A QoS-oriented Distributed Routing Protocol for Hybrid Networks

Ze Li and Haiying Shen
Dept. of Electrical and Computer Engineering
Clemson University, Clemson, SC 29631, USA
Email: {zel, shenh}@clemson.edu

***Abstract*—As wireless communication gains popularity, significant research has been devoted to supporting real-time transmission with stringent Quality of Service (QoS) requirements for wireless applications. At the same time, a wireless hybrid network that integrates a mobile wireless ad hoc network (MANET) and a wireless infrastructure network has been proven to be a better alternative for the next generation wireless networks. By directly adopting resource reservation-based QoS routing for MANETs, hybrids networks inherit invalid reservation and race condition problems in MANETs. How to guarantee the QoS in hybrid networks remains as an open problem. In this paper, we propose a QoS-oriented Distributed routing protocol (QOD) to enhance the QoS support capability of hybrid networks. Taking advantage of fewer transmission hops and anycast transmission features of the networks, QOD transforms the packet routing problem to a resource scheduling problem. QOD includes a QoS-guaranteed neighbor selection algorithm to meet the transmission delay requirement, a distributed packet scheduling algorithm to further reduce transmission delay, and mobility-based segment resizing algorithm that adaptively adjusts segment size according to node mobility in order to reduce transmission time. Analytical results show the QOD's properties of lower transmission delay and dynamism-resilience. Simulation results show that QOD can provide high QoS performance in terms of overhead, transmission delay, dynamism-resilience and scalability compared to a resource reservation-based mechanism.**

I. INTRODUCTION

The rapid development of wireless networks has stimulated numerous wireless applications that have been used in wide areas such as commerce, emergency services, military, education and entertainment. The number of WiFi capable mobile devices including laptops and handheld devices (e.g. smartphone and PDA) has been increasing rapidly. For example, the number of wireless Internet users has tripled world-wide in the recent three years, and the number of smartphone users has increased from 7.4 million in 2003 to 69.2 million in 2006, and will reach around 190 million in 2010 [1] and 300 million by 2013 [2]. Nowadays, people are more and more unsatisfied with only watching videos (e.g., Youtube and CNN video) on PCs at home. They also want to watch video via wireless mobile devices "on the go" for game playing, watching TV and long-distance conferencing, and so on. Therefore, video streaming applications such as Qik [3] and Flixwagon [4] on the infrastructure networks have received increasing attention recently. In these applications, a video recorded by a mobile phone is uploaded to the web, and users interact with each other in real time. The widespread use of wireless and mobile devices and the desire of "video watching on the go' are leading to a promising near future where wireless multimedia services (e.g., handheld game, on-line TV, and on-line conferences) are widely deployed. The emergence and the envisioned future of real-time and multimedia applications have stimulated the need of high Quality of Service (QoS) support in wireless and mobile networking environments [5]. The QoS support limits the end-to-end transmission delay and enhances the throughput to guarantee the seamless communication between wireless and infrastructure networks.

How to guarantee the QoS in wireless networks with high dynamism and fluctuating bandwidth still remains an open problem. QoS actually is a collection of characteristics or constraints that a connection must guarantee in order to meet the requirements of an application [6]. In infrastructure networks, QoS provision (e.g. Intserv [7], RSVP [8]) has been proposed for QoS routing, which often requires node negotiation, admission control, resource reservation, and priority scheduling of packets [9]. However, it is more difficult to guarantee QoS in MANETs due to their unique features including user mobility, channel variance errors and limited bandwidth. Thus, attempts to adapt the QoS solutions for infrastructure networks to MANETs generally do not have great success [10]. Numerous reservation-based QoS routing protocols have been proposed for MANETs [11]–[19]. Basically, by reserving resources, the protocols create routes formed by nodes and links possessing the resources required to fulfill QoS requirements. Although these protocols can increase the QoS of the MANETs to a certain extent, they suffer from reservation invalid [9] and race condition problems [9]. Reservation invalid problem means the reserved resources will be useless if the data transmission path between a source node and a destination node breaks. Race condition problem means a node reserves the same resource to two reservation requesters because of the delay of the reservation reply messages.

At the same time, hybrid wireless networks have been proven to be a better network structure for the next generation wireless networks [20]–[23], and can help to tackle the stringent end-to-end QoS requirements of different applications. Hybrid networks synergistically combine infrastructure wireless networks and MANETs to leverage each other. Specifically, infrastructure networks improve the scalability of MANETs, while MANETs automatically establish self-organizing networks, extending the coverage of the infrastructure networks. Let's take a vehicle opportunistic access network, a instance of hybrid networks, as an example. People in vehicles need to upload or download videos from remote Internet servers through base stations (i.e., access points) spreading out in a city. Since it is unlikely that the base stations can cover the whole city, the vehicles themselves can form a MANET to extend the coverage of the base stations, providing continuous network connections. However, little efforts has been devoted to specifically support QoS routing in hybrid networks. Direct adoption of the reservation-based QoS routing protocols of MANETs into hybrid networks inherits the invalid reservation and race condition problems.

In order to enhance the QoS support ability of hybrid networks, in this paper, we propose a QoS-oriented distributed routing protocol (QOD). Usually, a hybrid network has widespread base stations. For example, 97% of the area of US is covered by 3G base stations of $AT\&T$ [24]. The data transmission in hybrid networks has two features. First, a base station can be a source or a destination to any mobile node. Second, the transmission hops between a mobile node and a base station are small. The first feature allows a stream to have anycast transmission with multiple transmission paths to its destination, and the second feature enables a source node to acquire updated information of its neighbors with low overhead. Taking full advantage of the two features, QOD transforms the packet routing problem into a dynamic resource scheduling problem. Specifically, in QOD, a source node selects nearby neighbors that can provide QoS services to forward its packets to base stations in a distributed manner. The source node schedules the packet streams to neighbors based on their queuing condition, channel condition and mobility, aiming to reduce transmission time and increase network capacity. The neighbors will then forward packets to base stations, which further forward packets to the destination. In this paper, we focus on the neighbor node selection for QoS guarantee. QOD is the first work for QoS routing in hybrid networks. Specifically, we make three contributions in this paper.

- *QoS-guaranteed neighbor selection algorithm.* The algorithm selects qualified neighbor and employs deadline-driven scheduling mechanism to guarantee QoS routing.
- *Distributed packet scheduling algorithm.* After qualified neighbors are identified, this algorithm schedules packet routing. It makes the queuing of previous generated packets and the generating of new packets be conducted concurrently so that the packets of a stream can be transmitted at the same time in order to reduce the transmission delay.
- *Mobility-based segment resizing algorithm.* Although high network dynamism reduces the communication time between two nodes, it increases the meeting frequency among mobile nodes and base stations for data transmission. Taking advantage of this feature, the source node adaptively resizes its packet size for each neighbor node according to the neighbor's mobility in order to transmit more packets during meeting time while avoiding packet dropping.

The rest of this paper is organized as follows. Section II presents a review of QoS mechanisms in infrastructure networks and wireless networks. Section III details the QOD protocol, with an introduction of a network model and an emphasis on the routing method. Section IV shows the performance of the QOD protocol. Finally, Section V concludes the paper.

II. Related Work

QoS can be supported at each OSI layer from the application layer to the physical layer. We focus on the routing in the network layer, which is a critical component for QoS support in networks [9].

Existing approaches for providing guaranteed services in infrastructure networks are based on two models: Integrated Services (IntServ) [7] and Differentiated Service (DiffServ) [25]. IntServ is a stateful model that uses resource reservation for individual flow, and uses admission control [7] and a scheduler to maintain the QoS of traffic flows. The ReSerVation Protocol (RSVP) [8] is used in IntServ for resource reservation in all the routers in a path. In contrast, DiffServ is a stateless model which uses coarse-grained class-based mechanism for traffic management. It classifies QoS into a number of classes, and matches data packets to the classes according to their QoS requirements. Each traffic class is managed differently, ensuring preferential treatment for higher-priority traffic on the network. A number of queuing scheduling algorithms are proposed for DiffServ to further minimize packet droppings and bandwidth consumption [26]–[30]. Thus, the network traffic is differentiated by QoS classes rather than QoS of individual packet flow. High reservation overhead for resources over long paths makes IntServ not scalable. In contrast, DiffServ-based approaches are scalable since they don't require advance setup, reservation, and time-consuming end-to-end negotiation for each flow. However, dropping some low-priority packets to guarantee the QoS of high-priority packets wastes the resources previously consumed in carrying the dropped packets. In order to solve the problem, Stoica *et. al.* [31] proposed a Dynamic Packet Service (DPS) model to provide unicast IntServ-guaranteed service and Diffserv-like scalability.

The problem of guaranteeing QoS in MANETs is more complex than in infrastructure networks because of the highly dynamic nature of mobile nodes. A majority of QoS routing protocols are based on resource reservation [9], in which a source node sends probe messages to a destination in order to discover and reserve paths satisfying a given QoS requirement. Perkins *et al.* [17] extended Ad hoc On-Demand Distance Vector Routing protocol (AODV) [32] by adding information of the maximum delay and minimum available bandwidth of each neighbor in a node's routing table. A source node sends out message indicating the maximum time allowed for a source-destination transmission. The paths having higher delay than the indicated time will discard the message. After the destination node receives the message, it responds with the estimation of cumulative delay from the intermediate nodes to the destination. Shengming *et al.* [12] proposed to reserve the resources from the node with higher link stability in order to reduce the effects of node mobility. Each node uses a link caching scheme to record its time-out history used for link stability predication. Liao *et al.* [33] proposed an extension of Dynamic Source Routing (DSR) [34] by reserving resources based on time slots. Packets are transmitted among the neighbor nodes sharing the same time slots. Although these reservation-based routing protocols can improve the QoS performance of a packet flow to a certain extent, they may lead to invalid reservation and race condition [26] problems.

Some works consider providing multi-path routing to increase the robustness of QoS routing. Conti *et al.* [13] proposed to use nodes's local knowledge to estimate the reliability of routing paths and select reliable routes for packet forwarding. The works in [14], [15] balance traffic load among multiple routes to increase routing reliability. Shen *et al.* [16] proposed to use swarm intelligence technique to enhance routing reliability in multicast transmission. A source node fetches the lost packets from its neighbors to recover the multicast traffic. Shen and Thomas [18] proposed an unified mechanism to maximize both the QoS and security of the wireless ad hoc routing. The mechanism has three basic elements: a policy-based security framework to enhance routing security, a multilayer QoS routing to improve routing QoS and a controller to leverage the above two elements. Li *et al.* [19] proposed a centralized algorithm to optimize the QoS performance by considering cross-layer design among the physical layer, MAC layer and network layer.

In the filed of wireless sensor networks (WSNs), several protocols for QoS provision also have been proposed. RAP [35] and SPEED [36] give a high delivering priority to the packet with higher distance/delay rate to the destination. However, both methods require each sensor to know its own location. Thus, the methods are not suitable for a highly dynamic environment. Felemban *et al.* [37] and Deb *et al.* [38] proposed to improve routing reliability by multipath routing. However, the redundant transmission of the packets may lead to high power consumption. Most of these routing algorithms are proposed to increase the reliability of the QoS routing in MANETs or WSNs. As far as we know, QOD is the first work for QoS routing in hybrid networks.

TABLE I
LIST OF SYMBOLS

N	# of network nodes	T_p	Transmission time of packet p
m	# of neighbors of a node	T_a	Packet arrival interval
n_i	Node i	T_{QoS}	Delay QoS requirement
R_{n_i}	Transmission range of n_i	T_w	Queuing delay
R'_{n_i}	Interference range of n_i	d_{n_i,n_j}	Distance between n_i and n_j
C_i	Link capacity of node n_i	$\widetilde{T}_{U_s}$	Threshold of space utility
$\Re_I$	Interference region	D_p	Deadline of packet p
$\Re_T$	Transmission region	S_p	The size of packet p
$\phi_\Re$	Node density in region $\Re$	$\overline{T}$	Utility update interval
$S_\Re$	Area size of region $\Re$	U_c	Channel utility
U_s	Space utility	U_{as}	Available space utility

III. THE QOD PROTOCOL

A. Network and Service Models

We consider a hybrid wireless network with an arbitrary number of base stations spreading out over the network. N mobile nodes are wandering around in the network. Each node n_i $(1 \leq i \leq N)$ uses IEEE 802.11 interface with the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) protocol [39]. Since in a hybrid network where nodes are equipped with multi-interface and multi-channel generates much less interference than a hybrid network where nodes are equipped with a single WiFi interface, we assume that each node is equipped with one WiFi Interface in order to deal with more difficult problem. The WiFi Interface enables nodes to communicate with both base stations and mobile nodes. We use R_i and $R_i^{'}$ to denote the packet transmission range and transmission interference range of node n_i, respectively. We use $d_{i,j}$ to denote the distance between n_i and n_j. A packet transmission from n_i to n_j is successful if both conditions below are satisfied [40]: (1) $d_{i,j} \leq R_i$ and (2) any node n_k satisfying $d_{k,j} \leq R'_k$ is not transmitting packets, where $0 < k < N$ and $k \neq j$. For easy reference, Table I lists the symbols used in this paper.

We assume that queuing occurs only at the output ports of the mobile nodes [41]. A node inserts its

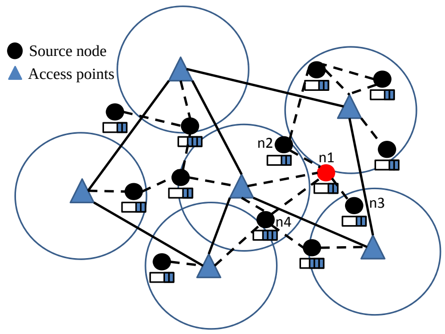


Fig. 1. The network model of the hybrid networks.

received packets into its output queue before sending them out. After a mobile node generates packets, it first tries to transmit the packets to nearby base stations that guarantee the QoS requirements. Then, it relies on its neighbors for relaying packets to base stations. The entire relaying process can be modeled as a process in which packets from a source node traverses a number of queuing servers to some base stations [42]. In this model, the problem of how to guarantee QoS routing can be transformed into the problem of how to schedule the neighbor resources between nodes to ensure QoS of packet routing.

The QoS performance requirements mainly include end-to-end delay bound, which is essential for many applications with stringent real-time requirement. While throughput guarantee is also important, it is automatically guaranteed by bounding the transmission delay for a certain amount of packets [42]. The source node conducts admission control to check whether there are enough resources to satisfy the requirements of QoS of the packet stream or not. Figure 1 shows the network model of a hybrid network. For example, when a source node n_1 wants to upload some files to an Internet server through base stations, it can choose to send packets to the base stations directly by itself or require its neighbor nodes n_2, n_3 or n_4 to assist the packet transmission.

B. An Overview of the QOD Protocol

As mentioned that when the QoS of the direct transmission between a source node and a base station can be guaranteed, the source node directly transmits packets to the base station by itself. Otherwise, the source node requests a number of qualified neighbors that can guarantee the QoS of packet transmission to assist the packet forwarding to base stations. *Scheduling feasibility* is the ability of a node to guarantee a packet to arrive at its destination within QoS requirements. The selected neighbor nodes periodically exchange the information of their status with the source node, which locally schedules the packet stream to the neighbor nodes in order to guarantee the scheduling feasibility of them. The packets are forwarded to the neighbor nodes in a round-robin fashion from a higher-delayed node to a lower-delayed node, aiming to make the queuing of previous generated packets and the generating of new packets be conducted concurrently. If there is no neighbor having the required resources for the packet delivery, the source node stops creating new connections based on the admission control policy [25].

Before introducing the details of QOD in the system, we justify that QOD is feasible to be used in a network with IEEE 802.11 protocol in Section III-C. We then present the details of QoS by answering the folllowing three questions in QoS routing in hybrid networks.

(1) **How to choose qualified neighbors for packet forwarding?** A QoS-guaranteed neighbor selection algorithm is proposed to choose neighbors in order to guarantee the routing QoS. (Section III-D)

(2) **How to schedule the packets to the qualified neighbor nodes?** A distributed packet scheduling algorithm is proposed to schedule the packets to the selected neighbors. (Section III-E)

(3) **How to guarantee the QoS transmission in a highly dynamic situation?** A mobility-based packet resizing algorithm is proposed to ensure the QoS of packet routing in a highly dynamic network. (Section III-F)

C. Applicability of QOD Distributed Scheduling Algorithm

The QOD distributed scheduling algorithm is developed based on the assumption that the neighboring nodes in the network have different channel utilities and workloads using IEEE 802.11 protocol. Otherwise, there is no need for packet scheduling in routing, since all neighbors produce comparative delay for packet forwarding. Therefore, we analyze the difference in node channel utilities and workloads in a network with IEEE 802.11 protocol in order to see whether the assumption holds true in practice.

1) Theoretical Analysis of Channel Utility and Workload Differences:

In order to avoid medium access contention and hidden terminal problem, IEEE 802.11 uses the CSMA/CA protocol as MAC access control protocol. Before a node sends out packets, it sends a Request To Send (RTS) message to the next hop node indicating the duration time of the subsequent transmission. The destination node responds with a Clear To Send (CTS) message to establish a connection with the source node. The neighbor nodes overhearing RTS and/or CTS set their Virtual Carrier Sense indicator (i.e., Network Allocation Vector (NAV)) to the transmission duration time, so that it can avoid transmitting data into the channel within the time duration. We define *channel utility* as the fraction of time a channel is busy over an unit time. Assume $\overline{T}$ is a constant time interval used for channel utility

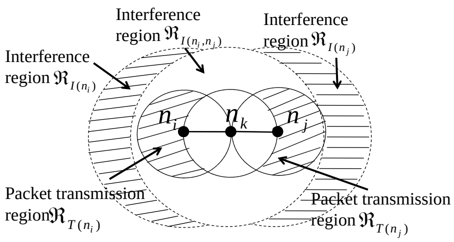


Fig. 2. Inference between two neighboring nodes.

updating, by referring to NAV and update interval $\overline{T}$, each n_i can statistically calculate its channel utility by $U_c(i) = \frac{T_{NAV}(i)}{\overline{T}}$. The available bandwidth for n_i is $(1-U_c(i))\cdot C_i$, where C_i is the transmission link capacity of node n_i.

Figure 2 shows a graph to show the inference between two neighboring nodes n_i and n_j. The solid circles around n_i and n_j denote their packet transmission ranges, and the dotted circles denote their interference ranges (sensing ranges). The nodes in shadow interference area $\Re_{I(n_i)}$ and $\Re_{I(n_j)}$ are independent from n_j and n_i respectively. By being independent, we mean when n_j is sending packets, a node in $\Re_{I(n_i)}$ can receive packets from other nodes at the same time. Therefore, the differences between the durations of generating packets of node n_i and node n_j lead to different channel utilities of the nodes in $\Re_{I(n_i)}$, $\Re_{I(n_j)}$ and $\Re_{I(n_i,n_j)}$. The workloads in n_i and n_j are determined by the packets received by n_i and n_j from the nodes in $\Re_{Tr(n_i)}$ and $\Re_{Tr(n_i)}$ respectively. We use $\phi_{\Re}$ to represent the node density in area $\Re$. Suppose node's transmission range is R, node's interference range is $R' = \alpha \cdot R$ $(\alpha > 1)$, distance between two neighboring nodes is $d = \beta \cdot R$ $(\beta < 1)$, we can get Lemma 3.1.

Lemma 3.1: Consider the inference regions of two neighboring nodes n_i and n_j, the ratio of the number of nodes with different channel utilities in $I(n_i)$, $I(n_i,n_j)$ and $I(n_j)$ is $\phi_{I(n_i)} : \eta \cdot \phi_{I(n_i,n_j)} : \phi_{I(n_j)}$, where $\eta = 2.46$.

Proof: From Figure 2, we can get

$$S_{\Re_{I(n_i)}} = S_{\Re_{I(n_j)}} = (\pi - 2\arccos(\frac{\beta}{2\alpha}))\alpha^2 R^2 + \frac{\beta R^2\sqrt{4\alpha^2-\beta^2}}{2}$$

and

$$S_{\Re_{I(n_i,n_j)}} = \pi \cdot \alpha^2 \cdot R^2 - S_{\Re_{I(n_i)}},$$

where $S_{\Re}$ denotes the size of $\Re$. Therefore, the ratio of the number of nodes in $S_{\Re_{I(n_i)}}$, $S_{\Re_{I(n_i,n_j)}}$ and $S_{\Re_{I(n_j)}}$ is $\phi_{n_i} : \eta \cdot \phi_{n_i,n_j} : \phi_{S_{\Re_{I(n_i)}}}$, where

$$\eta = \frac{S_{\Re_{I(n_i,n_j)}}}{S_{\Re_{I(n_i)}}} = \frac{\phi_{n_i,n_j} \cdot 2\arccos(\frac{\beta}{2\alpha}) \cdot \alpha^2 R^2 + \frac{\beta R^2}{2}\sqrt{4\alpha^2-\beta^2}}{((\pi - 2\arccos(\frac{\beta}{2\alpha})) \cdot \alpha^2 R^2 + \frac{\beta R^2}{2}\sqrt{4\alpha^2-\beta^2})}.$$

According to the specification of IEEE 802.11, $\alpha \approx 2$. Suppose $\beta \approx 1$, thus $\frac{S_{\Re_{I(n_i,n_j)}}}{S_{\Re_{I(n_i)}}} \approx 2.46$. ■

Theorem 3.1: The number of channel utility levels N_u in the system with N nodes are bounded by $\Theta(N) < N_u < \Theta(2^N)$.

Proof: According to Lemma 3.1, there are $3+1 = 4$ channel utility levels in a two neighboring nodes interference model as shown in Figure 2. Every time when we add a new node in the system, the number of utility levels increases at least linearly and at most exponentially. Therefore, for a system with N nodes, the number of different utilities N_u is bounded by $\Theta(N) < N_u < \Theta(2^N)$. ■

Suppose the probability of node n_i receiving a packet from nodes in $\Re_{T(n_i)}$ and n_j receiving a packet from $\Re_{T(n_j)}$ is q. The packet size is S_p. Then we can get

Theorem 3.2: The difference of the workloads in nodes n_i and n_j is

$$(\phi_{T(n_i)} - \phi_{T(n_j)})((\pi - 2\arccos(\frac{\beta}{2}))R^2 + \frac{\beta R^2}{2}\sqrt{4-\beta^2})q \cdot S_p.$$

Proof: Based on the geometric calculation, we can get

$$\Re_{T(n_i)} = \Re_{T(n_j)} = (\pi - 2\arccos(\frac{\beta}{2}))R^2 + \frac{\beta R^2}{2}\sqrt{4-\beta^2}.$$

The difference of the workload in nodes n_i and n_j is affected by the traffic from $\Re_{T(n_i)}$ and $\Re_{T(n_j)}$, which is equal to $(\phi_{T(n_i)} - \phi_{T(n_j)}) \cdot \Re_{T(n_i)} \cdot q \cdot S_p$. ■

The theoretical analysis show that if the source nodes are independent and identically distributed in a system with random packet generation rate, the nodes with IEEE 802.11 protocol can present diversified channel utilities and workload, which is suitable for distributed resource scheduling.

D. QoS-Guaranteed Neighbor Selection Algorithm

Similar to the Random Early Detection (RED) algorithm [43], in which a queue length threshold is set to avoid queuing congestion, we set up a space utility threshold $\widetilde{T}_{U_s}$ for each node as a safety line to deal with a sudden bandwidth decrease, which may make the queue scheduling infeasible. We define *space utility* as the fraction of time a node is busy with packet forwarding over an unit time. In QOD, after receiving a forward request from a source node, an intermediate node n_i with space utility less than threshold $\widetilde{T}_{U_s}$ replies the source node. The reply message contains information about the node's speed v_i, the metadata of its current workload, its available space utility $U_{as}(i)$, where $U_{as}(i) = 1 - \widetilde{T}_{U_s} - U_s(i)$. The metadata indicates the packet arriving interval and packet deadline of each flow being forwarded by the intermediate node. Based on this information, the source node chooses its neighbors for packet forwarding. A challenges here is how to select scheduling-infeasible neighbors.

Since delay is the major real-time QoS requirement for traffic transmission, we propose to use deadline-driven scheduling algorithm [44] for data traffic scheduling in intermediate nodes. This algorithm assigns the highest priority to the packet with the closest deadline, and assigns the lowest priority to the node with the furthest deadline. The packet with the highest priority is forwarded first.

Queuing deadline (deadline in short), denoted by D is defined as a time by when a packet must be sent out from an intermediate node to a base station to ensure the QoS requirement of the packet. Therefore, as Figure 3 shows, the queuing deadline is calculated as $D_p = t + T_{QoS} - T_{S\to I} - T_{I\to D}$, where T_{QoS} is the QoS delay requirement. Using W_S and W_I to denote the bandwidth of a source node and an intermediate node respectively, $T_{S\to I} = \frac{S_p}{W_S}$ denotes the transmission delay between a source node and an intermediate node, and $T_{I\to D} = \frac{S_p}{W_I}$ denotes the transmission delay between an intermediate node and a base station. t is the time when the packet is generated.

Liu *et. al* [44] proved that for a given set of $\overline{m}$ tasks for an operating system, the deadline-driven scheduling algorithm is feasible for the job scheduling iff

$$(\frac{T_{cp}(1)}{T_g(1)}) + (\frac{T_{cp}(2)}{T_g(2)}) + (\frac{T_{cp}(j)}{T_g(j)}) + ... + (\frac{T_{cp}(\overline{m})}{T_g(\overline{m})}) \leq 1, \quad (1)$$

where $T_g(j)$ denotes job arrival interval time period from task j and $T_{cp}(j)$ denotes the job computing time.

In a communication network, the packet transmission time of a packet in packet stream from node n_j is actually similar to the job computing time $T_{cp}(j)$ for a job from task j in in job scheduling model in [44]. Then, Equation (1) is equivalent to

$$(\frac{S_p(1)}{T_a(1)}) + (\frac{S_p(j)}{T_a(j)}) + ... + (\frac{S_p(m)}{T_a(m)}) \leq W_i, \quad (2)$$

where $W_i = (1 - U_c(i)) \cdot C_i$, $S_p(j)$ is the size of the packets from node n_j. Equation (2) shows that the scheduling feasibility of a queue is affected by packet size S_p, the number of packet streams from m neighbors, packet arrival interval T_a and bandwidth W_i of the intermediate node.

Based on Equation (2), given a certain packet size, the source node n_s can determine the packet generation interval of the packet $T_a(i)$ to each qualified intermediate node n_i. For example, an intermediate node n_i receives packet traffic from three different source nodes n_1, n_2 and n_3 periodically. The packets size of traffic from n_1,n_2 and n_3 are $1k$, $10k$ and $20k$ with arrival interval $0.1s$, $0.2s$ and $0.5s$, respectively. Then, $\frac{S_1}{T_1} + \frac{S_2}{T_2} + \frac{S_3}{T_3}$=70k. If the bandwidth W_i of the intermediate node n_i is larger than $70k/s$, the intermediate node can guarantee those packet flows are sent out within the given deadline. However, if the bandwidth is less than 70k/s, the QoS of the traffic cannot be guaranteed. It is called scheduling infeasible in this queue.

E. Distributed Packet Scheduling Algorithm

Section III-D solves the problem of how to select intermediate nodes that can guarantee the QoS of the packet transmission and how a source node assigns traffic to the intermediate nodes to ensure their scheduling feasibility. In order to further reduce the stream transmission time and balance the scheduling burdens on the queues of neighbors, a distributed packet scheduling algorithm is proposed for packet routing by making the queuing of previous generated packets and the generating of new packets be conducted in parallel. Such that, the packets of one stream can arrive at the destination at the same time.

As Figure 3 shows, a source node generates three packets p_1, p_2 and p_3 at times t_0, t_1 and t_2 ($t_0 < t_1 < t_2$), respectively. Since all these packets are generated from the same node, the transmission delay from the source node to each intermediate node $T_{S\to I}(1)$, $T_{S\to I}(2)$ and $T_{S\to I}(3)$ are almost the same. To make the analysis more clear, we suppose $T_{I\to D}(1) = T_{I\to D}(2) = T_{I\to D}(3)$. Let $T_w(i)$ denote the packet queuing time of n_i, if the queuing delay in each intermediate node satisfies $T_w(1) > T_w(2) > T_w(3)$, the final packet delivery time from the intermediate nodes to the destination node can be the same.

Therefore, after scheduling traffics to qualified intermediate nodes, the source node estimates the expected T_w of the packets in the queue of each intermediate node based on the received metadata from neighbors. When transmitting the packets, the earlier generated packet is transmitted to a node with longer queuing delay but still within the deadline bound. Taking advantage of the different T_w in different neighbor nodes, the transmission time of the entire traffic stream can be decreased by making the queuing of previous generated packets and the generating of new packets be conducted in parallel. Thus, a packet is waiting in one queue and the subsequential packets can concurrently be transmitted to base stations. T_w is estimated by

$$T_w = \sum_{j=1}^{x-1} T_{I\to D}(j) \cdot \lceil T_w / T_{I\to D}(j) \rceil + T_{I\to D}(x),$$

where x is denoted as a node with the ith priority in the queue. $T_{I\to D}(j)$ is denoted as the transmission delay of a node with the jth priority, where $0 < j < i$.

Theorem 3.3: Given a certain amount of packets to transmit, QOD can increase the throughput of the single shortest path transmission method, in which a source node always transmits packets through a single shortest path.

Proof: Suppose the packet generating rate of a source node n_i is λ. That is, the packet arrival interval is $T_a = \frac{1}{\lambda}$. The queuing delay time T_w in different intermediate nodes j are different, and $T_w(1) > T_w(2) >$

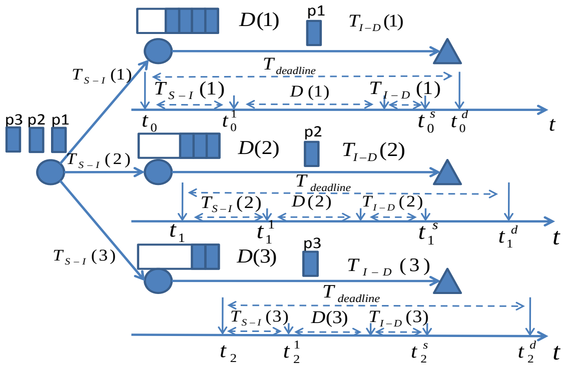


Fig. 3. Distributed queuing mechanism.

$T_w(3) > T_w(j) > ... > T_w(m)$ $(1 \le i \le m)$. According to the distributed packet scheduling algorithm, the time to transmit m packets to different base stations through m qualified intermediate nodes is $T_w^{(d)} = max(i \cdot T_a + T_w(i)) + T_{S\rightarrow I} + T_{I\rightarrow D}$ $(1 \le i \le m)$. As Equation (1) indicates, the packet queuing time is less than the packet arrival interval, i.e., $T_w(i) < T_a$. Therefore, in the single shortest path transmission method, the time for transmitting m packets is $T_w^{(s)} = m\cdot(T_a + T_w(m)) + T_{S\rightarrow I} + T_{I\rightarrow D}$. Then,

$$T_w^{(d)} - T_w^{(s)} = max(i \cdot T_a + T_w(i)) - m(T_a + T_w(m)).$$

Since $T_w(i) < T_a$, $max(i\cdot T_a + T_w(i)) = mT_a + T_w(m)$, then

$$T_w^{(d)} - T_w^{(s)} = (1 - m)T_w(m) \le 0.$$

Therefore,

$$T_w^{(d)} \le T_w^{(s)}.$$

■

F. Mobility-based Segment Resizing Algorithm

In a highly dynamic mobile wireless network, the transmission link between two nodes are frequently broken down. The delay incurred in the packet retransmission degrades the QoS of the transmission of a packet flow. On the other hand, a node in a highly dynamic network has higher probability to meet different mobile nodes and base stations, which is beneficial to the resource scheduling. Taking advantage of this benefit, a mobility-based segment resize algorithm is proposed in this section. As Equation (2) shows, the space utility of an intermediate node used for forwarding a packet p is $\frac{S_p}{W_i \cdot T_a}$. Reducing packet size can increase the scheduling feasibility of an intermediate node as well as reduces packet dropping probability. However, since more packets need to transmit when a packet is small, more overheads are created because of the packet head. Therefore, for a low mobility node, we prefer the packet with larger size. Therefore, in QOD, as the mobility of a node increases, the packet size of a node sent to its neighbor nodes i decreases as $S_p(new) = \frac{\gamma}{v_i} S_p(old)$, where γ is a scaling parameter and v_i is the comparative mobility speed of the source node and intermediate node. In order to compensate the decreased throughput of the source node, the packet generation rate T_a for the short size packet is increased by $T_a(new) = \frac{v_i}{\gamma} T_a(old)$.

Proposition 3.4: The QOD protocol can guarantee the QoS of packet routing in a highly dynamic network.

Proof: Suppose the density of the nodes in the system is ϕ, in a certain time period $\overline{T}$, a node can meet $N = \overline{T} \cdot \Theta(\phi \cdot v_i)$ nodes. Since the packet size of a packet sending from a source node to an intermediate node n_i is $\frac{\gamma}{v_i}$, and the average packet arrival interval is about $\Theta(\frac{N \cdot v_i}{\lambda})$, the space utility of a node n_i with bandwidth w_i is about $U_s = \Theta(\frac{\gamma \cdot \lambda}{\phi \cdot \overline{T} \cdot w_i \cdot v_i^2})$. As v_i increase, space utility U_s decrease. Therefore, based on QOD routing protocol, QoS of the traffic can be guaranteed in a highly dynamic situation. ■

IV. Performance Evaluation

This section demonstrates the distinguishing properties of QOD compared to E-AODV [17] through simulations built on NS-2 [45]. E-AODV is a resource reservation oriented routing protocol for QoS routing in MANETs. This protocol extends AODV by adding information of the maximum delay and minimum available bandwidth of each neighbor in a node's routing table. To apply E-AODV in hybrid networks, we let a source node searche for the QoS guaranteed path to a base station. The intermediate nodes along the path reserve the resources for the source node.

In the simulation, 30 nodes were independent and identically distributed in a 1500m $\times$ 1500m square. The transmission range of a node was set to 250m, interference range was set to 550m, link rate was set to 2M/s and packet size was set to 1024b. 6 base stations are uniformly distributed with IEEE 802.11 MAC protocol. Two source nodes are randomly selected to send packets to base stations in every ten seconds. A user's traffic is generated with CBR sources. The average generation rate of the CBR traffic is $100kb/s$. Unless otherwise specified, the speeds of the nodes are randomly selected from [1-40]m/s. Since packets arriving within a certain delay are valuable to video streaming applications, we define a new metric, namely QoS guaranteed throughput (QoS throughput in short). It is the throughput sent from a source node to a destination node satisfying a QoS delay requirement. This metric can simultaneously capture delay, throughput and jitter features of packet transmission. The collected results are the average value of 10 times of runs. The warmup time was set to 100s and the simulation time was set to 200s per round. We define QoS guaranteed throughput (QoS throughput in short) as the throughput sent from a source node to a destination node satisfying QoS requirement.

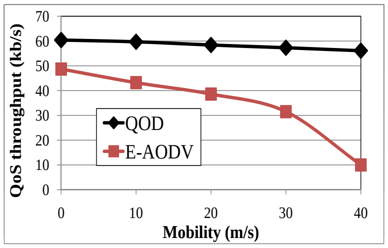


Fig. 4. QoS throughput versus mobility.

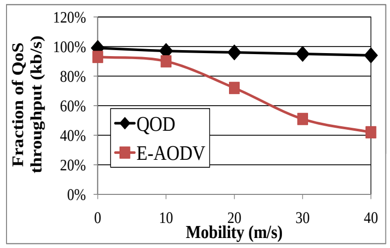


Fig. 5. Fraction of QoS throughput versus mobility.

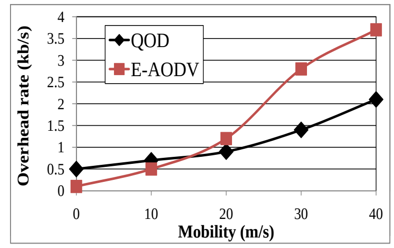


Fig. 6. Overhead versus mobility.

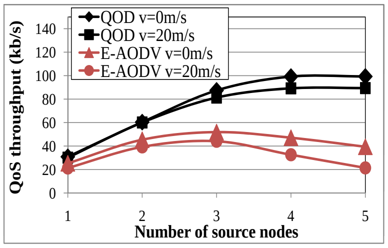
A. Performance in Different Mobility

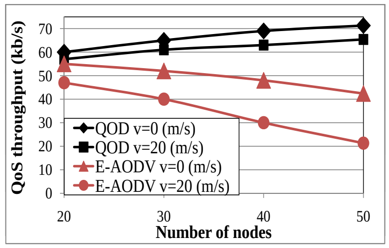
In this experiment, we varied the mobility of all nodes from 0m/s to 40m/s with 10m/s increment in each step. Figure 4 shows the QoS throughput of E-AODV and QOD versus the node mobility speed. As the mobility of the nodes increases, the QoS throughput of QOD slightly decreases, but that of E-AODV decreases sharply. It is because in E-AODV, the routing resources in each link are reserved for QoS traffic. In a highly dynamic network, the reserved links constantly break down, forcing the source node searching for a new path to a base station. The delay resulted from the path searching degrades the ability to meet the QoS requirements. Therefore, the QoS of the packet traffic in E-AODV is very difficult to be guaranteed in a highly dynamic network.

In contrast, in QOD, rather than reserving the resources in each transmission link, the intermediate nodes periodically report their queuing status to the source node. The source node adaptively schedules the packets to the neighbor nodes based on their current space utilities. In this way, there is no need for retransmission caused by invalid resource reservation. Moreover, since every reporting node can receive scheduled packets for the forwarding transmission, the race contention problem can be avoid. Furthermore, a packet resizing algorithm is used for traffic scheduling in QOD as the mobility of the nodes increases. A small packet size can reduce the probability of a packet being dropped due to link breakdown. Meanwhile, it can increase the scheduling feasibility of an intermediate node on this packet as Proposition 3.4 indicates. Because of the increased overhead in QOD in a higher dynamic network, its QoS throughput decreases slightly. We can also see that the QoS throughput of QOD is better than E-AODV even in a low dynamic scenario. It is because QOD makes the queuing of previous generated packets and the generating of new packets be conducted in parallel, which reduces the transmission time of a packet stream. Therefore, the QoS throughput of nodes in QOD is greatly improved. This experiment result is consistent with Theorem 3.3.

Figure 5 shows the high QoS-guaranteed feature of QOD than E-AODV. We define the *fraction of QoS throughput* as the ratio of QoS throughput to total packet throughput. The figure shows that when the network topology is stable, the fraction of QoS throughput of E-AODV is high. That is, most of the received packets meet their QoS requirements. Due to the same reason of Figure 5, the fraction of QOD's QoS throughput is higher than E-AODV. We can also see that as the mobility of the nodes in the system increases, the fraction of QoS throughput in QOD decreases marginally. QOD's distributed packet scheduling algorithm can avoid race contention, and the packet resizing algorithm can increase the scheduling feasibility of the intermediate nodes. The decrease fraction of the QoS throughput is due to the increasing overhead in the system. More specifically, since the high mobility of nodes leads to low packet size for each packet, given a certain amount of data, more packets are needed to transmit these data to the destination nodes. Therefore, the increasing number of packets lead to increasing extra overhead in packet head. That is why the overhead of QOD is increased in a highly dynamic scenario. In contrast, the fraction of the QoS throughput decreases sharply in E-AODV as node mobility increases. It is due to the new path discovering process that prevents the packets in E-AODV from arriving at base stations in time.

We define the overhead rate as the overhead generated by the system in an unit time. Figure 6 plots the transmission overhead rate of QOD and E-AODV as the node mobility increases. The figure shows that in a low mobility environment, QOD generates higher overhead than E-AODV. It is because the routing control overhead takes up most of the overhead in E-AODV. If the topologies of the system are comparably stable, the control overhead is small. The overhead in QOD consists of two parts. The first part is overhead for periodic queuing information exchange. In QOD, a source node exchanges its status information with its neighbor nodes periodically during the packet transmission time for the packet scheduling. Although the information exchange is reactive conducted and is only conducted when a source node has packets to be sent out, the overhead is still larger than E-AODV in a low mobility scenario. The second part of the overhead is resulted from the packet head. Although the packet size of each packet is reduced as the mobility of the nodes increases, more





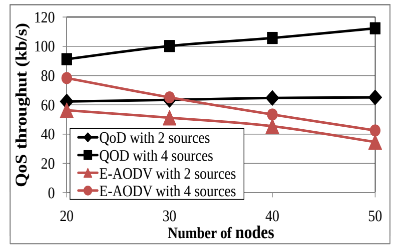


Fig. 7. QoS throughput versus workload. Fig. 8. QoS throughput versus network size. Fig. 9. QoS throughput versus network size.

packets need to generate in order to transmit the stream to the destination node. The extra packet head increase the overhead of QOD. That is why as the mobility of nodes in the system increases, the overhead of QOD also increase. However, the increasing mobility of the nodes leads to more control overhead for topology maintenance in E-AODV. Also, the overhead increasing rate is much higher than QOD. Therefore, the overhead of QOD is less than E-AODV in a highly dynamic environment.

B. Performance in Different Workload

Figure 7 plots the QoS throughput of QOD and E-AODV with different number of source nodes. The more source nodes means more workload in the system. It is very interesting to see that as the number of source nodes increase from 0 to 3, the QoS performance of QOD increases linearly. It is because the capacity of the system is not saturate at this moment. When the number of source nodes increases to 5, the QoS throughput stops increasing. In QOD protocol, when a source node finds all of its neighbor nodes cannot guarantee the QoS of its packets, it stops generating new packet flows into the system based on the admission control policy. However, in E-AODV, as the number of source nodes increases, the QoS throughput is increasing initially but decreases later. It is because in E-AODV, when the workload of the system increases, the probability that two or more QoS routings reserves the same resources at a node simultaneously increases, leading to the race condition problem. Therefore, the QoS throughput of E-AODV decreases in a highly loaded system. The figure also shows that the increasing mobility of the nodes in the system leads to a decrease of QoS throughput of both E-AODV and QOD protocol. However, the performance decrease of QOD is much less than E-AODV due to the same reason in Figure 4.

C. Performance in Different Network Size

Figure 8 illustrates the QoS throughput of QOD and E-AODV with different number of nodes in the system. The figure shows that as the number of nodes in the system increases, the QoS performance of QOD increases, but the QoS performance of E-AODV decreases. The reason is that the increasing number of nodes in the system lead to a increasing number of neighbors of a node. Thus, more resources are available for a source node to conduct packet traffic scheduling. We can see that the QoS throughput of E-AODV decreases as the number of nodes in the system increases. This is because that the average transmission hops between a source node and a base station grow, and more transmission hops produce a high probability of path breakdown in a dynamic network. That is also why as the mobility of the nodes increases, the QoS performance of E-AODV decreases sharply especially when the number of nodes in the system is large. For QOD, the increase of mobility does not affect its QoS performance significantly because of its mobility resilient feature, due to the same reason as Figure 4.

Figure 9 shows the comparison results of the QoS throughput of QOD and E-AODV with different number of source nodes and different number of all nodes in the system. The figure shows that in QOD, as the number of source nodes increases from 2 to 4, the QoS throughput increases. By scheduling the traffic rather than reserving the resource in the neighbors of the source node, QOD relieves the race contention in E-AODV. The QoS throughput of QOD increases until the network has no resource for the packet transmission. The figure also shows that in QOD, a system with larger number of nodes and source nodes has higher throughput increasing rate. The reason is that a larger number of neighbor nodes can provide more resources for the packet scheduling. However, in E-AODV, as the number of nodes in the system increases, the QoS throughput of E-AODV decreases. Moreover, as the number of source nodes in the system increases, the throughput exhibits dramatic decrease. A large number of source nodes produce a high workload in the system, resulting in a high probability of race contention occurrence. Moreover, a larger number of nodes generate more transmission hops, resulting in a high probability of link breakdown.

V. Conclusions

Hybrid wireless networks that integrate MANETs and infrastructure wireless networks have been proved to be a better network structure for the next generation networks, and have higher QoS support capability than MANETs and infrastructure wireless networks. However, little effort have been devoted to supporting QoS routing in hybrid networks. Direct adoption of the QoS routing techniques in MANETs into hybrid networks inherits their drawbacks. In this paper, we propose a QoS-based distributed routing protocol (QOD) for hybrid networks to provide QoS services in a highly dynamic scenario. Taking advantage of the unique features of hybrid networks, i.e., anycast transmission and short transmission hops, QOD transforms the packet routing problem to the packet scheduling problem. In QOD, a source node directly transmits packets to a base station if the direct transmission can guarantee the QoS of the traffic. Otherwise, the source node schedules the packets to a number of qualified neighbor nodes. Specifically, QOD incorporates three algorithm. The QoS-guaranteed neighbor selection algorithm is used to choose qualified neighbors for packet forwarding. The distributed packet scheduling algorithm is used to schedule the packet transmission to further reduce the packet transmission time. The mobility-based packet resizing algorithm is used to ensure the QoS of the traffic in a highly dynamic environment. Theoretical analysis proves the high performance of QOD. Experiment results show that QOD outperforms a resource reservation-based algorithm in terms of dynamism-resilience, scalability and contention reduction.

ACKNOWLEDGEMENTS

This research was supported in part by U.S. NSF grants CNS-1025652, CNS-1025649, and CNS-0917056, Microsoft Corporation grant 8300751, and Sandia National Laboratories grant 10002282.

REFERENCES

[1] The state of the smartphone market. http://www.allaboutsymbian.com.
[2] Next Generation Smartphones Players, Opportunities & Forecasts 2008-2013. Technical report, Juniper Research, 2009.
[3] Qik. http://qik.com/.
[4] Flixwagon. http://www.flixwagon.com/.
[5] H. Wu and X. Jia. QoS multicast routing by using multiple paths/trees in wireless ad hoc networks. *Ad hoc Networks*, 2009.
[6] E. Crawley, R. Nair, B. rajagopalan, and H. Sandick. RFC2386: A framework for QoS-based routing in the internet. *IETF*, 1998.
[7] R. Braden, D. Clark, and S. Shenker. RFC1633: Integrated services in the internet architecture: an overview. *IETF*, 1994.
[8] E. Crawley, R. Nair, B. rajagopalan, and H. Sandick. RFC2205: Resource reservation protocol RSVP. *IETF*, 1998.
[9] I. Jawhar and J. Wu. Quality of service routing in mobile ad hoc networks. *Springer, Network Theory and Applications*, 2004.
[10] T. B. Reddy, I. Karthigeyan, B. S. Manoj, and C. S. R. Murthy. Quality of service provisioning in ad hoc wireless networks: a survey of issues and solutions. *Ad hoc Networks*, 2006.
[11] X. Du. QoS routing based on multi-class nodes for mobile ad hoc networks. *Ad hoc Networks*, 2004.
[12] J. Shengming, L. Yaoda, J. yuming, and Y. Qinghe. Provisioning of adaptability to variable topologies for routing schemes in MANETs. *IEEE J-SAC*, 2004.
[13] M. Conti, E. Gregori, and G. Maselli. Reliable and efficient forwarding in ad hoc networks. *Ad hoc Networks*, 2004.
[14] G. Chakrabarti and S. Kulkarni. Load banlancing and resource reservation in mobile ad hoc networks. *Ad hoc Networks*, 2006.
[15] A. Argyriou and V. Madisetti. Using a new protocol to enhance path reliability and realize load balancing in mobile ad hoc networks. *Ad hoc Networks*, 2006.
[16] C. Shen and S. Rajagopalan. Protocol-independent multicast packet delivery improvement service for mobile ad hoc networks. *Ad Hoc Networks*, 2007.
[17] C. E. Perkin, E. M. Royer, and S. R. Quality of service in ad hoc on-demand distance vector routing. *IETF*, 2001.
[18] Z. Shen and J. P. Thomas. Security and QoS self-optimization in mobile ad hoc networks. *IEEE TMC*, 2008.
[19] Y. Li and A. Ephremides. A joint scheduling, power control, and routing algorithm for ad hoc wireless networks. *Ad Hoc Networks*, 2008.
[20] H. Luo, R. Ramjeey, P. Sinhaz, L. Liy, and S. Lu. UCAN: A unified cell and ad-hoc network architecture. In *Proc. of MOBICOM*, 2003.
[21] P. K. Mckinley, H. Xu, A. Esfahanian, and L. M. Ni. Unicast-based multicast communication in wormhole-routed direct networks. *IEEE TPDS*, 1992.
[22] H. Wu, C. Qiao, S. De, and O. Tonguz. Integrated cell and ad hoc relaying systems: iCAR. *J-SAC*, pages 2105– 2115, 2001.
[23] J. Zhou and Y. R. Yang. PAR CelS: Pervasive ad-hoc relaying for cell systems. In *Proc. of Med-Hoc-Net*, 2002.
[24] At&t coverage. http://www.att.com/truthabout3g/?WT.srch=1.
[25] Y. E. Sung, C. Lund, M. Lyn, S. Rao, and S. Sen. Modeling and understanding end-to-end class of service policies in operational networks. In *Proc. of SIGCOMM*, 2009.
[26] J. F. Kurose and K. W. Ross. Computer networking: A top-down approach featuring the internet. *Addison Wesley*, 2004.
[27] L. Kleinrock. Queueing system. *Computer Applications*, 1976.
[28] A. Parekh and R. Gallager. A generalized processor sharing approach to flow control - the single node case. In *Proc. of Infocom*, 1992.
[29] J. C. R. Bennett and H. Zhang. WF^2Q: Worst-case fair weighted fair queueing. In *Proc. of Infocom*, 1996.
[30] S. Golestani. A self-clocked fair queueing scheme for broadband applications. In *Proc. of Infocom*, 1994.
[31] I. Stoica and H. Zhang. Providing guaranteed services without per flow management. In *Proc. of SIGCOMM*, 1999.
[32] C. Perkins, E. Belding-Royer, and S. Das. RFC 3561: Ad hoc on demand distance vector (AODV) routing. *IETF*, 2003.
[33] W. H. Liao, Y. C. Tseng, and K. P. Shih. A tdma-based bandwidth reservation protocol for QoS routing in a wireless mobile ad hoc network. In *Proc. of ICC*, 2002.
[34] D. B. Johnson and D. A. Maltz. Dynamic source routing in ad hoc wireless networks. *IEEE Mobile Computing*, 1996.
[35] C. Lu, B. M. Blum, T. F. Abdelzaher, J. A. Stankovic, and T. He. Rap: A real-time communication architecture for large-scale wireless sensor networks. In *Proc. Of RTAS*, 2002.
[36] T. He, J. Stankovic, C. Lu, and T. Abdelzaher. Speed: A stateless protocol for real-time communication in sensor networks. In *Proc. Of ICDCS*, 2003.
[37] E. Felemban, C. G. Lee, and E. Ekici. Mmspeed: Multipath multi-speed protocol for qos guarantee of reliability and timeliness in wireless sensor networks. *IEEE TMC*, 2006.
[38] B. Deb, S. Bhatnagar, and B. Nath. Reinform: Reliable information forwarding using multiple paths in sensor networks. In *Proc. Of LCN*, 2003.
[39] IEEE-SA Standards Board. Wireless lan medium access control (mac) and physical layer (phy) specifications.
[40] P. Gupta and P. R. Kumar. The capacity of wireless networks. *IEEE TIT*, 2000.
[41] M. Karol, M. Hluchyj, and S. Mogan. Input versus output queueing on a space-division packet switch. *IEEE TOC*, 1987.
[42] Hui Zhang. Service disciplines for guaranteed performance service inpacket-switching networks. In *Proc. of IEEE*, 1995.
[43] D. Lin and R. Morris. Dynamics of random early detection. In *Proc. of SIGCOMM*, 1997.
[44] C. L. Liu and J. W. Layland. Scheduling algorithms for multiprogramming in a hard real-time enviroment. *JACM*, 1973.
[45] The network simulator ns-2, version 2.23. http://www.isi.edu/nsnam/ns/.