sheu. "A Multi-Path QoS Routing Protocol in a Wireless Mobile Ad Hoc Network", IEEE Int'l Conf. on Networking (ICN), 2001.

[13] C. R. Lin and M. Gerla, "Asynchronous Multimedia Multihop Wireless Networks," IEEE INFOCOM 1997.

[14] J. Liu and S. Singh, "ATCP: TCP for mobile ad hoc networks," IEEE Jou. of Selected Areas of Communications, 19(7):1300-1315, July 2001.

[15] L. Qian, D.L. Jones, K. Ramchandran, and S. Appadwedula, "A general joint source-channel matching method for error resilient wireless video transmission," Data Compression Conference, 1999, pp. 414-423.

[16] E. M. Royer and C.-K. Toh. "A Review of Current Routing Protocols for Ad-Hoc Mobile Wireless Networks," IEEE Personal Communications Magazine, April 1999, pp. 46-55.

[17] Special Issue on Wireless Ad-Hoc Networks, IEEE Journal on Selected Areas in Communications, Aug. 1999.

[18] D. Thomson, N. Schult, and M. Mirhakkak, "Dynamic Quality-of-Service for Mobile Ad Hoc Networks," MobiHoc 2000, Boston, Massachusetts.

[19] A. Veres, A. T. Campbell, M. Barry and L. H. Sun, "Supporting Service Differentiation in Wireless Packet Networks Using Distributed Control," IEEE Journal of Selected Areas in Communications, Oct. 2001.

[20] Z. Wang and J. Crowcroft, "QoS Routing for Supporting Resource Reservation," IEEE Journal on Selected areas in Communications, pp. 1228-1234, September 1996.

[21] C. Zhu and M. S. Corson, "QoS Routing for Mobile Ad Hoc Networks," INFOCOM 2002.