Wireless Sensor Network, **2010, 2, 274-284**
doi:10.4236/wsn.2010.24038 Published Online April 2010 (http://www.SciRP.org/journal/wsn)

Comparative Review of QoS-Aware On-Demand Routing in Ad Hoc Wireless Networks

Ning Zhang, Alagan Anpalagan
WINCORE Lab, Department of Electrical and Computer Engineering, Ryerson University, Toronto, Canada
E-mail: alagan@ee.ryerson.ca

Received February 23, 2010; *revised March* 2, 2010; *accepted March* 5, 2010

Abstract

In this paper, a representative set of QoS models and QoS-aware on-demand routing protocols are reviewed with emphasis on their ability to support QoS in mobile ad-hoc networks (MANETs) possibly used in WSNs. In particular IntServ, DiffServ, FQMM, and SWAN QoS models are reviewed followed by different QoS-aware on-demand routings in MANETs from different perspectives such as the challenges, classifications, algorithmic aspects in QoS provisions. Tradeoff in providing support to real time (RT) and best effort (BE) traffic is highlighted. Finally, a detailed and comprehensive comparison table is provided for better understanding of QoS provision in MANETs.

Keywords: MANET, QoS, QoS Models, QoS-Aware Routing, Real Time Traffic, Best Effort Traffic, Comparison

1. Introduction

Mobile ad-hoc network (MANET) [1] is a collection of wireless mobile nodes, dynamically forming a temporary network without pre-existing network infrastructure or centralized administration. MANETs for wireless sensing and networking have certain unique characteristics that pose several difficulties in provisioning quality of service (QoS). They are: dynamic network topology, lack of precise state information, lack of central control, error-prone shared radio channels, limited resource availability and hidden terminal problems [2]. Most routing protocols for mobile ad-hoc networks, such as OLSR [3], DSDV [4]; DSR [5] and AODV [6]; and ZRP [7] are designed without explicitly considering QoS of the routes they generate. QoS-aware routing requires to find not only a route from a source to a destination, but a route that satisfies the end to end QoS requirement such as bandwidth, delay, jitter or probability of packet loss. Though QoS in MANETs has been researched extensively, it is a rapidly growing area of research interest due to the rising popularity and necessity of multimedia applications.

It is important to understand both QoS models and QoS routings together in MANETs and there are number of papers in the literature that discuss comparisons between them (see [8,9] and references therein). However, in this paper, they are reviewed with emphasis on different perspectives such as the challenges, classifications, algorithmic aspects for QoS-aware on-demand routing. The paper is organized as follows: In the next section, five QoS models are described in detail and then in Section 3, challenges in providing QoS are discussed with the summary of some related work in the literature. Section 4 covers various QoS-aware routing algorithms with their advantages and disadvantages. Finally, the paper concludes with the summary of the paper.

2. Different QoS Models

A QoS model specifies the network service architecture that enables us to offer better services than the best effort (BE) model and plays an important role in providing QoS support in MANETs. The QoS architecture should adapt to dynamic topology and time-varying links of MANETs. In the following, a representative set of QoS models namely: IntServ, DiffServ, FQMM, and SWAN are discussed due to their popularity in MANET research community. The authors in [8,9] provided a good review of the models and we provide them here for completeness in the following.

2.1. IntServ

The Integrated Service (IntServ) [10] QoS model includes four components: the classifier, the packet scheduler, the admission control routing, and the reservation setup protocol. An important concept “flow” is defined as distinguishable stream of related datagrams that results from a single user activity and requires the same QoS. It is the finest granularity of packet stream distinguishable by the IntServ. The basic idea of the IntServ model is that the flow-specific states are kept in every IntServ-enabled router. A flow-specific state should include bandwidth requirement, delay bound, and cost of the flow. IntServ architecture allows sources to communicate their QoS requirements to routers and destinations on the data path by means of a signaling protocol such as ReSerVation Protocol-RSVP [11]. IntServ proposes two service classes in addition to BE service. One is guaranteed service; the other is controlled load service. The guaranteed service is provided for applications requiring strict delay bound. The controlled load service is for applications requiring reliable and enhanced BE service.

Figure 1 shows how these components work together to provide integrated services. For each packet, an internet forwarder executes a suite-dependent classifier and then passes the packet and its class to the appropriate output driver. A classifier must be both general and efficient. The output driver implements the packet scheduler. If admission control gives the "OK" for a new request, the appropriate changes are made to the classifier and packet scheduler database to implement the desired QoS.

Because every router keeps the flow state information, the quantitative QoS provided by IntServ is for every individual flow. In the absence of state aggregation, the amount of state on each node scales in proportion to the number of concurrent reservations, which can be potentially large on high-speed links. This model also requires application support for the RSVP signaling protocol.

IntServ/RSVP model is not suitable for MANETs due to the resource limitations in MANETs. There are several factors which prohibit the use of that model over a MANETs. 1) Scalability: IntServ/RSVP based on per-flow resource reservation is not appropriate for MANETs because of the frequently changing topology and limited resources in MANETs resulting in more signaling overhead and unaffordable storage and computing process for mobile nodes. 2) Signaling: The RSVP reservation and maintenance process is a network consuming procedure. Thus, RSVP signaling packets will grapple with the data packets for resources and more specifically for bandwidth. This happens because RSVP is an out-of-band signaling protocol. 3) Router Mechanisms: IntServ imposes high requirement on routers. All routers must have the four basic components: RSVP, admission control routine, classifier, and packet scheduler. Consequently, the processing overheads of routers are high which is undesirable in power-constrained MANETs.

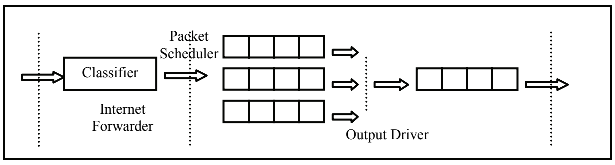

Figure 1. IntServ architecture foundations.

2.2. DiffServ

Differential Service (DiffServ/DS) [12] QoS architecture is based on a simple model where the traffic entering a network is classified and possibly conditioned at the boundaries of the network, and assigned to different behavior aggregates. Each behavior aggregate is identified by a single DS codepoint (DSCP). Within the core of the network, packets are forwarded according to the per-hop behavior (PHB) associated with the DS codepoint. The key components within a differentiated services region are traffic classification and conditioning functions. Thus, DiffServ is scalable but it does not guarantee services on end-to-end basis. This hinders DiffServ deployment in the Internet as well as in MANETs. **Figure 2** shows the architecture for DiffServ mechanism. The difference between an edge router and a core router is that the edge router is required to do the traffic conditioning as defined by a traffic conditioning agreement between their DS domain and their peer domain they are connecting to.

DiffServ on the other hand is a lightweight model for the interior routers since individual state flows are aggregated into a set of flows. This makes routing a lot more easy in the core of the network.

However, since DiffServ is originally designed for fixed wired networks, we still face some challenges to implement DiffServ in MANETs. First, it is ambiguous as to what the boundary routers in MANETs are. Intuitively, the source nodes play the role of boundary routers. Other nodes along the forwarding paths from sources to destinations are interior nodes. But every node should have the functionality as both boundary router and interior router because the source nodes can not be predefined. This drawback would again take us back to the IntServ model where several separate flow states are maintained, causing a heavy storage cost in every node. Moreover the concept of the service level agreement, defined in wire-based QoS models is not more applicable.

Three QoS classes are defined for the destination to select the best available path. The first class has the highest priority and corresponds to applications with real-time (RT) traffic such as voice. This class is for applications with high delay constraints. The corresponding service of this class in DiffServ is referred to as

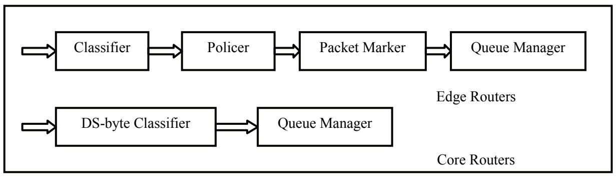

Figure 2. DiffServ architecture foundations.

"expedited forwarding" and in IntServ to guaranteed service. The second class has less priority than the first class. It is suitable for applications requiring high throughput such as video applications. This service class is referred to as "assured forwarding" in DiffServ and controlled load in IntServ. The least priority class has no specific constraint. This class is referred to as best effort in both architectures. **Table 1** shows the defined QoS classes together with their mappings to IntServ and DiffServ services. **Table 2** is a comparison of IntServ and DiffServ architecture for QoS.

2.3. FQMM

Flexible QoS Model for MANETs (FQMM) [13] is the first QoS model proposed for MANETs. It can be viewed as a hybrid of IntServ and DiffServ model. The basic idea of this model is that it uses both the per-flow state property of IntServ and the service differentiation property of DiffServ. In other words, this model proposes that highest priority is assigned per-flow provisioning and other priority classes are given per-class provisioning. This model is based on the assumption that not all packets in the network are actually seeking for highest priority. FQMM is for small to medium size MANETs using a flat non-hierarchical topology. This hybrid model defines three types of nodes, as in DiffServ: a) ingress, if it is transmitting data, b) core, if it is forwarding data and c) egress, if it is receiving data. The difference 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.

Figure 3 illustrates the FQMM architecture. A traffic conditioner is put at the ingress node where the traffic originates. It polices the traffic according to the traffic profile after a valid route is found. Components of the conditioner include traffic profile, meter, marker and dropper. For FQMM, the absolute traffic profile is not applicable since the effective bandwidth of a wireless link between nodes is time-varying. Thus, the traffic profile is defined as the relative percentage of the effective link capacity, in order to keep the differentiation between classes predictable and consistent under the dynamics of the network. In addition, the profile should be adaptive to the dynamics of the network.

Table 1. QoS classes and mappings.

Priority Class	IntServ	DiffServ
1st class e.g. voice, low delay	Guaranteed	Expedited Forwarding
2nd class e.g. video, high throughput	Controlled Load	Assured Forwarding
3rd class e.g. data, no constraint	BE	BE

Table 2. Comparison of IntServ and DiffServ.

Criteria	IntServ	DiffServ
Granularity	Individual flow	Aggregate of flows
State in routers	Per-flow	Per-aggregate
Classification	Header fields	DS field
Signaling	Required(RSVP)	Not required
Coordination	End-to-end	Per-hop
Scalability	< # of flows	< # of classes

FQMM is the first attempt at proposing a QoS model for MANETs but with the following problems: 1) without an explicit control on the number of services with per-flow granularity, the scalability problem still exists, 2) to make a dynamically negotiated traffic profile is a very difficult problem, 3) it is difficult to code the PHB in the DS field of IP, if the PHB includes per-flow granularity considering the DS field is at most 8 bits without extension. A downside of this approach is that the source stations have to take great care in regulating their traffic, since the rate of in-profile traffic must be processable in all network regions, including bottleneck areas where traffic from different sources accumulates. However, FQMM actually lacks the counterpart to DiffServ's service level agreements, and it remains an open question how the source stations should determine the dynamic parameter for their token bucket metering.

2.4. SWAN

Service Differentiation Stateless Wireless Ad-hoc Networks (SWAN) [14] is a stateless network QoS model which uses distributed control algorithms with additive increase multiplicative decrease (AIMD) rate control mechanism to deliver service differentiation in mobile wireless ad-hoc networks. The SWAN model includes a number of mechanisms used to support rate regulation of BE traffic and admission control regulation of RT traffic, as illustrated in **Figure 4**. A classifier and a shaper operate between the IP and MAC layers. The classifier is capable of differentiating RT and BE packets, forcing the shaper to process BE packets but not RT packets. The

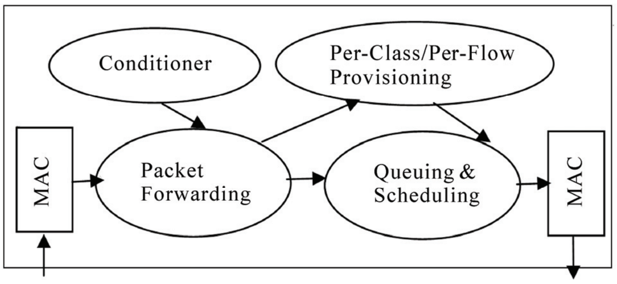

Figure 3. FQMM architecture.

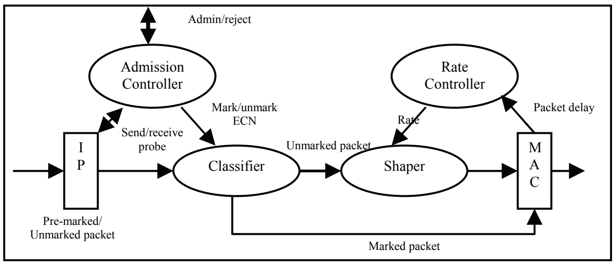

Figure 4. SWAN model adapted from [14].

shaper represents a simple leaky bucket traffic shaper. The goal of the shaper is to delay BE packets in conformance with the rate calculated by the rate controller. What makes such a stateless approach work is that all nodes independently regulate BE traffic and each source node uses admission control for RT sessions. When a new RT session is admitted, the packets associated with the admitted flow are marked as RT. The classifier looks at the marking and, if the packet is marked as RT, the packet will bypass the shaper mechanism, remaining unregulated. Here, there is an implicit assumption that a source node regulates its RT sessions based on its admission control decision.

It is unclear how the amount of bandwidth available for RT traffic should be chosen in a sensible way. Choosing larger value results in a poor performance of RT flows and starvation of BE flows, and choosing it too low results in the denial of RT flows for which the available resource would have sufficed. There would also be no flexibility to tolerate channel dynamics. The total rate of aggregated RT traffic may be dynamic due to node changes in traffic patterns and node mobility. Due to node mobility, for example, intermediate nodes may need to maintain RT traffic in excess of resources set-a-side for RT traffic. An intermediate router making this observation sets the explicit congestion notification flag in RT packets' headers. Thus, though SWAN can be a candidate QoS model, it can not be a complete QoS solution for a highly dynamic network like MANETs. In [15], a hybrid approach (incorporating DiffServ) is proposed to achieve performance gain in most of the traffic load conditions using SWAN model. A comprehensive simulation study is reported in [16] that is based on the SWAN model in MANET environment. We can conclude that SWAN tries to maintain delay and bandwidth requirements of RT traffic by admission control of UDP traffic and rate control of TCP and UDP traffic.

In the above discussed QoS models, certain routing protocols, algorithms and implementation are not specified, but the methodology and architecture to provide certain types of services were presented. There are also other architectures that could adopt a hybrid mechanism to guarantee the QoS provisions in MANETs. Since achieving QoS in MANETs not only rely on these models, all the components such as QoS-aware routing algorithms, QoS signaling and QoS MAC protocol must also work together to ensure this. In the next section, different QoS-aware routing mechanisms are presented and compared.

3. QoS-Aware Routing Mechanism and RELATED WORK

In any given network, there are two types of flows in general: BE flows which requires the data to be reliably delivered to the destination, and QoS flows (such as RT flows) which apart from reliability, requires some additional constraints such as available bandwidth, delay, etc. to be satisfied. Reusing BE routing methods for QoS-aware routing is not feasible since BE routing performs these tasks based on a single measure, usually hop-count while QoS-aware routing, however, must take into account multiple QoS measures and requirements. This section discusses different QoS-aware routings in MANETs from different perspectives including its challenges, classifications, algorithms and comparisons.

3.1. Challenges of QoS-Aware Routing

Routing in general consists of two entities, namely the routing protocol and the routing algorithm. The routing protocol has the task of capturing the state of the network and its available network resources, and disseminating this information throughout the network. The routing algorithm uses this information to compute shortest paths. Providing QoS is more difficult for MANETs due to at least two reasons. First, unlike wired networks, radios have broadcast nature. Thus, each link's bandwidth will be affected by the transmission/receiving activities of its neighboring links. Second, unlike cellular networks where only one-hop wireless communication is involved, MANETs need to guarantee QoS on a multi-hop wireless path. Further, mobile hosts may join, leave, and rejoin at any time and at any location; existing links may disappear and new links may be formed "on-the-fly". All these raise challenges to QoS-aware routing in MANETs. Next, we discuss some the challenges in designing routing protocols.

Dynamic Network Topology: A key challenge in studying protocol behavior lies in how to represent the underlying topology and traffic patterns. The constantly changing and decentralized nature of current networks results in a poor understanding of these characteristics and makes it difficult to define a "typical" configuration. For example, random graphs can result in unrealistically long paths between certain pairs of nodes, "well-known" topologies may show effects that are unique to particular configurations, and regular graphs may hide important effects of heterogeneity and non-uniformity. The performance of QoS-aware routing depends heavily on the underlying network topology. The dynamic nature of

MANETs may make the flow stop receiving QoS provisions due to path disconnections. And also new paths must be established because of the disconnections and hence will be causing data loss and delays.

Imprecise State Information: In the link-state routing algorithms, the source router selects a path based on the connection traffic parameters and the available resources in the network. The routing protocol distributes topology and load information throughout the network, and a signaling protocol for processing and forwarding connection establishment requests from the source. In MANETs, the link state changes continuously; hence, the QoS-aware routing protocols can impose a significant bandwidth and processing load on the network. Because, each router must maintain its own view of the available link resources, distribute link-state information to other routers, and compute and establish routes for new connections.

Hidden Route Problem: This problem arises at the time as the route discovery procedure of a QoS-aware routing protocol is executed. It is because the admission decision in a route discovery procedure considers only the local information, e.g., local capacity of the radio coverage of the node.

Error-Prone Shared Medium: Loss in wired networks is typically caused by excessive congestion that causes packets to be dropped at routers in the network. A wireless link, however, typically suffers much more loss due to error-prone shared medium. One cause of loss in wireless transmission is fading, in which multiple versions of the same signal are received at the destination. If these signals are out-of-phase with each other or Doppler-shifted, they can interfere with each other. Other types of interference may also cause problems in wireless transmissions including electrical noise, or possibly even intentional communication jamming. Propagation delay can also be a tremendous burden to all communication, especially to communication that requires a guarantee on total delay.

Hidden and Exposed Terminal Problem: Consider the scenario in **Figure 5**, where there are four common free time slots between *A* and *B* (1, 2, 3, 4) and four free time slots between *B* and *C* (3, 4, 5, 6), if we reserve slots (1, 2, 3) for *A* to transmit and slots (4, 5, 6) for *B* to transmit, the path bandwidth is only three which is the maximum number. Suppose there is another pair, *D* and *E*, which are currently using slot 2 to communicate. Then two cases will occur. If *D* is a receiver on slot 2, *A* will not be allowed to send on slot 2 because otherwise collision will occur at *A*. This is the hidden-terminal problem. So in the example of **Figure 5**, the common free time slots between *A* and *B* should be reduced to (1, 3, 4) and the path bandwidth from *A* to *B* has to be downgraded to 2 slots. On the contrary, if *D* is a sender on slot 2, *A* will still be allowed to send on slot 2, because this is an exposed-terminal problem. Then the common free time slots between *A* and *B* (and thus the path bandwidth) remain the same. This simple example shows the complication of the bandwidth reservation problem in QoS-aware routing in MANETs.

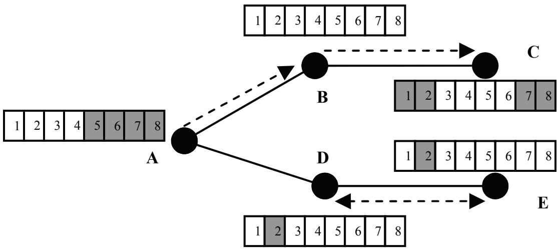

Figure 5. Hidden and exposed terminal problem in bandwidth calculation.

Lack of Central Control: Because of the lack of central controller which can account for and control MANETs' limited resources, nodes must negotiate with each other to manage the resources required for QoS routes. This is further complicated by frequent topology changes. Due to these constraints, QoS-aware routing is more demanding than BE routing.

Limited Resources Availability: In wireless networks, there are additional considerations to be taken into account. The difficulty of satisfying the QoS requirement is aggravated by further constraints on energy reserves and available bandwidth, and signal degradation by noise and limited transceiver resources. Therefore, instead of a traditional layered network control approach, a joint optimization scheme affecting both the link and the routing layer may be necessary.

QoS Signaling Support: INSIGNIA [17] provides in-band signaling support for QoS in MANETs and it is more suitable than explicit out-of-band approaches for supporting end-to-end QoS in highly dynamic environments where network topology, node connectivity and end-to-end QoS are highly time-varying. In [18], a hybrid QoS model for MANETs, called HQMM, which combines the per-flow granularity of INSIGNIA [17] and the per-class granularity of Diffserv [12], is proposed to provide scalable QoS support for MANETs. In [9], pros and cons of using cross-layer design approach to QoS provision in MANETs are discussed.

3.2. Related Work on QoS-Aware Routing

The routing protocols for MANETs may be broadly classified as table driven protocols [3,4] and on-demand driven protocols [5,6]. Table driven protocols need to maintain the global routing information about the network in every mobile node for all the possible source-destination connection and acquire to exchange routing information periodically. This kind of protocol has the property of lower latency and higher overhead.

On-demand routing protocol creates routes only when the source nodes request. When a node requires a route to a destination, it initiates a route discovery process within the network. On-demand routing protocols are characterized as having higher latency and lower overhead. A majority of existing research on the QoS-aware routing in MANETs is based on two kinds of route protocols. However, the table-driven QoS protocols request globe network state information which is not good for scalability and on-demand QoS protocols need initiates a route discovery based on flooding, which are not fit the dynamic and capability constrain in MANETs. In [19], a load-balanced AODV (LB-AODV) is proposed to control the overhead of on-demand routing in MANETs.

QoS-aware routing has received much attention recently for providing QoS in wireless ad-hoc networks and some work has been carried out to address this critical issue. Here, we provide a brief review of existing work addressing the QoS-aware routing issues in wireless ad-hoc networks. There have already been several surveys and overviews regarding the QoS-aware routing issues and solutions. Authors in [2] summarized the important QoS-related issues in MANETs in 2001, and the issues that required further attention. They updated and expanded their article in 2004 [20]. A fairly comprehensive overview of the QoS in networking could be found in [21-23]. The main conclusions from the literature are highlighted below:

1) Many of the underlying algorithmic problems, such as multi-constraint routing, have been shown to be NP-complete [20].

2) QoS and BE, routing can only be successfully achieved if the network is combinatorially stable. The dynamic topology, the error-prone channel, the lack of central control and the insecure medium have always been roadblocks for the development of QoS-aware routings [22].

3) Different techniques are required for QoS provisioning when the network size becomes very large, since QoS state updates would take a relatively long time to propagate to distant nodes [23].

4) The amount of state propagation and topology update information must be kept to a minimum. In particular, every change in available bandwidth should not result in updated state propagation [20].

5) QoS-aware routing protocol is designed without considering the situation when multiple QoS routes are being setup simultaneously. If two QoS routes cannot be fully established because they are blocking each other, both will be deleted. Hence how to setup QoS routes when there are multiple competing requests needs further study [24].

6) The protocols should be designed to accommodate multiple classes of traffic, in particular, to ensure that lower-class traffic is not starved of network resources in the presence of RT traffic [23].

4. QoS-Aware On-Demand Routing Protocols

The problem that concerned the QoS-aware routing protocol designers was that of discovering the paths that satisfy the different QoS requirements such as throughput, delay and jitter in the networks. To find a QoS route in a MANET is to establish a path that satisfies the QoS requirement given the knowledge of the available channel information at each forwarding node. In this section, some of the main QoS-aware on-demand routing protocols in MANETs are presented and the merits and deficiencies of each protocol are discussed.

4.1. Bandwidth Constraint QoS-Aware Routing

Like DSR [5] and AODV [6], the on-demand QoS-aware routing protocol [25-27] conforms to a pure on-demand rule. It neither maintains any routing table nor exchange routing information periodically. When a source node wants to communicate with another node for which it has no routing information, it floods a route request (RREQ) packet to its neighbors. A bandwidth routing protocol usually consists of three components: an end-to-end path bandwidth calculation algorithm to inform the source node of the available bandwidth to any destination; a bandwidth reservation algorithm to reserve sufficient number of free slots for the QoS flow; and a standby routing algorithm to re-establish the QoS flow in case of path breaks.

In this protocol, all packets contain following uniform fields: <*packet_type, source_addr, dest_addr, sequence#, route_list, slot_array_list, data, TTL*>. For a source node, in order to send a stream of packets to a destination node, a virtual connection (VC), to that node has to be established. The VC establishment process includes route discovery, path bandwidth calculation and bandwidth reservation components. When a node receives a RREQ packet in the route discovery process, it records the status of available slots in the *slot_arrary_list*. When the destination node receives one RREQ packet, it returns a RREP packet by unicasting back to the source following the route recorded in the *route_list*. The destination node selects the path with least cost among them and copies the fields {*route_list, slot_array_list*} from the corresponding RREQ packet to the QoS route reply (RREP) packet and sends the RREP packet to the source along the path recorded in *route-list*. As the RREP traverses back to the source, each node recorded in *route_list* reserves the free slots that have been recorded in the *slot_array_list* field. Finally, when the source receives the RREP, the end-to-end bandwidth reservation process gets completed successfully and starts sending data packets in the data phase. The reservations made are soft state in nature in order to avoid resource lock-up.

The disadvantages of these protocols are: 1) when the RREP travels back to the source, the reservation operation may not be successful. This may result from the fact that the slots which we want to reserve are occupied a little earlier by another VC or the route breaks. If this is the case, the route has to be given up and the destination re-starts the reservation process again along the next feasible route which incurs longer delay; 2) once a VC is established, the source can begin sending datagrams in the data phase. At the end of the session, all reserved slots must be released. These free slots will be contended by all new connections. However, if the last packet is lost, we will not know when the reserved slots should be released; 3) the QoS path discovered in this process may satisfy the QoS provisions but not necessarily the shortest path.

4.2. Delay Constraint QoS-Aware Routing

The On-Demand Delay-Constrained Unicast Routing Protocol (ODRP) is proposed in [28]. For ODRP to work correctly, each node is required to maintain a distance vector consisting of $|V|–1$ entries where $|V|$ is the number of nodes in the network. The entry for node v at node u ($u! = v$) contains the following information: the identifier of node v, the shortest distance from u to v (in hop count), and the next hop of u along this path to v. This vector can be provided by running a proactive wide-area distance vector routing protocol in the network.

The process of discovering a QoS-aware routing includes two phases: 1) Probing the feasibility of min-hop routing. The source sends a packet along the min-hop routing to the destination and starts a timer. If the min-hop routing satisfied the delay requirement, this delay constraint routing has been identified; 2) Destination initiated route discovery for delay-constraint path. If the minimum hop path does not satisfy the delay constraint, the destination initiates a directed and limited flood search by broadcasting a RREQ packet. Intermediate nodes only forward the RREQ with the least delay value and ignore any further RREQs. When a copy of the RREQ reaches the source with a path that meets the delay constraint, the route discovery process is complete.

The advantages of this routing protocol are: 1) the path discovery restricted flooding only when the min-hop routing does not satisfy the QoS requirements, which helps to reduce the communication overhead; 2) the route searching process is restricted and limited in a predetermined searching range and each node only forwards RREQ packet once which further limits the communication overhead. The disadvantages are: 1) the restricted searching process may lower the probability of finding a feasible path; 2) the on-demand nature of route discovery process leads to higher connection setup time; 3) while the aim of the directed flooding is to avoid global flooding, thereby reducing overhead compared to protocols that are based on that, extra overhead is incurred by the proactive distance-vector protocol which maintains the routing tables.

4.3. Location Based QoS-Aware Routing

In [29], a predictive location-based QoS-aware routing protocol is proposed. This protocol includes three components: update protocol, predictions (location prediction and delay prediction) and QoS-aware routing.

The update protocol includes two types of updates. Type 1 update is generated periodically at a constant frequency or can vary linearly between a maximum *f(max)* and minimum *f(min)* threshold with the velocity of the node. Consequently, the distance traveled between successive Type 1 updates remains constant. Type 2 update is generated when there is a considerable change in the node's velocity or direction of motion. In establishing a connection to a particular destination *B*, source *A* has to first predict the geographic location of the destination *B* as well as the intermediate hops, at the instant when the first packet will reach the respective nodes. Hence, this step involves a location as well as propagation delay prediction. The location prediction is used to determine the geographical location of some node (either an intermediate node or the destination *B*) at a particular instant of time *t* in the future when the packet reaches it.

As a result of the location-resource updates, each node has information about the entire topology of the network. It can thus compute a source route from itself to any other node using the information it has, and can include this source route in the packet to be routed. The QoS requirements are in the form of a tuple <*estimated duration of connection, maximum delay, maximum delay jitter*>. The maximum delay QoS requirement can be mapped onto the end-to-end delays observed for the updates from *B* to *A*. Thus, given the resource availability at the nodes and the QoS requirements of the connection, admission control can be performed. To search for a QoS path from *A* to *B*, *A* first runs a location-delay prediction on each node in its proximity list and obtains a list of its neighbors at the current time. It determines which of these neighbors have the resources to satisfy the QoS requirements of the connection. The next step at *A*'s network level is to perform a depth-first search for the destination starting at each of these candidate neighbors to find all candidate routes. From the resulting candidate routes, the geographically shortest one is chosen and the connection is established.

Some of the disadvantages of this protocol include: 1) it relies on accurate location awareness, which limits its usefulness to devices that are capable of being equipped with GPS receivers or such; 2) the update protocol in this paper involves flooding of location and resource information pertaining to a node to all the other nodes in the network. Ordinarily, such a full flooding of the network involves a very large overhead. However, with schemes such as the multipoint relay scheme, the overhead associated with flooding can be considerably reduced.

4.4. Hierarchical Routing: CEDAR

CEDAR [30], a core-extraction distributed ad-hoc routing algorithm for QoS-aware routing in ad-hoc network environments, has three key components: 1) the establishment and maintenance of a self organizing routing infrastructure called the core for performing route computations; 2) the propagation of the link state of high bandwidth and stable links in the core through increase/decrease waves; and 3) a QoS-route computation algorithm that is executed at the core nodes using only locally available state.

Core Extraction: The core structure is used to limit the number of nodes that must participate in the exchange of topology and available bandwidth information. The goal of setting up the core is to proactively create a core set such that every node is either a core node or a neighbor of a core node. As the route computation is done by the core nodes, minimizing the number of core nodes is desirable. Since core computation is local, it makes core computation in CEDAR scalable as the core can be computed in a constant amount of time. When a node is electing a dominator, it gives preference to core nodes already present in its neighborhood (including itself). This provides stability to the core computation algorithm, though it might have implications on the optimality of the number of core nodes. Each core node maintains local topology information and performs route discovery, route maintenance and call admission on behalf of these nodes.

Core Broadcast: In order to achieve efficient core broadcast, each node temporarily caches every request to send (RTS) and clear to send (CTS) packet that it hears on the channel for core broadcast packets only. The purpose of caching RTS/CTS is to use them for the elimination of duplicate packet reception for broadcasts.

QoS State Propagation: To propagate state information (available bandwidth) among the core nodes, increase waves and decrease waves are used. These waves are generated when a core node's available bandwidth has changed by a threshold value. A slow-moving increase wave denotes an increase of bandwidth on a link, and a fast-moving decrease wave denotes a decrease of bandwidth on a link. For low-bandwidth links, it makes sense to have as few nodes as possible contending for the link, while for stable high bandwidth links, it makes sense to have as many core nodes as possible know about the link in order to compute good routes. In other words, the maximum distance that the link state can travel (*i.e.* the time-to-live field) is an increasing function of the available bandwidth of the link. And because every core node that caches information corresponding to a link can potentially use the bandwidth of the link, the number of core nodes that cache the state of a low bandwidth link should be less compared to a stable high bandwidth link to reduce the contention for a low bandwidth link.

QoS-Aware Routing Setup: Briefly, QoS route computation in CEDAR is an on-demand routing algorithm which proceeds as follows: when a source node s seeks to establish a connection to a destination node d, provides its dominator node $dom(s)$ with a (s, d, b) tuple, where b is the required bandwidth for the connection. If $dom(s)$ can compute an admissible available route to using its local state, it responds to immediately. Otherwise, if $dom(s)$ already has the dominator of d cached and has a core path established to $dom(d)$, it proceeds with the QoS route establishment phase. If $dom(s)$ does not know the location of d, it first discovers $dom(d)$, simultaneously establishes a core path to d , and then initiates the route computation phase. A core path from s to d results in a path in the core graph from $dom(s)$ to $dom(d)$; $dom(s)$ then tries to find the shortest-widest-furthest admissible path along the core path. Based on its local information, $dom(s)$ picks up the farthest reachable domain until that which it knows is an admissible path. It then computes the shortest-widest path to that domain, once again based on local information.

The advantages of this routing protocol includes: route computation does not involve the maintenance of global state and only a few nodes are involved in state propagation and route computation. If the topology stabilizes, then routes will converge to the optimal routes. Disadvantages include: As far as the nature of state maintained at each core node is concerned, at one extreme is the minimalist approach of only storing local topology information at each core node. This approach may result in a poor routing algorithm (*i.e.*, the routing algorithm may fail to compute an admissible route even if such routes exist in the ad-hoc network). At the other extreme is the maxima list approach of storing the entire link state of the ad-hoc network at each core node. This approach computes optimal routes for stable networks, but incurs a high state management overhead for dynamic networks and potentially computes stale routes based on an out-of-date cached state when the network dynamics are high.

4.5. Application-Aware QoS-Aware Routing

A unique approach to QoS-aware routing is presented in [31]. It is unique because instead of using lower layer information, it is based on the aid of the transport layer. The protocol assumes the use of RT transport protocol (RTP) and the RT streams are delivered in the RTP packets. The delay between two nodes is estimated statistically by examining the difference between timestamps on transmission and receipt of RTP packets between those two nodes. The delay variance is also calculated. Furthermore, each node records the throughput requirement of RTP sessions which are flowing through it. Subtracting the total of these throughput values from the raw channel capacity gives an estimate for the total remaining capacity at that node.

Figure 6 shows a MANET including eight mobile nodes. The dashed line between two nodes represents the wireless connection. The number tag of each dashed line denotes the estimated transmission delay in the unit of 10 ms.

The route discovery is performed in the following steps:

Step 1: Using Floyd's algorithm, define a set I= $\{ R_1, R_2, ..., R_k \}$ including all the routes with the short- est delays satisfying the delay requirements.

Step 2: Select the subset A whose elements satisfy the bandwidth requirement. If set A is null, then go to Step 4.

Step 3: From set A, select the route R with the minimum variance of the transmission delays during a predefined period.

Step 4: Select the route R with the maximum allocated available bandwidth. If there is sufficient available bandwidth for a multimedia application, the most robust QoS route is selected using this scheme. If there are no routes that meet the bandwidth requirement, the route with the highest available channel capacity, which satisfies the delay constraint, is selected.

Using Floyd's algorithm, we could compute four routes that satisfy the delay requirement from Vs to Vt: 1, Vs→V_2→Vt; 2, Vs→V_1→V_2→Vt; 3, Vs→V_2→V_4→ Vt and 4, Vs→V_3→V_2→Vt. From Step 2, we could eliminate those routes that do not satisfy the bandwidth requirement. We assume that route 1 and 2 can satisfy the bandwidth requirement. Then from Step 3, we could choose the route that has the minimum delay variance as the QoS route. If none of the routes satisfy the bandwidth requirement, the route with the maximum available bandwidth will be selected.

A major advantage of this routing protocol is that no extra overhead is incurred for QoS-aware routing, since the existing transport layer packets are used for QoS

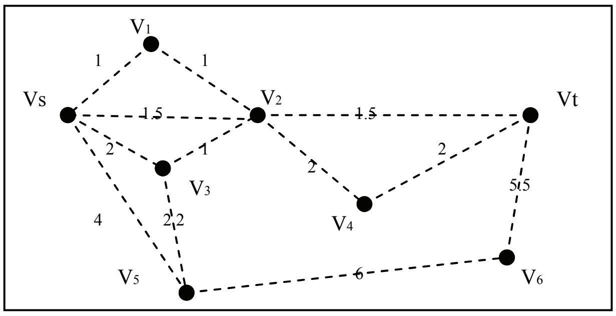

Figure 6. Network topology of application-aware routing adapted from [31].

Table 3. Comparison of QoS-aware on-demand routing protocols.

Routing Algorithms	QoS Metrics	Architecture & Reactive	Network/Node Information	MAC Layer	Other Assumptions
Bandwidth guaranteed Routing (CCBR) [23]	Bandwidth	Flat/Proactive	Time slot schedule Neighbor nodes status	CDMA over TDMA; resource reservation	DSDV routing, call admission control
On-demand QoS-aware routing [32]	Bandwidth	Flat/Reactive	Node states Neighbor nodes status	CDMA over TDMA; resource reservation	AODV routing
Delay constrained routing (ODRP) [25]	Bounded delay	Flat/Reactive	Distance vector consisting of |V|-1 entries (identifier of V, shortest path, next hop)	Resource reservation	AODV routing but proactive state dissemination
Bandwidth Reservation Multi-path QoS-aware routing [27]	Bandwidth	Flat/Reactive	Link state information	Resource reservation	AODV routing
Predictive location-based QoS-aware routing [29]	Improved link and path longevity	Flat/Reactive	Node relative positions and velocities	None	Relative location awareness; relative speed awareness; source-routing
CEDAR (Core Extraction Distributed Ad-hoc Routing) [26]	Bandwidth	Hierarchical/ Partially	Link residual capacity	Link residual capacity estimation	RTS/CTS is cached for the purpose of core broadcasting
Application-aware QoS-aware routing [31]	Bounded delay	Flat/Reactive	RTCP information	None	RTP is needed

metric estimation. Additionally, both delay and throughput constraints may be considered. However, the use of RTP is assumed, and therefore the range of application scenarios for this protocol is obviously limited.

Comparison of QoS-aware Routing Protocols

There are different ways to classify the QoS-aware routing protocols in MANETs. Some classify the protocols by the network topology (flat, hierarchical, hybrid). Some classify the protocols by different approaches to solve the QoS issues (ticket-based probing, predictive, more node state information). Some classify the protocols by route discovery approach (proactive, reactive, hybrid). Other typical classifications include by the interaction with MAC layer (independent or dependent), and also by the QoS requirements (delay, bandwidth, security, energy). In this paper, the classification of QoS-aware routing protocols is based on the approaches to QoS-aware routing in MANETs. **Table 3** lists the representative QoS-aware routing mechanisms discussed in this paper. It includes the QoS metrics, the node information, the requirement from MAC layer and other assumptions to make the protocols feasible.

5. Conclusions

In this paper, we presented a representative set of QoS-models and QoS-routings for MANETs with an emphasis on QoS-aware on-demand routing and their support for QoS provision. Although most of the research focus on different problems, they are related to each other and have to deal with some common difficulties, which include mobility, limited bandwidth and power consumption, and broadcast characteristic of radio transmission in MANETs. A detailed and comprehensive comparison table is also provided for better understanding of QoS provision in MANETs through on-demand routing mechanisms. Cross-layer approach to QoS provision in MANETs has to be carefully researched to exploit the layer interactions effectively. Though layers with strict and well defined boundaries lend themselves to modular design/testing and interoperability, they could pose some challenges in the overall system implementation [33,34].

6. References

[1] S. Corson and J. Macker, "Mobile Ad-hoc Networking (MANET): Routing Protocol Performance Issues and Evaluation Considerations," *Network Working Group Request for Comments*, No. 2501, 1999.

[2] S. Chakrabarti and A. Mishra, "QoS Issues in Ad-Hoc Wireless Networks," *IEEE Communications Magazine*, Vol. 39, No. 2, February 2001, pp. 142-148.

[3] P. Jacquet, *et al.*, "Optimized Link State Routing Protocol for Ad-Hoc Networks," *IEEE INMIC Multi Topic Conference*, Lahore, December 2001, pp. 62-68.

[4] C. E. Perkins and P. Bhagwat, "Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers," *ACM press*, Vol. 24, No. 4, October 1994, pp. 234-244.

[5] J. Broch, *et al.*, "The Dynamic Source Routing Protocol for Mobile Ad-hoc Networks," *IETF Internet-Draft*, March 1998.

[6] C. E. Perkins, "Ad-hoc On-demand Distance Vector Routing," *IETF Internet-Draft*, November 1997.

[7] Z. J. Haas and M. R. Pearlman, "ZRP: A Hybrid Framework for Routing in Ad-Hoc Networks," *Ad-hoc Networking*, 2001, pp. 221-253.

[8] K. Wu and J. Harms, "QoS Support in Mobile Ad Hoc Networks," *Interdisciplinary Journal of Crossing Boundaries*, Vol. 1, No. 1, 2001, pp. 92-106.

[9] N. Sarma and S. Nandi, "Enhancing QoS Support in Mobile Ad Hoc Networks, Advances in Computer, Information, and System Sciences and Engineering," Springer, 2006, pp. 267-273.

[10] R. Braden, D. Clark and S. Shenker, "Integrated Services in the Internet Architecture: An Overview," *Network Working Group Request for Comments*, June 1994.

[11] L. Zhang, *et al.*, "RSVP: A New Resource ReSerVation Protocol," *IEEE Network*, Vol. 7, No. 5, September 1993, pp. 8-18.

[12] S. Blake, "An Architecture for Differentiated Services," *Network Working Group Request for Comments*, December 1998.

[13] X. Hannan, *et al.*, "A Flexible Quality of Service Model for Mobile Ad-Hoc Networks," *IEEE Vehicular Technology Conference*, Vol. 1, No. 15-18, May 2000, pp. 445-449.

[14] G. Ahn, *et al.*, "Supporting Service Differentiation for RT and Best effort Traffic in Stateless Wireless Ad-Hoc Networks (SWAN)," *IEEE Transaction on Mobile Computing*, Vol. 1, No. 3, July-September 2002, pp. 192-207.

[15] N. Sarma and S. Nandi, "QoS Support in Mobile Ad Hoc Networks," *IFIP International Conference on Wireless and Optical Communications Networks*, April 2006, pp. 1-5.

[16] N. Zhang and A. Anpalagan, "Sensitivity of SWAN QoS Model in MANETs with Proactive and Reactive Routing: A Simulation Study," *Telecommunication Systems*, 2008.

[17] S-B. Lee and A. T. Campbell, "INSIGNIA: In-band Signaling Support for QoS in Mobile Ad-Hoc Networks," *International Workshop on Mobile Multimedia Communication*, 1998.

[18] Y. He and H. Abdel-Wahab, "HQMM: A Hybrid QoS Model for Mobile Ad-Hoc Networks," *IEEE Symposium on Computers and Communications*, June 2006, pp. 194-200.

[19] J. H. Song, V. W. S. Wong, V. C. M. Leung, "Efficient On-demand Routing for Mobile Ad-Hoc Wireless Access Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 22, No. 7, September 2004, pp. 1374-1383.

[20] S. Chakrabarti and A. Mishra, "Quality of Service Challenges for Wireless Mobile Ad-Hoc Networks," *Wiley J. Wireless Communications and Mobile Computing*, Vol. 4, March 2004, pp. 129-153.

[21] Z. Chenxi and M. S. Corson, "QoS-Aware Routing for Mobile Ad-Hoc Networks," *IEEE International Conference on Computer Communications*, Vol. 2, pp. 958-967.

[22] R. Renesse, *et al.*, "QoS Conflict Resolution in Ad-Hoc Networks," *IEEE International Conference on Communications*, Vol. 8, June 2006, pp. 3826-3831.

[23] T. B. Reddy, *et al.*, "Quality of Service Provisioning in Ad-Hoc Wireless Networks: A Survey of Issues and Solutions," *Ad-Hoc Networks*, Vol. 4, No. 1, January 2006, pp. 83-124.

[24] C. R. Lin, "QoS-aware Routing in Ad-Hoc Wireless Networks," 23*rd Annual Conference on Local Computer Networks*, October 1998, pp. 31-40.

[25] C. R. Lin and J. Liu, "QoS-Aware Routing in Ad-Hoc Wireless Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 17, No. 8, August 1999, pp. 1426-1438.

[26] C. R. Lin, "On-demand QoS-aware Routing in Multihop Mobile Networks," *IEEE International Conference on Computer Communication*, Vol. 3, April 2001, pp. 1735-1744.

[27] W.-H. Liao, Y.-C. Tseng, S.-L. Wang and J.-P. Sheu, "A Multi-path QoS-Aware Routing Protocol in a Wireless Mobile Ad-Hoc Network," 1*st International Conference on Networking-Part* 2, London, Vol. 2094, 2001, pp. 158-167.

[28] B. Zhang and H. T. Mouftah, "QoS-Aware routing for Wireless Ad-Hoc Networks: Problems, Algorithms and Protocols," *IEEE Communications Magazine*, Vol. 43, No. 10, October 2005, pp. 110-117.

[29] S. H. Shah and K. Nahrstedt, "Predictive Location-Based QoS-Aware routing in Mobile Ad-Hoc Networks," *IEEE International Conference on Communications*, New York, Vol. 2, May 2002, pp. 1022-1027.

[30] P. Sinha, R. Sivakumar and V. Bharghavan, "CEDAR: A Core-Extraction Distributed Ad-Hoc Routing Algorithm," *IEEE Journal of Selected Areas of Communications*, Vol. 17, No. 8, August 1999, pp. 1454-1465.

[31] W. Min and K. Geng-Sheng, "An Application-Aware QoS-Aware routing Scheme with Improved Stability for Multimedia Applications in Mobile Ad-Hoc Networks," *IEEE Vehicular Technology Conference*, Stockholm, Vol. 3, No. 25-28, September 2005, pp. 1901-1905.

[32] S. Chen and K. Nahrstedt, "Distributed Quality-of-service Routing In Ad-Hoc Networks," *IEEE Journal of Selected Areas of Communications*, Vol. 17, No. 8, August 1999, pp. 1488-1505.

[33] H. Tian, *et al.*, "CLA-QOS: A Cross-Layer QoS Provisioning Approach for Mobile Ad-Hoc Networks," *IEEE TENCON*, November 2005, pp. 1-6.

[34] Q. Liu, *et al.*, "A Cross-Layer Scheduling Algorithm with QoS Support in Wireless Networks," *IEEE Transactions on Vehicular Technology*, Vol. 55, No. 3, May 2006, pp. 839-847.