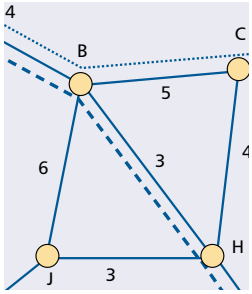
QoS in Mobile Ad Hoc Networks

Prasant Mohapatra, Jian Li, and Chao Gui, University of California



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 need for supporting QoS in these networks is becoming essential.

This work was supported in part by the National Science Foundation under the grants CCR-0296070 and ANI-0296034, and a generous gift from Hewlett Packard Corporation.

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 operate as end hosts as well as routers. They intercommunicate through single-hop and multihop paths in a peer-to-peer fashion. With the expanding scope of applications of MANETs, the need to support QoS in these networks is becoming essential. This article 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 the physical and going up to the application layer. A few proposals on interlayer approaches to QoS provisioning are also addressed. The article concludes with a discussion on the future directions and challenges in the areas of QoS support in MANETs.

Introduction

Wireless mobile networks and devices are becoming increasingly popular as they provide users access to information and communication anytime and anywhere. Conventional wireless mobile communication is usually supported by a wired fixed infrastructure, such as asynchronous transfer mode (ATM) or the Internet. The mobile devices use single-hop wireless radio communications to access a base station that connects 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 multihop paths in a peer-to-peer fashion. Intermediate nodes between a pair of communicating nodes act as routers. Thus, the nodes in MANETs operate as both hosts and routers. The nodes are mobile, 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 MANET.

MANETs are useful in many application environments and do not need any infrastructure support. Collaborative computing and communications in smaller areas (buildings, 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 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 uses of MANETs.

With the increase in quality of service (QoS) needs in evolving applications, it is also desirable to support these services in MANETs. The resource limitations and variability further add to the need for QoS provisioning in such networks. However, the characteristics of these networks make QoS support a very complex process. QoS support in MANETs encompasses issues at the application layer, transport layer, network layer, medium access control (MAC) layer, and physical layer of the network infrastructure. This article provides a detailed survey of the issues involved in supporting QoS across all the protocol layers in MANETs. We classify different approaches, discuss various techniques, and outline the future issues and challenges related to QoS provisioning in MANETs.

The rest of the article is organized as follows. We define the QoS metrics and review the basics of QoS support in MANETs. The QoS issues at all layers of IP are discussed. Interlayer design approaches are described, followed by an outline of future challenges.

Issues in QoS-Aware MANETs

In order to facilitate QoS support in MANETs, we first need to define the metrics to quantify QoS, and understand the difficulties or issues in provisioning QoS in MANETs. In this section we first define QoS and its metrics, followed by an outline of the generic issues and compromising principles in QoS-aware MANETs.

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 is expected to guarantee a set of measurable prespecified service attributes to users in terms of end-to-end performance, such as delay, bandwidth, probability of packet loss, and delay variance (jitter). Power consumption and service coverage area are two other QoS attributes that are more specific to MANETs. QoS metrics can 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

1536-1284/03/$17.00 © 2003 IEEE

delay jitter are additive. The end-to-end delay (jitter) is the accumulation of all delays (jitters) of the links along the path. Furthermore, QoS metrics could be defined in terms of one of the parameters or a set of parameters in varied proportions. Multiconstraint QoS aims to optimize multiple QoS metrics while provisioning network resources, and is an admittedly complex problem.

QoS Support in MANETs: Issues and Difficulties

Mobile multihop wireless networks differ from traditional wired Internet infrastructures. The differences introduce unique issues and difficulties for supporting QoS in MANET environments. These issues include *features* and *consequences*. Examples of features include unpredictable link properties, node mobility, and limited battery life, whereas hidden and exposed terminal problems, route maintenance, and security can be categorized as consequences. These issues 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 multipath cancellation. All these properties make measures such as bandwidth and delay of a wireless link unpredictable.

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

Limited battery life: Mobile devices generally depend on finite battery sources. Resource allocation for QoS provisioning must consider residual battery power and rate of battery consumption corresponding to resource utilization. Thus, all the techniques for QoS provisioning should be power-aware and power-efficient.

Hidden and Exposed Terminal Problems: In a MAC layer with the traditional carrier sense multiple access (CSMA) protocol, multihop packet relaying introduces the "hidden terminal" and "exposed terminal" problems. The hidden terminal problem happens when signals of two nodes, say A and B, that are out of each other's transmission ranges collide at a common receiver, say node C. With the same nodal configuration, an exposed terminal problem will result from a scenario where node B attempts to transmit data (to someone other than A or C) while node C is transmitting to node A. In such a case, node B is exposed to the transmission range of node C and thus defers its transmission even though it would not interfere with the reception at node A.

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 MANETs have to operate with inherently imprecise information. Furthermore, in ad hoc networking environments, nodes can join or leave at any time. The established routing paths may be broken even during the process of data transfer. Thus, the need arises for maintenance and reconstruction of routing paths with minimal overhead and delay.

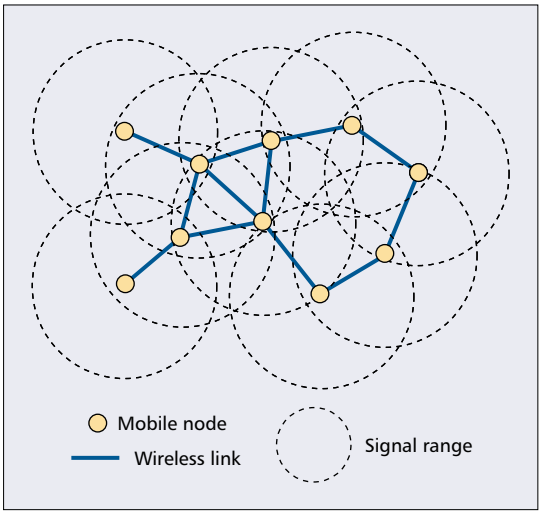


Figure 1. *A mobile ad hoc network.*

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, reservation maintenance with updates in the routing path becomes cumbersome.

Security: Security can be considered a QoS attribute. Without adequate security, unauthorized access and usage may violate QoS negotiations. The nature of broadcasts in wireless networks potentially results in more security exposure. The physical medium of communication is inherently insecure, so we need to design security-aware routing algorithms for MANETs.

Compromising Principles

The dynamic nature of MANETs is attributed to several inherent characteristics, such as variable link behavior, 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 [1]. Soft QoS means that after 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 [2], 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 reallocation. A good example of this case is layered real-time video, which requires a minimum bandwidth assurance and allows for an enhanced level of QoS when additional resources are available. QoS adaptation can be also done at various

Soft QoS means that after 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.

One of the major challenges in supporting QoS communication over wireless media is channel estimation, which 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.

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). A more sophisticated technique involves an adaptive error correction mechanism that increases or decreases the amount of error correction coding in response to changes in transmission quality or desired QoS. 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.

QoS from a Layered Perspective

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

QoS Support in Physical Channels

The wireless channel in a MANET is time-varying, which means that the signal-to-noise ratio in channels fluctuates with respect to time. Thus, adaptive modulation that can tune many possible parameters according to current channel state (e.g., instantaneous signal-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, which involves accurate channel estimation at the receiver and then reliable feedback of the estimation to the transmitter so that the transmitter and receiver can be properly synchronized. Perfect synchronization, although highly desirable, is almost impossible to achieve in MANETs. The time-varying fading channel also makes coding schemes designed for a fixed channel model unsuitable for use in MANETs. Wireless channel coding needs to address the problems introduced by channel or 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 tight integration with upper layers, such as source compression algorithms at the application layer. Use of higher source coding rates (less data compression) can decrease the final end-to-end distortion. Similarly, 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 trade-off between these two rates. Joint source-channel coding takes both source characteristics and current channel situation into consideration.

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 and exposed terminal problems. Multihop access collision avoidance (MACA) [3] is proposed to solve these problems through the request-to-send/clear-to-send (RTS/CTS) dialogs, but does not completely eliminate the hidden terminal problem. MACAW [4] was proposed as an extension to MACA to provide faster recovery from hidden terminal collisions. The IEEE 802.11 standard specification includes the collision avoidance feature of MACA and MACAW by its distributed control function (DCF). Its fundamental access method is carrier sense multiple access with collision avoidance (CSMA/CA), which solves the hidden terminal problem completely. However, it does not provide real-time traffic support. In this section we survey the MAC layer QoS issues proposed for MANETs.

IEEE 802.11 DCF and Its Extension — IEEE 802.11 is a CSMA/CA protocol. In DCF mode, after the node has sensed the medium to be idle for a time period longer than distributed interframe space (DIFS), it begins transmitting. Otherwise, the node defers transmission and starts to back off. Each node holds a value called a 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 set to a random value between 0 and CW. Whenever the medium becomes idle for a period longer than DIFS, the backoff timer is decremented. As soon as the timer expires, the node starts transmission. To improve performance by reducing packet collisions, the sender will first send a short RTS packet if the data packet is longer than a threshold value. If the intended receiver grants the request, it will return a short CTS packet. 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 a best effort control algorithm. It has no notion of service differentiation and no support for real-time traffic. Veres *et al.* [1] have proposed a scheme to extend IEEE 802.11 DCF with the ability to support at least two service classes, premium service (i.e., high-priority) 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 by Sobrinho *et al.* [5] avoids packet collision in a very novel way, 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 a fairness guarantee. Nodes contend for the medium after it has been idle for a period longer than the interframe space. Nodes with best effort traffic and those with real-time traffic use different interframe space values. This makes real-time traffic as a group have higher priority over data nodes. A BB contention scheme is added to any CSMA/CA protocol in the following manner. Right before sending their packets when the medium remains idle long enough, real-time

nodes first contend for transmission rights by jamming the media with pulses of energy called BBs. The novelty of this scheme is that each contending node uses a BB of different length. The length of each BB is an integral number of black slots, each slot of a given length. The number of slots that forms a BB is an increasing function of the contention delay experienced by the node, measured from the instant an attempt to access the channel is scheduled until the node starts transmission of its BB. Following each BB transmission, a node senses the channel for an observation interval. Since distinct nodes contend with BBs of different length, each node can determine without ambiguity whether its BB is of longest duration. Thus, only one winner is produced after this contention, and it 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 — Multihop access collision avoidance with piggyback reservation (MACA/PR) [6] 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 a 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 that hears the RTS will defer long enough so the sender can receive the returning CTS. Any node near the receiver that hears the CTS will avoid colliding with the following data packet. The RTS/CTS dialog is used only for the first packet to set up 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 that hear the data packet can learn about the next packet transmission time. Likewise, neighbors at the receiver side that hear the ACK will avoid sending at the time the receiver is scheduled to receive the next packet. Notice that the ACK serves to renew a reservation rather than to recover from packet loss. In fact, if the ACK is not received, the packet is not retransmitted. Instead, if the sender consecutively fails to receive ACKs for a certain number of transmissions, 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 a protector for the given time window, and a mechanism to inform the sender if something is wrong on the link.

QoS-Aware Routing at the Network Layer

Several routing protocols have been proposed for MANETs, which can be classified into three broad categories:

- Proactive table-based routing schemes
- Reactive on-demand source-based routing schemes
- Constraint-based routing schemes

The proactive 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 schemes 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) [7]. Flat routing schemes require maintenance of the state of the entire network at all nodes, which limits its scalability. In the hierarchical approach, the state of only a subset of the network is maintained at all nodes, and routing is facilitated through another level of state information, which is stored in fewer nodes [7].

In the case of on-demand source-based routing schemes, 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 the Temporary Ordered Routing Algorithm (TORA) [7]. These algorithms focus on finding the shortest path between the source and destination nodes by considering the node status and network configuration when a route is desired.

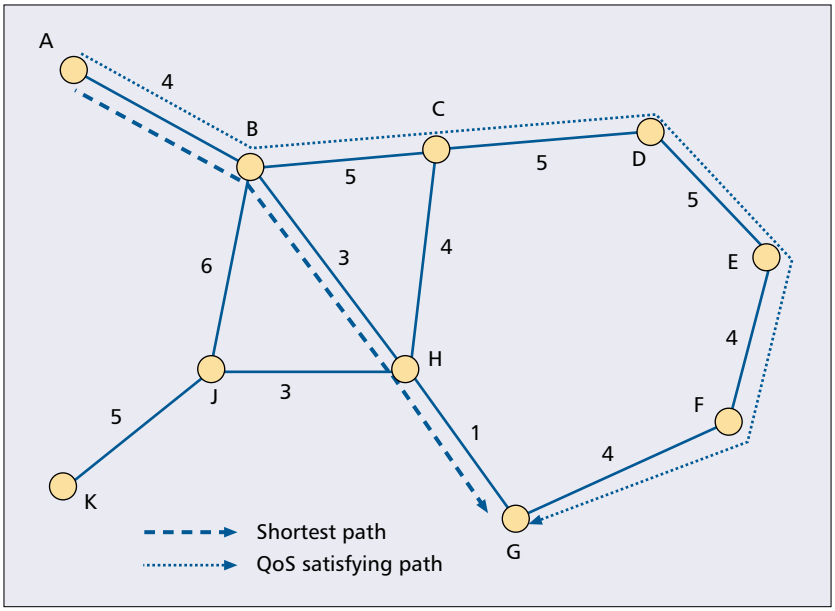
Constraint-based routing protocols use metrics other than the shortest path to find a suitable and feasible route. Associativity-based routing (ABR) and signal stability routing (SSR) [7] take into account the node's signal strength and location stability so that the path chosen is more likely to be long-lived. Dynamic load-aware routing (DLAR) [8] considers the load of intermediate nodes as the primary route selection metric.

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 network topology derived from Fig. 1. The mobile nodes are labeled 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 source node A to destination node G. For conventional routing using the shortest path (in terms of the number of hops) as a metric, the route A-B-H-G would be chosen. However, the QoS-based route selection process could select a completely different path. 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 to provide the required bandwidth.

The primary goal of the QoS-aware routing protocols is to determine a path from a source to the

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.



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

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

CEDAR — The Core Extraction Distributed Ad Hoc Routing (CEDAR) algorithm is proposed as a QoS routing scheme for small to medium-sized ad hoc networks consisting of tens to hundreds of nodes [5]. It dynamically establishes the core of the network, and then incrementally propagates the link states of stable high-bandwidth links to the core nodes. Route computation is on demand, and is performed by the core nodes using only local state. CEDAR has three key components:

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 to perform route computations. The core nodes are elected by approximating a minimum dominating set[1] of the ad hoc network.

Link 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.

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 low-overhead infrastructure to perform routing, while the state propagation mechanism ensures 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 [4]. 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 its 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 levels of route redundancy. It also uses rerouting 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 rerouting, path repairing does not find a completely new path. Instead, it tries to repair the path using local reconstructions.

Another approach to integrating QoS in the flooding-based route discovery process is proposed in [9]. The proposed positional attribute-based next-hop determination approach (PANDA) discriminates next-hop nodes based on their location or capabilities. When a route request is broadcast, instead of using a random rebroadcast delay, the receivers opt for a delay proportional to their abilities to meet 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 [5]. 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

[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.*

minimum available bandwidth of the links along the path. In time-slotted ad hoc networks, however, bandwidth calculation is much harder. In general, we need not only to know the free slots on the links along the path, but also to determine how to assign the free slots at each hop. A simple example is illustrated in Fig. 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, from A to B and from B to C.

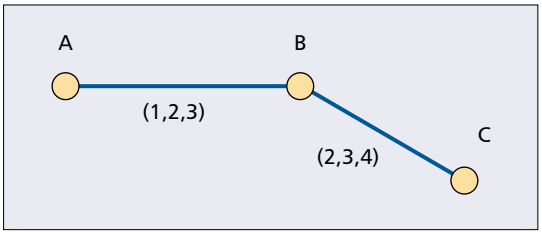
In TDMA systems, time is divided into slots, which in turn are grouped into frames. Each frame contains two phases: control and data. During the control phase, each node takes turns broadcasting its information to all its neighbors in a predefined slot. So at the end of the control phase, each node has learned the free slots between itself and its neighbors. Based on this information, bandwidth calculation and assignment can be performed distributedly. Determining slot assignments while searching for the available bandwidth along the path is an NP-complete problem. So Lin *et al.* have proposed a heuristic approach to resolve this issue [5].

An on-demand QoS routing protocol based on AODV is developed for TDMA-based MANETs in [10]. 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 short routes.

Multi-path QoS Routing — Liao *et al.* have proposed a multipath QoS routing protocol [11]. 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 a certain bandwidth requirement. The multiple paths collectively satisfy the required QoS. This protocol also adopts the idea of ticket-based probing discussed earlier. The multipath 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.

Transport Layer Issues for QoS Provisioning

The transport layer also plays an important role in delivering QoS communications, mainly involving UDP and TCP protocols. Some real-time applications, such as interactive audio/video streams, may be preferably built on top of UDP, which assumes only minimum 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 the 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 activates its congestion control and avoidance algorithms, which makes TCP perform poorly in terms of end-to-end throughput in MANETs.



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

TCP performance improvement in mobile wireless networks has been addressed by a variety of techniques, such as local retransmissions, split-TCP connections, and forward error correction. These schemes attempt to either hide noncongestion losses from the TCP sender or make the TCP sender aware of the existence of wireless hops so that the sender can avoid invoking congestion control algorithms when noncongestion losses occur. Most of these protocols, however, are designed for infrastructured wireless networks (e.g., cellular networks) and attempt to take advantage of base stations' capabilities in dealing with packet losses 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 [12–14], which are dependent on explicit feedback mechanisms to distinguish error losses from congestion losses so that appropriate actions can be taken when packet losses occur.

Incorporation of appropriate techniques for performance and resource management in the transport layer protocols helps in provisioning end-to-end QoS in MANETs.

Application Layer Issues

As mentioned earlier, adaptive strategies play a very important role in supporting QoS in MANETs. Application-level QoS adaptation belongs to these adaptive strategies, including 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 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

In the 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.

Inter-node communication in MANETs is an expensive operation in terms of bandwidth and energy consumption. Thus it is critical to design efficient intercommunication protocols to conserve scarce resources — something difficult to achieve following the architectural philosophy of the strict separation of the protocol layer functionalities.

application in achieving the goal of flexible and adaptive QoS services in MANETs.

As proposed in [2], 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 or configured by the user according to their intended use in application. Second, at runtime, the application should be able to adapt its behavior based on 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 the Internet. Considering the unique characteristics 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 (e.g., 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.

Interlayer Design Approaches

Internode communication in MANETs is an expensive operation in terms of bandwidth and energy consumption. Thus, it is critical to design efficient intercommunication protocols to conserve scarce resources — difficult to achieve following the architectural philosophy of strict separation of the protocol layer functionalities. To achieve better efficiency while conserving resources in internode communications, interlayer or cross-layer issues must be explored. A few efforts have been directed to the design and implementation of interlayer QoS frameworks for MANETs.

In this section we describe two noteworthy attempts in this direction: INSIGNIA [15] and the iMAQ framework [16].

INSIGNIA — The primary design goal of the INSIGNIA QoS framework is to support adaptive services that 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 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, route discovery and resource reservation are integrated in the QoS routing protocols. Noting that the timescales over which session setup and routing (i.e., computing new routes) operate are distinct and functionally independent, the INSIGNIA designers consider that MANET routing protocols should not be burdened with 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 control information being carried along with data packets. The term *out-of-band signaling* refers to the control information being typically carried in separate control packets and on channels that may be distinct from the data path. In general, out-of-band signaling systems are not good at responding to fast timescale dynamics, because they need to maintain source route information and respond to topology changes by directly notifying the affected nodes to allocate/deallocate 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, QoS is guaranteed for the duration of the session. However, these techniques are not suitable in MANETs, where route discovery and resource reservation need to adapt to topology changes in a timely manner. In MANETs, a soft state approach to 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 is 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. The INSIGNIA 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 requests. Packet forwarding classifies incoming packets as signaling or data packets, and forwards them to an appropriate module. The routing protocol adapts to the dynamics of the network and provides a 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 can provide assured adaptive QoS levels to real-time applications, based on the QoS requested by applications and the resource availability in the MANET.

iMAQ Framework — The integrated mobile ad hoc QoS framework (iMAQ) is a cross-layer architecture to support transmission of multimedia data over a MANET. A model of the framework is

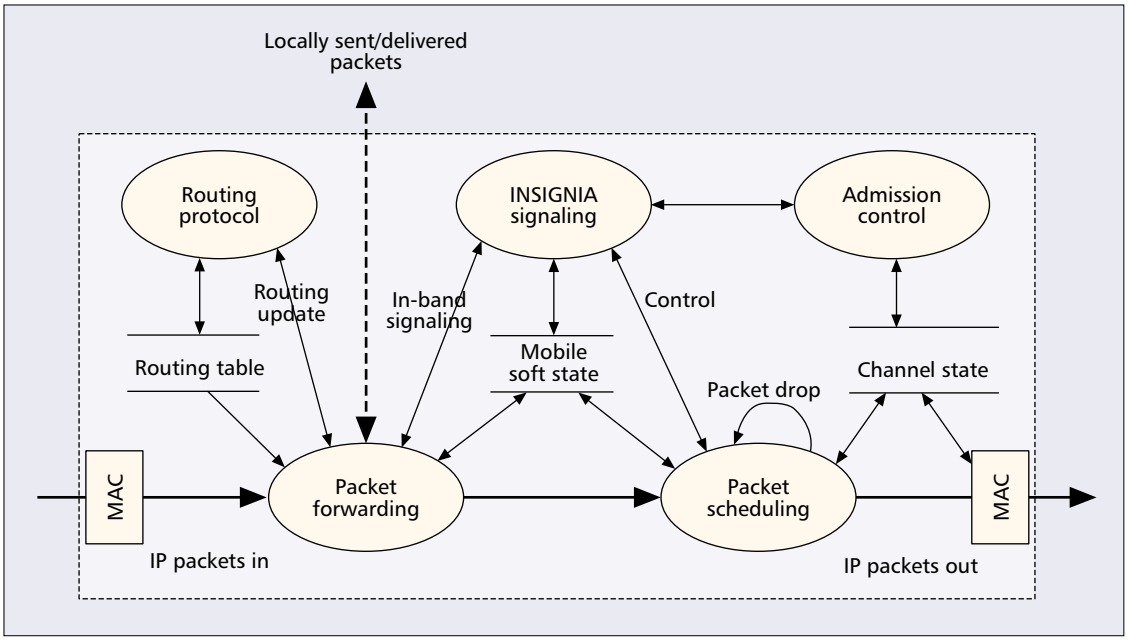


Figure 4. *The INSIGNIA framework.*

shown in Fig. 5. The framework involves the network layer (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 data accessibility services are discussed next.

In a MANET where mobile nodes may move relatively fast and change direction frequently, update information may be obsolete when it reaches the correspondent node (in a table-based routing protocol). Even in the case of an 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, 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 a 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 renegotiate 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 data lookup and data replication services. On each node the data lookup service maintains a data availability table. Messages advertising 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, certain nodes in different groups are chosen intelligently, and data replication is performed in advance.

Future Challenges

MANETs are likely to expand their presence in future communication environments. 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 currently being examined and needing further exploration. It

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.

MANETs are likely to expand their presence in 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.

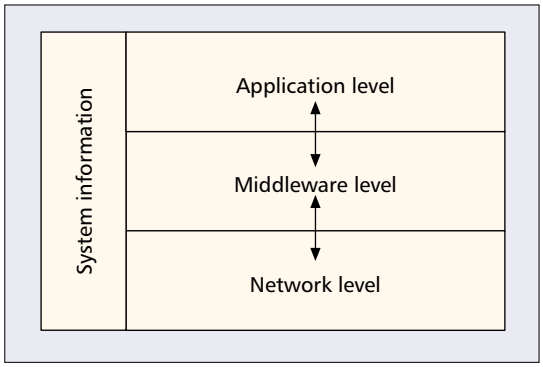


Figure 5. *The iMAQ framework model.*

is generally assumed that all nodes in a MANET are homogeneous in terms of both capacity and functionality. QoS issues in heterogeneous MANETs should be investigated as the issues in heterogeneous MANETs are different than those of homogeneous MANETs. 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, security, and support for multiple levels of services in QoS routing schemes. Many 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] A. Veres *et al.*, "Supporting Service Differentiation in Wireless Packet Networks Using Distributed Control," *IEEE JSAC*, Oct. 2001.

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

[3] P. Karn, "MACA -a New Channel Access Method for Packet Radio," *ARRL/CRRL Amateur Radio 9th Comp. Net. Conf.*, 1990, pp. 134–40.

[4] V. Bharghavan *et al.*, "MACAW: A Media Access Protocol for Wireless LANs," *Proc. ACM SIGCOMM 1994*.

[5] Special Issue on Wireless Ad Hoc Networks, *IEEE JSAC*, Aug. 1999.

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

[7] E. M. Royer and C.-K. Toh. "A Review of Current Routing Protocols for Ad-Hoc Mobile Wireless Networks," *IEEE Pers. Commun.*, Apr. 1999, pp. 46–55.

[8] S.-J. Lee and M. Gerla, "Dynamic Load-Aware Routing in Ad Hoc Networks," *Proc. ICC*, Helsinki, Finland, June 2001.

[9] J. Li and P. Mohapatra, "PANDA: A Positional Attribute-Based Next-hop Determination Approach for Mobile Ad Hoc Networks," Tech. rep., Dept. of Comp. Sci., UC Davis, 2002.

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

[11] W. H. Liao *et al.*, "A Multi-Path QoS Routing Protocol in a Wireless Mobile Ad Hoc Network," *IEEE ICN*, 2001.

[12] G. S. Ahu *et al.*, "SWAN: Service Differentiation in Stateless Wireless Ad Hoc Networks," *Proc. INFOCOM 2002*.

[13] K. Chandran *et al.*, "A Feedback-Based Scheme for Improving TCP Performance in Ad Hoc Wireless Networks," *IEEE Pers. Commun.*, vol. 8, no. 1, Feb. 2001, pp. 34–39.

[14] G. Holland and N. Vaidya, "Analysis of TCP Performance over Mobile Ad hoc Networks," *ACM MobiCom*, Seattle, WA, Aug. 1999.

[15] S. B. Lee *et al.*, "INSIGNIA: An IP-Based Quality of Service Framework for Mobile Ad Hoc Networks," *J. Parallel and Distrib. Comp.,* Special issue on Wireless and Mobile Computing and Communications, vol. 60 no. 4, Apr. 2000, pp. 374–406.

[16] K. Chen, S. H. Shah, and K. Nahrstedt, "Cross Layer Design for Data Accessibility in Mobile Ad Hoc Networks," *J. Wireless Commun.*, vol. 21, 2002, pp. 49–75.

Biographies

PRASANT MOHAPATRA (prasant@cs.ucdavis.edu) is currently an associate professor in the Department of Computer Science, University of California, Davis (UC Davis). In the past he was on the faculty at Iowa State University and Michigan State University. He has also held visiting scientist positions at Intel Corporation and Panasonic Technologies. He received his Ph.D. in computer engineering from the Pennsylvania State University in 1993. He is currently on the editorial board of *IEEE Transactions on Computers*. He was also co-editor of the January 2003 issue of *IEEE Network*. His research interests are in the areas of wireless mobile networks, Internet protocols and QoS, and Internet servers.

JIAN LI (lijian@cs.ucdavis.edu) received B. Eng. and M. Eng. degrees from Tsinghua University, Beijing, China, in 1997 and 2000, respectively. He is currently working toward a Ph.D. degree in Networks Research Laboratory, Department of Computer Science, UCDavis. His research interests include computer networks, MANETs, and sensor networks.

CHAO GUI (guic@cs.ucdavis.edu) received a B.S. degree from Huazhong University of Science and Technology, China, in 1996, and an MS degree from the University of Central Florida in 2001. Currently, he is a Ph.D. student in the Computer Science Department, UCDavis, where he works as a research assistant in the Networks Laboratory. His research interests include MANETs and wireless sensor networks.