A Comparative Analysis of Performance and QoS Issues in MANETs

Javed Parvez\textsuperscript{1} and Mushtaq Ahmad Peer\textsuperscript{2}

Abstract—Mobile Ad hoc networks (MANETs) are collections of wireless mobile nodes dynamically reconfiguring and collectively forming a temporary network. These types of networks assume existence of no fixed infrastructure and are often useful in battle-field tactical operations or emergency search-and-rescue type of operations where fixed infrastructure is neither feasible nor practical. They also find use in ad hoc conferences, campus networks and commercial recreational applications carrying multimedia traffic. All of the above applications of MANETs require guaranteed levels of performance as experienced by the end-user. This paper focuses on key challenges in provisioning predetermined levels of such Quality of Service (QoS). It also identifies functional areas where QoS models are currently defined and used. Evolving functional areas where performance and QoS provisioning may be applied are also identified and some suggestions are provided for further research in this area. Although each of the above functional areas have been discussed separately in recent research studies, since these QoS functional areas are highly correlated and interdependent, a comprehensive and comparative analysis of these areas and their interrelationships is desired. In this paper we have attempted to provide such an overview.

Keywords—Bandwidth Reservation, Congestion, Dynamic Network Topology, End-to-End Delay, Flexible QoS Model for MANET (FQMM), Hidden Terminal, Mobile Adhoc Network (MANET), Packet Jitter, Queuing, Quality-of-Service (QoS), Relative Bandwidth Service Differentiation (RBSD), Resource ReSerVation Protocol (RSVP).

I. INTRODUCTION

QoS is the performance level of a service offered by the network to the user. The primary goal of QoS provisioning is to achieve more deterministic behavior by proper utilization of the network resources. A network or a service provider can offer different kinds of services to the users based on a set of service requirements such as minimum bandwidth, maximum delay, maximum variance in delay and maximum rate of packet loss. After accepting a service request from the user, the network is expected to ensure the committed service requirements of the users throughout the communication. QoS provisioning is challenging due to the key characteristics of MANETs i.e. lack of central coordination, mobility of hosts and limited availability of resources.

QoS support for MANETs encompasses four major functionally interdependent areas i.e. QoS models, QoS resource reservation signaling, QoS Routing and QoS Medium Access Control (MAC). A QoS model defines the service architecture of the overall QoS framework. QoS routing focuses on identification of network paths with sufficient capacity to meet end-user service requirements. QoS MAC provides mechanisms for resolving medium contention and supports reliable unicast links.

II. BACKGROUND

From the perspective of the roles that they play in supporting QoS in MANETs, the fundamental building blocks of a QoS architectural framework can be broken down into the following modules.

1. Admission Control

Admission control policies are generally tied to ISP or service level agreements (SLA) between a subscriber and the ISP. They may be additionally based on the availability of adequate network resources in an attempt to meet the performance objectives of a particular service request. Policies may be parameter based, if predefined hard-QoS guarantees are desired; otherwise measurement-based policies are used for soft-QoS i.e. relative service assurance. Regulation of new traffic to ensure that it does not lead to network overload or service degradation in existing traffic is the primary responsibility of this module.

2. Traffic Classification & Scheduling

Scheduling is based on a service rate allocation to classes of traffic that share a common buffer. It is the mechanism that selects a packet for transmission from the packets waiting in the transmission queue. Packet scheduling thus controls bandwidth allocation to different nodes or types of applications. The desired service guarantees are realized independently at each router via proper scheduling.

3. QoS – Hard vs Soft State

Maintaining the QOS of adaptive flows in MANETs is one of the most challenging aspects of the QOS framework. Typically, wire line networks have little quality of service or state management where the route and the reservation between
source-destination pairs remain fixed for the duration of a session. This style of hard-state connection-oriented communications (e.g. virtual circuit) guarantees QoS for the duration of the session holding time. However, these techniques are not flexible enough in MANETs where the paths and reservations need to dynamically respond to topology changes in a timely manner. Therefore, soft-state approach to state management at intermediate routing nodes is a suitable approach for the management of reservations in MANETs. It relies on the fact that a source sends data packets along an existing path. If a data packet arrives at a mobile router and no reservation exists, then admission control and resource reservations attempt to establish soft state. Subsequent reception of data packets (associated with the reservation) at that router are used to refresh the existing soft-state reservation. This is called a soft-connection when considered on an end-to-end basis and in relation to the virtual circuit hard-state model. When an intermediate node receives a data packet that has an existing reservation, it reconfirms the reservation over the next interval.

4. Buffer Management
Buffer management deals with the task of either storing or dropping a packet awaiting transmission. The key mechanisms of buffer management are the backlog controller and the dropper. The backlog controller specifies the time instances when traffic should be dropped, and the dropper specifies the traffic to be dropped. Buffer management is often associated with congestion control. As an example, consider one of the UDP segments generated by an IP phone application. The UDP segment is encapsulated in an IP datagram. As the datagram wanders through the network, it passes through buffers (i.e. queues) in the routers in order to access outbound links. It is possible that one or more buffers in the route from the sender to receiver is full and cannot admit the IP datagram. In this case, the IP datagram is discarded, never to arrive at the receiving application. Therefore, a mechanism to deal with the packet loss is desired.

5. Resource Reservation
Resource reservation is typically performed with a signaling mechanism such as RSVP or INSIGNIA. Using such a mechanism, the network sets aside the required resources on demand for delivering the desired network performance. This is in general closely associated with admission control. Since charges are normally based on the use of reserved resources, resource reservation requires the support of authentication, authorization, accounting and settlement between ISPs.

6. Packet Jitter
A crucial component of end-to-end delay is the random queuing delays in the routers. Because of these varying delays within the network, the time from when a packet is generated at the source until it is received at the receiver can fluctuate from packet to packet. This phenomenon is called jitter.

7. End-to-End Delay
End-to-End delay is the accumulation of transmission, processing and queuing delays in routers; propagation delay in the links; and end-system processing delays. For a highly interactive application such as IP phone, end-to-end delays smaller than 150 ms are not perceived by human listeners. Lesser end-to-end delay implies better performance.

III. QOS PROVISIONING CHALLENGES IN MANET
Several research studies have been conducted on providing QoS support in conventional wireless networks. Such wireless networks often require a fixed wireline backhaul through which mobile hosts can connect to the wireline base stations in a one-hop radio transmission. In MANETs no such fixed infrastructure may exist. Thus, providing QoS support in MANETs is more challenging than conventional wireless networks. A summary of the major challenges in providing QoS support in MANETs has been presented in [26, 48].

1. Dynamic network topology: In MANETs there is no restriction on mobility. Thus the network topology changes dynamically causing hosts to have imprecise knowledge of the current status. A QoS session may suffer due to frequent path breaks, thereby requiring re-establishment of new paths. The delay incurred in re-establishing a QoS session may cause some of the packets belonging to that session to miss their delay targets and/or deadlines, which is not acceptable for applications that have stringent QoS requirements.

2. Error-prone wireless channel: The wireless radio channel by nature is a broadcast medium. The radio waves suffer from several impairments such as attenuation, thermal noise, interference, shadowing and multi-path fading effects during propagation through the wireless medium. This makes it difficult to ensure QoS commitments like hard packet delivery ratio or link longevity guarantees.

3. Lack of central coordination: Like wireless LAN and cellular networks, a MANET does not have central controllers to coordinate the activity of the nodes. A MANET may be set up spontaneously without planning and its members can change dynamically, thus making it difficult to provide any form of centralized control. As a result communications protocols in MANET utilize only locally available state and operate in a distributed manner [43]. This generally increases the overhead and complexity of an algorithm as QoS state information must be disseminated efficiently.

4. Imprecise state information: The nodes in a MANET mostly maintain link-specific as well as flow-specific state information. The link-specific state information comprises bandwidth, delay, delay jitter, loss rate, error rate, stability, cost and distance values for each link. The flow-specific information includes session ID, source address, destination address and QoS requirements of the flow. Due to dynamic changes in network topology and channel characteristics, these state information are inherently imprecise. This may result in inaccurate routing decisions resulting in some packets missing their deadlines, leading to violation of real-time QoS commitment.

5. Limited availability of resources: Although mobile
devices are becoming increasingly powerful and capable, it still holds true that such devices generally have less computational power, less memory and a limited (battery) power supply, compared to devices such as desktop computers typically employed in wired networks. This factor has a major impact on the provision of QoS assurances, since low memory capacity limits the amount of QoS state that can be stored, necessitating more frequent updates incurring greater overhead. Often QoS routing problems, most of which are NP-complete and require complicated heuristics for solving them, place an excessive strain on mobile node processors leading to higher consumption of limited battery power.

6. Hidden Terminal Problem: The hidden terminal problem is inherent in MANET. This problem occurs when packets originating from two or more sender nodes, which are not within the direct transmission range of each other, collide at a common receiver node. It necessitates retransmission of packets, which may not be acceptable for flows that have strict QoS requirements. Some control packet exchange mechanisms [30, 32] reduce the hidden terminal problem only to a certain extent.

IV. QoS PROTOCOL PERFORMANCE ISSUES/FACTORS

Even after overcoming the challenges of MANET, a number of factors [26] have major impacts when evaluating the performance of QoS protocols. Some of these parameters are of particular interest considering the characteristics of the MANET environment. They can be summarized as follows:

Node mobility: This parameter has been the focus of research studies such as [11]. This factor generally encompasses several parameters: the nodes' maximum and minimum speed, speed pattern and pause time. The node's speed pattern determines whether the node moves at uniform speed at all times or whether it is constantly varying, and also how it accelerates, for example, uniformly or exponentially with time. The pause time determines the length of time nodes remain stationary between each period of movement. Together with maximum and minimum speed, this parameter determines how often the network topology changes and thus how often network state information must be updated.

Network size: Since QoS state has to be gathered or disseminated in some way for routing decisions to be made, the larger the network, the more difficult this becomes in terms of update latency and message overhead. This is the same as with all network state information, such as that used in best-effort protocols [43];

Number, type and data rate of traffic sources: Intuitively, a smaller number of sources can be constant bit rate (CBR) or may generate bits or packets at a rate that varies with time according to the Poisson distribution, or any other mathematical model. The maximum data rate affects the number of packets in the network and hence the network load. All of these factors affect performance significantly [46];

Node transmission power: Some nodes may have the ability to vary their transmission power. This is important, since at a higher power, nodes have more direct neighbors and hence connectivity increases, but the interference between nodes increases as well. Transmission power control can also result in unidirectional links between nodes, which can affect the performance of routing protocols. This factor has also been studied extensively, e.g. [40, 13, 22];

Channel characteristics: As detailed earlier, there are many reasons for the wireless channel being unreliable, i.e. many reasons why bits, and hence data packets, may not be delivered correctly. These all affect the network's ability to provide QoS.

V. QoS MODELS

QoS model defines the service architecture of the total QoS framework. That is, it defines the type of services a QoS model should provide and the classifications of the services. It also sets the system level goal that all other components like QoS signaling, QoS Routing and QoS MAC layer of QoS framework should implement.

A lot of work is done to support QoS in the Internet. A QoS model for MANET, however, should be able to overcome the challenges of MANET, e.g. dynamic topology and time-varying link capacity. The QoS model for MANET has extended the traditional Internet models to make them suitable for MANET. We are thus motivated to elaborate the existing QoS models for the Internet such as IntServ [9] and DiffServ [7] as background and highlight their incompatibility for MANET. We then describe FQMM and SWAN - the QoS models for MANET.

Integrated Service (IntServ) Model

IntServ is an existing QoS framework developed within the IETF to provide QoS guarantees to individual application sessions within the Internet. The IntServ model merges the advantages of two different paradigms: datagram networks and circuit switched networks. It can provide a circuit switched service in packet-switched networks by treating an application session between a pair of end users as a flow. The basic idea of the IntServ model is that the flow-specific states are kept in every IntServ-enabled router [27].

IntServ proposes two basic service classes: Guaranteed Service [53] and Controlled Load Service [18]. The Guaranteed Service is provided for applications requiring fixed delay-bound. The Controlled Load Service is for applications requiring reliable and enhanced best effort service. In an IntServ-enabled router, IntServ is implemented with four main components [27]: the signaling protocol, the admission control routine, the classifier and the packet scheduler. Other components, such as the routing agent and management agent, which are the essential mechanisms within the routers are kept intact. The Resource ReSerVation Protocol (RSVP) [10] is used as the signaling protocol to reserve resources in IntServ. Applications with Guaranteed Service or Controlled-Load Service requirements use RSVP to reserve resources before transmission.
Admission control, which is used to decide whether to accept the resource requirement, is invoked at each router to make a local accept/reject decision at the time that a host requests a real-time service along some paths through the Internet. Admission control notifies the application through RSVP if the QoS requirement can be granted and the application accordingly transmits its data. When a router receives a data packet, the classifier will perform a Multi-Field (MF) classification [25], which classifies a packet based on multiple fields such as source and destination addresses, source and destination port numbers, Type Of Service (TOS) bits and protocol ID in the IP header. Then the classified packet will be put into a corresponding queue according to the classification result. Finally, the packet scheduler reorder the output queue to meet different QoS requirements.

IntServ/RSVP model is not ideally suitable for MANETs due to the resource limitation in MANETs: (1) the amount of state information increases proportionally with the number of flows. Keeping flow-state information will cost a huge storage and processing overhead for the mobile host whose storage and computing resources are scarce. Although the scalability problem is unlikely to arise in current MANETs due to their limited bandwidth and relatively small number of flows compared with wired networks, it may arise in the near future due to the development of fast radio technology and potentially large number of users; (2) the RSVP signaling packets will contend for bandwidth with the data packets and consume a substantial percentage of bandwidth in MANETs; (3) every mobile host must perform the processing of admission control, classification and scheduling which imposes a heavy burden for the resource-limited mobile hosts.

**Differentiated Service (DiffServ) Model**

The DiffServ architecture, on the other hand, is designed to overcome the difficulty of implementing and deploying IntServ and RSVP in the Internet backbone [27]. DiffServ also addresses the scalability issues of IntServ by providing a limited number of aggregated classes.

In DiffServ architecture, an edge router controls the traffic entering the network by classifying, marking, policing and shaping mechanism. The policy manager, in the router, assures that no one will violate the type of service it is pre-assigned. DiffServ also defines the layout of the Type of Service (TOS) bits in the IP header, called the DS field, and a base set of packet forwarding rules, called Per-Hop-Behavior (PHB). When a data packet enters a DiffServ-enabled domain, an edge router marks the packet’s DS field, and the interior routers along the forwarding path forward the packet based on its DS field. Since the DS field only codes very limited service classes, the processing of the core routers is very simple and fast.

Core routers in DiffServ do not need to keep per-flow state information (Fig. 1). Many services, such as Premium Service [41] and Assured Service [18,28 ] (Nichols et al. 1999) can be supported in the DiffServ model. Premium Service is supposed to provide low loss, low delay, low jitter, and end-to-end assured bandwidth service. Assured Service is for applications requiring better reliability than Best Effort Service. Its purpose is to provide guaranteed or at least expected throughput for applications.

Because of easy deployment and lightweight core node requirement, DiffServ may be a possible solution to the MANET QoS model. In addition, it provides Assured Service, which is a desirable service context in MANETs.

![Fig. 1 Flow management in DiffServ. Interior flows are aggregated in the core router and the core routers do not keep per flow information.](image)

However, since DiffServ is designed for fixed, wired networks, there are still some residual challenges in implementing DiffServ in MANETs:

(a)The identification of edge routers in MANETs is ambiguous (see Fig. 2). Intuitively, the source nodes play the role of edge routers. Other nodes along the forwarding paths from sources to destinations are core nodes. But every node should have the functionality as both boundary router and interior router because the source nodes cannot be predefined.

![Fig. 2: Ambiguous edge/core routers in DiffServ. In scenario (a) nodes 2, 3, 6 act as core nodes and 1, 7 as edge routers; In scenario (b) only node 3 acts as a core node, and 2, 6 act as edge routers.](image)
Flexible QoS Model for MANET (FQMM)

Flexible QoS Model for MANET (FQMM) is the first QoS Model proposed for MANETs in 2000 by [20]. The idea of the paper is to combine knowledge from the solutions offered in the wired networks and apply them to a new QoS model, which will take into consideration the characteristics of MANETs.

The basic idea of this model is that it uses both the per-flow state property of IntServ and the service differentiation of DiffServ. In other words, this model proposes to assign highest priority to per-flow provisioning and other priority classes are given per-class provisioning. This model is based on the assumption that not all packets in our network are actually seeking for highest priority because then this model would result in a similar model with IntServ where we have per-flow provisioning for all packets.

As in DiffServ, three kinds of nodes (ingress, core and egress nodes) are defined in FQMM. An ingress node is a mobile node that sends data. Core nodes are the nodes that forward data for other nodes. An egress node is a destination node (Fig. 3). The difference though is that in FQMM the type of a node has nothing to do with its physical location in the network, since this would not make any sense in a dynamic network topology. A traffic conditioner is placed at the ingress nodes where the traffic originates. It is responsible for re-marking the traffic streams, discarding or shaping packets according to the traffic profile, which describes the temporal properties of a traffic stream such as rate and burst size.

FQMM is the first attempt at proposing a QoS model for MANETs. However, some problems still need be solved. First, how many sessions could be served by per-flow granularity? Without an explicit control on the number of services with per-flow granularity, the scalability problem still exists. Second, just as in DiffServ, the core nodes forward packets according to a certain PHB that is labeled in the DS field. We argue that it is difficult to code the PHB in the DS field if the PHB includes per-flow granularity, considering the DS field is at most 8 bits without extension. Finally, making a dynamically negotiated traffic profile is very difficult.

Very recently, the FQMM model has been extended to Relative Bandwidth Service Differentiation (RBSD) scheme in order to realize relative service differentiation in MANET [20]. It is termed in [17] as the relative bandwidth service differentiation scheme, where a service profile $\gamma$ for a traffic session is defined as a relative target rate, which is in fact a fraction of the effective capacity of a link and ranges between 0 and 1. The relative target rate of a session is normalized over time according to the traffic distribution in the MANET.

In order to estimate the effective link capacity, RSBD proposes two methods - parameter-based and measurement-based. Like its predecessor (i.e. the FQMM), the present model also suffers from the flaws like how many sessions could be served by per-flow granularity and how to distinguish between core and border nodes.

Service Differentiation in Wireless Ad Hoc Network (SWAN)

SWAN is a stateless approach dealing with service differentiation in mobile ad hoc networks [3]. SWAN assumes the use of best-effort MAC (for example IEEE 802.11 DCF) and uses feedback-based control mechanisms to support real-time services and service differentiation in ad hoc networks. In order to ensure that the bandwidth and delay requirements of real-time UDP traffic are met, distributed rate control of TCP and UDP best-effort traffic is performed at every node. Rate control is designed to restrict best-effort traffic - thus yielding the necessary bandwidth required to support real-time traffic. In addition, SWAN uses an Additive Increase Multiplicative Decrease (AIMD) rate control mechanism to improve the performance of real-time UDP traffic. Unlike TCP that uses packet loss as a feedback to avoid network congestion, SWAN uses MAC delay as a feedback to local rate controllers.

The SWAN architecture consists of three key elements, namely admission controller, packet classifier, and rate controller as depicted in Fig. 4.

The classifier and the shaper operate between IP and MAC layers. The classifier is capable of differentiating real-time and best effort packets, forcing the shaper to process best-effort packets but not real-time packets. The goal of the shaper is to delay best-effort packets in conformance with the rate calculated by the rate controller. An admission test regarding whether to admit a new real-time session is made only at the source node, and hence intermediate nodes do not perform admission tests. For this purpose, a given source is required to probe the network between itself and its desired destination in order to measure the instantaneous end-to-end bandwidth availability. Based on this probing, the source makes the sole decision. In case of false source-based admission control or
traffic violations brought on by the re-routing of real-time sessions, Explicit Congestion notification (ECN) is used to control and regulate UDP real-time traffic.

The weak side of SWAN approach is that it can only provide weak service guarantees and, although it is claimed to be stateless, intermediate nodes may be required to remember whether the flows that traverse them are new or old in order to regulate traffic [3]. In addition, source-based admission control using probing packets is again unrealistic and ineffective in a dynamic environment such as a MANET, as conditions and network topology tend to change fairly frequently. Furthermore, bandwidth calculations in SWAN do not take best effort traffic into consideration, and hence may lead to a false estimation of the available bandwidth.

VI. QoS SIGNALING/INSIGNIA

QoS signaling is used to reserve the resources required for a QoS session. A signaling protocol should handle the resource reservation, release, call setup, call tear down and renegotiation of flows in the network. A good QoS signaling should have two distinct mechanisms. First, the QoS signaling information must be reliably carried between the routers. Second, the QoS signaling information must be correctly interpreted and the relative processing should be activated. Based on the first mechanism, the QoS signaling system can be divided into in-band signaling and out-of-band signaling. The in-band signaling refers to the fact that control information is carried along with data packets [34]; the out-of-band signaling refers to the approach that uses explicit control packets.

Both in-band and out-of-band signaling have merits and demerits in their favor. For MANET, however, out-of-band signaling is not suitable since the signaling overhead of out-of-band signaling protocol like RSVP, as used in IntServ, is heavy for the mobile hosts. The signaling control message will contend with data packets for the channel and cost a large amount of bandwidth. Further, it does not adapt to the time-varying topology because it has no mechanism to rapidly respond to the topology change in MANETs. However, RSVP can be modified to make to adapt to MANET challenges.

INSIGNIA [2, 34] is an in-band signaling system that supports QoS in MANETs. It is the first signaling protocol designed solely for MANETs. The signaling control information is carried in the IP option of every IP data packet, which is called the INSIGNIA option (Fig. 5). Like RSVP, the service granularity supported by INSIGNIA is per-flow management. Each flow-state information is established, restored, adapted and removed over an end-to-end session in response to topology change and end-to-end quality of service condition.

<table>
<thead>
<tr>
<th>Request Mode</th>
<th>Service Type</th>
<th>Payload Indicator</th>
<th>Bandwidth Indicator</th>
<th>Bandwidth Request</th>
</tr>
</thead>
<tbody>
<tr>
<td>Req/Res</td>
<td>RT/BE</td>
<td>RT/BE</td>
<td>Max/Min</td>
<td>Max</td>
</tr>
</tbody>
</table>

1 bit       1 bit       1 bit       1 bit       ---       16-bit wide ---

![Fig. 5 INSIGNIA IP option field](image)

Fig.6 shows the position and the role of INSIGNIA in wireless flow management [2] at a mobile host.

The packet-forwarding module classifies the incoming packets and forwards them to the appropriate modules (routing, INSIGNIA, local applications and packet scheduling modules) [34]. If a received IP packet includes an INSIGNIA option, the control information is forwarded to and processed by the INSIGNIA module. In the meantime, the received packet is delivered to a local application or forwarded to the packet scheduling module according to the destination address.
in the IP head. If the mobile host is the destination of the packet, the packet is processed by a local application. Otherwise, the mobile host will forward the packet to the next hop determined by the MANET routing protocol. Before the packets are sent through the MAC component, a packet-scheduling module is used to schedule the output of the flows in order to fairly allocate the resource to different flows. In INSIGNIA, a Weighted Round-Robin (WRR) discipline that takes location-dependent channel conditions into account [38] is implemented. Note that a wide variety of scheduling disciplines could be used to realize the packet scheduling. The INSIGNIA module is responsible for establishing, restoring, adapting and tearing down real-time flows. It includes fast flow reservation, restoration and adaptation algorithms that are specifically designed to deliver adaptive real-time service in MANET [34]. The flow-state information is managed in soft-state method, that is, the flow-state information is periodically refreshed by the received signaling information. Coordinating with the admission control module, INSIGNIA allocates bandwidth to the flow if the resource requirement can be satisfied. Otherwise, if the required resource is unavailable, the flow will be degraded to best-effort service.

To keep the processing simple and lightweight, INSIGNIA does not send rejection and error messages if the resource request is not satisfied. For fast responding to the changes in network topology and end-to-end quality of service conditions, INSIGNIA uses QoS reports to inform the source node of the status of the real-time flows. The destination node actively monitors the received flows and calculates QoS statistical results such as loss rate, delay, and throughput. The QoS reports are periodically sent to the source node. Through this kind of feedback information, the source node can take corresponding actions to adapt the flows to observed network conditions.

On the whole, INSIGNIA is an effective signaling protocol for MANETs. In combination with other network components (viz. routing protocol, scheduling and admission control), INSIGNIA can efficiently deliver adaptive real-time flows in MANETs. However, since the flow-state information should be kept in the mobile hosts, the scalability problem may hinder its deployment in the future.

VII. QOS MAC PROTOCOL

QoS supporting components at upper layers, such as QoS signaling and QoS routing, assume the existence of a MAC protocol, which solves the problems of medium contention, supports reliable unicast communication and provides resource reservation for real-time traffic in a distributed wireless environment. A lot of MAC protocols [32, 57, 6] have been proposed for wireless networks. Unfortunately, their design goals are usually to solve medium contention and hidden/exposed terminal problems and improve throughput. Most of them do not provide resource reservation and QoS guarantees to real-time traffic.

The first problem that a MAC protocol in wireless networks should solve is the hidden/exposed terminal problem. For convenience in later discussion, we simply describe the problem and the Request-To-Send (RTS) - Clear-To-Send (CTS) dialogue as its basic solution. As shown in Fig. 7, host A and host C cannot hear each other. When A is transmitting a packet to B, C cannot sense the transmission from A. Thus C may transmit a packet to B and cause a collision at B. This is referred as the hidden terminal problem since A is hidden from C. Similarly, when B is transmitting a packet to C, A cannot initiate a transmission to D, since this can potentially cause collisions of the control packets at both B and A, thereby disrupting both transmissions. This is called the exposed terminal problem since A is exposed to B. An RTS-CTS dialogue can be used to solve the hidden/exposed terminal problem. In Fig. 7, when C wants to send a data packet to B, it first sends a RTS message to B. When B receives the RTS, it broadcasts a CTS message to C and A. When C receives the CTS, it begins to transmit the data packet. Upon receiving the CTS, A will defer its data transmission because it knows B will receive data from C. This method avoids the possible collisions at host B and thus solves the hidden terminal (A is hidden from C) and exposed terminal (A is exposed to B) problems.

Another dialogue frequently used in MAC protocols is the PKT-ACK dialogue, which means the sender sends a data packet (PKT) to the receiver and the receiver immediately responds with an acknowledgement packet (ACK) to the sender if the data packet is correctly received. Failure to receive the ACK will prompt a retransmission after a short timeout. Besides dealing with the hidden/exposed terminal problems, a QoS MAC protocol must provide resource reservation and QoS guarantees to real-time traffic. There are some proposed protocols such as the Multiple Access Collision Avoidance with Piggyback Reservation (GAMA/PR) protocol and the newly proposed Black-Burst (BB) contention mechanism [55], which can provide QoS guarantees to real-time traffic in a distributed wireless environment. However, they are supposed to work in a wireless LAN in which every host can sense each others transmission, or in a wireless network without hidden hosts.

A MAC layer protocol for QoS support in MANET was proposed by C.R.Lin and M.Gerla in [36]. They proposed Multiple Access Collision Avoidance with Piggyback Reservation (MACA/PR) for multi-hop wireless networks. MACA/PR provides rapid and reliable transmission of non-realtime datagrams as well as guaranteed bandwidth support.
QoS routing will then select A-B-C-E over route A-D-E, although the latter has fewer hops. In this paper we have only focused on MANET QoS routing protocols. QoS routing is difficult in MANET due to several reasons. First, the overhead of QoS routing is too high for the bandwidth-limited MANETs because the mobile host should have some mechanisms to store and update link state information. We have to balance the benefit of QoS routing against the bandwidth consumption in MANETs. Second, because of the dynamic nature of MANETs, maintaining the precise link state information is very difficult. Third, the traditional meaning that the required QoS should be ensured once a feasible path is established is no longer true. The reserved resource may not be guaranteed because of the mobility-caused path breakage or power depletion of the mobile hosts. QoS routing should rapidly find a feasible new route to recover the service. Addressing the above mentioned difficulties, there exist a number of QoS routing protocols for MANET. There also exist different classifications of these protocols based on (i) interaction with MAC layer, (ii) interaction between route discovery and QoS provisioning mechanism, and (iii) approach to route discovery. The basic principles of different routing protocols are summarized in following sections under the light of different classification schemes.

Fig. 8 QoS routing selects route A-D-E-C over route A-B-C to ensure bandwidth guarantee of 4 Mbps, disregarding the total hop count.

VIII. QOS ROUTING MECHANISM

Due to node mobility and limited wireless communication range of nodes in a multi-hop MANET, communication with other node must depend on the neighbor nodes to forward the data packet to the destination node. Hence, a routing protocol for MANET is a protocol that will execute on every node and therefore subject to the limit of the resources at each mobile node. The challenges of MANET, as discussed in Section III, also make the routing protocol more difficult to implement. Existing literature provides plenty of solutions to the routing problem of MANET. However, not all of them have the capability to assure QoS. A QoS routing protocol can be defined as: Given a source node s, a destination node d, a set of QoS constraints C and a possible optimization goal, a QoS routing algorithm finds the best feasible path s to d which satisfies C. For example, consider Fig. 8 where the numbers next to the radio links represent their respective bandwidth (e.g. megabits per second). To minimize delay and better utilize network resources, minimizing the number of intermediate hops is one of the principal objectives in determining suitable routes. However, suppose that the packet flow from A to E requires a bandwidth guarantee of 4 Mbps.

QoS Routing Protocol (MAC Layer Dependence)

Contention Free
Contended
Independent

Classification Based on MAC Layer Interaction
Routing protocols are often classified based on the reliance of routing protocols on MAC layer. Three classes of QoS routing solutions are presented in [26]:

- Routing protocols that rely on contention-free MAC protocol
- Routing protocols that rely on contention-based MAC protocol
- Routing protocols that do not require any MAC layer interaction at all.

These major classes (Fig. 9) are briefly explained in the following sections.
depends on an underlying contention-free MAC protocol, (2) it can operate with a contended MAC protocol, and (3) it is completely independent of the MAC protocol

(a) Protocols Relying on Contention-Free MAC
Routing protocols that rely on accurately quantified resource (commonly channel capacity) availability and resource reservation, and therefore require a contention-free MAC solution such as TDMA, belong to this group. Such protocols are able to provide, what we term, pseudo-hard QoS. Hard QoS guarantees can only be provided in a wired network, where there are no unpredictable channel conditions and node movements. Due to these unpredictable conditions, a MANET is not a suitable environment for providing truly hard QoS guarantees. Some examples of routing protocols that belong to this group include (i) QoS Routing in a CDMA over TDMA network [15, 37, 29], (ii) ticket-based multi-path routing (Chen and Nahrstedt 1998) [44], (iii) on-demand SIR and bandwidth-guaranteed routing [33], and Node state routing [56].

(b) Protocols Based on Contended MAC
Routing protocols that rely only on a contended MAC protocol and therefore only on the available resources or achievable performance to be statistically estimated belong to this group. Such protocols typically use these estimations to provide statistical or soft guarantees. Implicit resource reservation may still be performed, by not admitting data sessions that are likely to degrade the QoS of previously admitted ones. However, all guarantees are based on contended and unpredictable channel access or are given only with a certain probability and are therefore inherently soft. Routing protocols that belong to this group includes (i) core extraction distributed ad hoc routing [54], (ii) interference-aware QoS routing [23], (iii) cross-layer multi-constraint QoS routing [23], (iv) on-demand delay-constrained unicast routing [50], and (v) QoS greedy perimeter stateless routing for ultra wide band MANET [1].

(c) Protocols Independent of the Type of MAC
This group consists of those routing protocols that do not require any MAC layer interaction at all and are thus independent from the MAC protocol. Such protocols cannot offer any type of QoS guarantees that rely on a certain level of channel access. They typically estimate node or link states and attempt to route using those nodes and links for which more favorable conditions exist. However, the achievable level of performance is usually not quantified or is only relative, and therefore no promises can be made to applications. The aim of such protocols is typically to foster a better average QoS for all packets according to one or more metrics. This comes often at the cost of trade-offs with other aspects of performance, increased complexity, extra message overhead or limited applicability. The routing protocols of the group includes (i) QoS optimized link state routing [31, 4], (ii) link stability-based routing [49], (iii) hybrid ad hoc routing protocol [42], delay-sensitive adaptive routing protocol [52], (iv) application-aware QoS routing [14], (v) genetic algorithm-based QoS routing [5] and (vi) energy- and reliability-aware routing [39].

Classification Based on Routing Protocol : the QoS Provisioning Mechanism Interaction
Based on the interaction between the routing protocol and the QoS provisioning mechanism, QoS routing protocols can be classified into two categories (Fig. 10), (i) coupled and (ii) decoupled QoS approaches. In the case of the coupled QoS approach, the routing protocol and the QoS provisioning mechanism closely interact with each other for delivering QoS guarantees. If the routing protocol changes, it may fail to ensure QoS guarantees in coupled category. Some well-known QoS routing protocols that belong to this category include: (i) Ticket-Based QoS routing protocol (TDR) (Chen and Nahrstedt 1998),

Fig. 10 Classification of QoS routing protocols based on the interaction between routing protocol and the QoS provisioning mechanism: (1) coupled and (2) decoupled

(ii) Predictive Location-Based QoS Routing protocol (PLBQR) [51], (iii) trigger-based (on-demand) distributed QoS routing protocol [19] (iv) Bandwidth Routing (BR) protocol [37], (v) On-demand QoS routing (OQR) protocol [35], (vi) On-demand link-state multi-path QoS routing (OLMQR) protocol [16], (vii) Asynchronous QoS Routing (AQR) scheme [59] and (viii) Core Extraction Distributed Ad Hoc Routing (CEDAR) [54].

In the case of decoupled approach, the QoS provisioning mechanism does not depend on any specific routing protocol to ensure QoS guarantees. Routing protocols that belong to this group include INSIGNIA [2], SWAN [3] and proactive real-time MAC [12].

Classification Based on the Routing Information Update Mechanism Employed
Based on the routing information update mechanism employed, QoS approaches can be classified into three categories namely, (i) table-driven, (ii) on-demand and (iii) hybrid QoS approaches, shown in Fig. 11.

Fig. 11 Classification of QoS routing protocols based on the routing information update message employed.
In the table-driven approach, each node maintains routing information to every other node (or nodes located in a specific part) in the network. The routing information is usually kept in a number of different tables. These tables are periodically updated if the network topology changes. Due to the overhead of periodic route update message, this approach is seldom directly used in practice and the predictive location-based QoS routing protocol (PLBQR) [51] is an example of this kind. In the on-demand approach, no such tables are maintained at the nodes, and hence the source node has to discover the route on the fly. Therefore, on-demand routing protocols were designed to reduce the overheads in table-driven approach by maintaining information for active routes only. Some of the well-studied QoS routing protocols that belong to this class includes (i) the trigger-based (on-demand) distributed QoS routing (TDR) protocol [19] (ii) QoS version of AODV [45] (iii) On-Demand QoS routing (OQR) protocol [35] (iv) the On-Demand Link-State Multi-Path QoS routing (OLMQR) protocol [16], (v) Asynchronous QoS Routing (AQR) [59] and (vi) INORA [21]. A hybrid approach which is both table-driven and on-demand in nature may be designed to increase scalability by allowing nodes with close proximity to work together in order to form a backbone which reduces the route discovery overheads. This is mostly achieved by proactively maintaining routes to nearby nodes and determining routes to far away nodes using a route discovery strategy. Broadband Routing (BR) [37] and Core Extraction Distributed Ad hoc Routing (CEDAR) [54] are examples of this approach.

IX. EMERGING ISSUES IN MANET PERFORMANCE AND QoS

The support of QoS in MANETs combines several routing concepts with the mechanisms for QoS support, at the same time making assumptions as have been made for the QoS support in wired networks. However, the practical applicability of these approaches is limited due to several inherent reasons as stated below.

Bandwidth Reservation in MANET

Guaranteering a certain amount of bandwidth for a certain flow or service class requires that the station providing that guarantee is in control of that bandwidth. In wired network with full-duplex point-to-point links this can be easily assumed. It is also possible to agree on a determined share of bandwidth in a shared wired medium. Since a wired network is comprised of well-defined sub-networks, bandwidth guarantees for flows or service classes can be met by enforcing them in every involved sub-network.

The situation is completely different for wireless ad hoc networks consisting of devices with a single network interface. Wireless mobile ad hoc networks can be based on two different MAC technologies. With a single-channel protocol (e.g. IEEE 802.11 [30]), all stations communicate on the same channel and therefore potentially interfere with each other. With a multi-channel protocol in contrast (e.g. Bluetooth [8] or CDMA [47]), stations can communicate on several channels (piconets in Bluetooth terminology) simultaneously.

In MANETs, for a multi-channel MAC, a closed collision domain could easily be formed if any two wireless nodes are in same transmission range. To separate the transmission from different domains, a different channel is assigned for each of the domains. Since a station does not have a separate interface for each sub-network it participates in (as is the case in wired networks), as a result the devices have to switch channels regularly. Consequently, a well-defined, fixed sub-network structure is absent. On the other hand, in the single-channel case, the attempt to identify collision domains fails altogether. These domains would span entire connected components of the MANET, since any two neighbors belong to the same collision domain. The potentially interfering devices depend on the exact sender-receiver pair. In both of the above cases the bandwidth reservation mechanism requires a transmission schedule defining time slots, which take their turns periodically [24]. This need for a transmission schedule is the fundamental limitation in contrast to wired networks. As pointed out in [58] the problem of finding an optimal schedule is in fact NP-complete.

Service Differentiation

Apart from the general difficulties related to the bandwidth reservation outlined earlier, there is significant complexity involved in integrating classical QoS (based on bandwidth reservations) in MANET, which are often deployed with off-the-shelf hardware [24]. As a consequence, the IntServ and DiffServ based approaches are not ideally suited for MANETs. However, implementing SWAN [3] or FQMM [20] for resource reservation, though feasible, also has some practical limitations.

Queuing Issues

In service differentiation, the queuing techniques play the most important factor. Implementing traditional priority queuing strategy in MANET is significantly difficult. For example, as proposed in FQMM [20], a simple priority queue ensures that high-priority packets are given unconditional preference over low-priority packets. Secondly, they consider a FIFO queue which they enhance with a mechanism called random early discard with IN/OUT buffer management. Similarly, SWAN [3] also conceptually utilizes a priority queue, but limits the amount of real-time traffic in order to protect the lower-priority traffic from starvation.

Dealing with Congestion

FQMM [20] tries to limit network congestion by policing the traffic at the traffic sources. The sources are the equivalent of ingress routers in DiffServ networks. To regulate the traffic, a source node implements a token bucket that determines whether a packet is in-profile or out-out-profile.

SWAN [3] uses a strict admission control scheme for real-time packets. Real-time traffic is admitted by the source node depending on the outcome of probing the network for resources. If the probe packet passes a link on which the total amount of real-time traffic exceeds a certain threshold, the session will not be admitted. Thus real-time traffic is prevented from crowding out the bulk traffic. Further, large delays in real-time traffic are also prevented. It results in a quality degradation of high-priority flows as their volume increases. However, it is likely that multimedia applications will respond by changing their coding scheme to a higher compression to mitigate the effects of delay or lost packets.

Dealing with Excessive Delay

Certain applications have stringent delay bounds for their traffic. This means that packets arriving too late are useless. From the application's point of view, there is no difference...
between late and lost packets. This implies that it is actually useless to forward real-time packets that stay in a router for more than a threshold amount of time, because they will be discarded at the destination anyway. Dropping those packets instead has the advantage of reducing the load in the network [24].

X. CONCLUSIONS & FUTURE SCOPE

In this paper we have presented a comprehensive overview of the state-of-the-art research work on QoS support in MANETs. We have presented the issues and challenges involved in providing QoS in MANETs in terms of the research work on QoS models, QoS resource reservation signaling, QoS routing and QoS MAC, which are required to ensure high levels of QoS.

Related areas for further research include power consumption, resource availability, location management, interlayer integration of QoS services, support for heterogeneous MANETs, as well as robustness and security. Continued growth is expected in this area of research in order to develop, test and implement the essential building blocks for providing efficient and seamless communications in wireless mobile ad hoc networks.

REFERENCES


