QoS Routing for Wireless Ad Hoc Networks: Problems, Algorithms, and Protocols

Baoxian Zhang and Hussein T. Mouftah, University of Ottawa

ABSTRACT

QoS routing plays an important role for providing QoS in wireless ad hoc networks. The goals of QoS routing are in general twofold: selecting routes with satisfied QoS requirement(s), and achieving global efficiency in resource utilization. In this article we first discuss some key design considerations in providing QoS routing support, and present a review of previous work addressing the issue of route selection subject to QoS constraint(s). We then devise an on-demand delay-constrained unicast routing protocol. Various strategies are employed in the protocol to reduce the communication overhead in acquiring cost-effective delay-constrained routes. Simulation results are used to verify our expectation of the high performance of the devised protocol. Finally, we discuss some possible future directions for providing efficient QoS routing support in wireless ad hoc networks.

INTRODUCTION

Quality of service (QoS) provisioning is becoming a critical issue in designing wireless ad hoc networks due to the necessity of providing multimedia applications in such networks. These applications are typically delay-sensitive and have high bandwidth requirements. Providing QoS in wireless ad hoc networks is challenging because the wireless channel is shared among adjacent hosts, and network topology can change as hosts move. Recently, much effort has been focused on designing medium access protocols (e.g., the IEEE 802.11e medium access control (MAC) specifications [1] for efficiently managing and allocating resources at the MAC layer) and designing QoS signaling for efficient resource management such as INSIGNIA [2]. However, decoupling routing from QoS provisioning can result in the selection of inefficient routes and thus reduce the likelihood of meeting the QoS requirement(s) of arriving communication requests.

The goals of QoS routing are twofold: selecting paths that can satisfy given QoS requirements of arriving communication requests, and achieving global efficiency in resource utilization. Typical routing metrics for providing QoS include delay, jitter, bandwidth, and loss rate. QoS routing plays an important role in providing QoS-aware services in wireless ad hoc networks. Here, we give some key design considerations in QoS routing. First, a designed routing protocol should scale well with respect to the network size, and with respect to computation, communication and storage overhead. Second, it is difficult to guarantee an initial QoS contract with a session that has specific QoS requirement(s), due to network dynamics caused by node mobility. There may exist transient time when the required QoS is not guaranteed due to path break or network partition. Third, a designed protocol should be able to effectively absorb routing information inaccuracy. Fourth, it is necessary to prevent QoS traffic from starving best effort traffic in networks wherein the two traffic types coexist.

Recently, QoS routing has received attention for providing QoS in wireless ad hoc networks, and some work has been carried out to address this critical issue. In this article we first present a review of existing work addressing the issue of route selection subject to QoS constraint(s) in wireless ad hoc networks, and discuss their merits and deficiencies. We devise an on-demand QoS-based unicast routing protocol that works to discover cost-effective routes subject to a delay constraint of low overhead.

The rest of this article is organized as follows. We give the network to be studied and routing problems to be addressed in this article. We present a review of existing work and discuss their merits and deficiencies. We devise a delay-constrained on-demand routing protocol. We provide simulation results to evaluate the performance of the devised protocol. We conclude this article and give possible future directions in the QoS routing area.

NETWORK MODEL AND ROUTING PROBLEMS

In this section we first describe the communication network under study and then present the QoS routing problems to be addressed.

0163-6804/05/$20.00 © 2005 IEEE

A network is modeled as a set V of nodes that are interconnected by a set E of communication links. V and E change over time when nodes move. Nodes have the same maximum transmission range. Each node is equipped with an omnidirectional antenna. Two nodes are immediate neighbors, and an undirected link connecting them exists if they are in the transmission range of each other. Hereafter, we use the terms *host*, *router*, and *node* interchangeably unless otherwise stated.

The state information associated with a link $(i,j) \in E$ includes:

- $Cost(i,j)$, which can be simply one as hop count or a function of link utilization
- $Delay(i,j)$, the delay a packet experiences when it goes through the link
- $Width(i,j)$, the residual (available) bandwidth on the link

For a directed path p, the end-to-end cost (or delay) of path p is the sum of the cost (or delay) of its constituent links. Calculating path bandwidth largely depends on the medium access protocols used for channel access and resource management at the MAC layer.

Before presenting the routing problems to be studied in designing QoS routing protocols, we first discuss some general assumptions made in the design of such protocols. First, as for *resource availability*, each node is assumed to be able to monitor the available resources (e.g., delay and cost) on each of its outgoing links. Second, as for *resource reservation*, a medium access protocol is assumed to be able to resolve media contention and support resource reservation (as necessary) at the MAC layer.

In this article we study the following two routing problems: delay-constrained least-cost (DCLC) unicast routing and bandwidth-constrained least-cost (BCLC) unicast routing. The DCLC issue is to select the path connecting a source-destination pair and with the minimal cost among those meeting the given delay requirement. The DCLC problem is known to be NP-complete. The BCLC issue is to select the path with the minimal cost among those meeting the given bandwidth requirement. The BCLC issue is known to be polynomial in networks wherein path bandwidth is decided by the available bandwidth on its bottleneck link while being NP-complete in time-slotted wireless ad hoc networks [3].

Related Work

QoS routing has received 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 routing issues in wireless ad hoc networks. In general, QoS routing can be classified into two basic paradigms: source QoS routing and hop-by-hop QoS routing. Hereafter, the term *routing* will refer to QoS routing unless otherwise specified.

With source routing, the source node of a communication request locally computes the entire constrained path to the intended destination with the global state information that it locally maintains. Gathering and maintaining global state information can introduce excessive protocol overhead in dynamic networks and thus have the scalability issue. Moreover, the calculation of constraint(s)-based routes would be computationally intensive for the calculating nodes.

In [4] Shah *et al.* designed a predictive location-based QoS routing protocol. This protocol is mainly to alleviate the scalability issue with respect to communication overhead in implementing source routing. Instead of disseminating the state of each link network-wide, each node broadcasts its node status (including its current position, velocity, moving direction, and available resources on each of its outgoing links) across the network periodically or upon a significant change. With such information, at any instant each node can locally depict an instant view of the entire network. To accommodate a QoS request, the source locally computes a QoS-satisfied route (if available) and route data packets along the calculated path. Moreover, the source can predict route break and predictively compute a new route before the old route breaks by using the global state it stores. This routing protocol is suitable for providing soft QoS in small or medium-sized networks wherein mobile hosts are equipped with Global Positioning System (GPS) receivers and their moving behavior is predictable.

In [5] Sinha *et al.* presented the Core-Extraction Distributed Ad Hoc Routing (CEDAR algorithm. CEDAR is designed to select routes with sufficient bandwidth resources. CEDAR dynamically manages a core network, on which the state information of those stable high-bandwidth links is incrementally propagated. Each core node is responsible for maintaining its local topology as well as calculating routes on behalf of nodes in its vicinity. CEDAR selects QoS routes upon request. In detail, a core path is first established by flooding a control message across the core network. A QoS route calculation is then performed to shorten this core path using the partial topology information a core keeps. The performance of CEDAR largely depends on how well core nodes can manage their local resources. In [5] the load factor at core nodes and how it can affect the network performance are not addressed.

With hop-by-hop routing, the task of route selection is shared among intermediate nodes between the source and the destination. There is no centralized computational burden on any node, which enables hop-by-hop routing to scale well. To implement hop-by-hop routing, there are currently the following three routing strategies employed in existing work: shortest path routing, flooding, and multiple paths routing.

The shortest path routing strategy simply returns the shortest path if this path meets the QoS requirement(s) of an arriving request, or otherwise rejects the request. An example protocol following this strategy can be found in [3]. This protocol works to make a routing decision on accepting an arriving request with a specific bandwidth requirement or not by simply checking the feasibility of the min-hop path connecting the source-destination pair of the request. Advantages of this strategy include simplicity, fast route acquisition, and low control overhead.

With hop-by-hop routing, the task of route selection is shared among intermediate nodes between the source and the destination. There is no centralized computational burden on any node, which enables hop-by-hop routing to scale well.

The design of ODRP focuses on the operations at the network layer and it assumes the capabilities of determining resource availability on neighboring links and the availability of resource reservation functions at nodes.

It can work well when traffic demand is light such that the min-hop path probably has enough resources to accommodate a QoS request. However, this strategy can suffer from low request acceptance rate as traffic demand increases. This is simply because the infeasibility of the min-hop path does not mean the nonexistence of feasible paths in the network.

Flooding is another strategy that supports QoS routing, which works by flooding a route-searching message across the entire network to search for a QoS route on demand. Intermediate nodes forward nonduplicate route searching messages that they receive provided that the given QoS requirement(s) has not been violated yet. Once the intended destination receives a route searching message, a QoS route is discovered. In [6, 7] this strategy was used to acquire bandwidth-constrained paths in time-slotted ad hoc networks. In [7] the issue of power assignment at intermediate nodes was further studied to ensure satisfied signal-to-interference (SIN) in the selection of such bandwidth-constrained paths. Moreover, to reduce blocking probability, multiple routes can be used to accommodate a request with a high bandwidth requirement. Searching for QoS routes through pure flooding can achieve good success probability in acquiring QoS routes due to its wide searching scope. However, pure flooding, an expensive operation in resource-scarce ad hoc networks, can introduce excessive communication overhead.

Multiple paths routing aims at achieving a good trade-off between success probability in route acquisition and protocol overhead. It works by searching multiple paths in parallel for a QoS path. In [8] Chen *et al.* designed a protocol named Ticket-Based Probing (TBP). TBP requires multiple routing daemons to run in parallel in the network, one for each of the concerned metrics, to obtain the distance from each node to all other nodes with respect to each of the metrics. To search for a QoS route, the source issues a fixed number of probe packets, each carrying a ticket. Each *probe* is in charge of searching for a path, if possible. The maximum number of *probes* at any time is bounded by the number of tickets. TBP can largely absorb information inaccuracy due to its multipronged nature. Although elegant in concept, TBP has the following difficulties in its implementation. First, proactively running multiple wide-area routing daemons in a network can generate excessive communication overhead. Second, storing one-hop neighbors' routing tables at each node can be a big concern in terms of storage overhead.

In [9] Zhang *et al.* designed an alternate QoS routing mechanism. It works by searching for alternate QoS-satisfied paths if the shortest path is not qualified to accommodate a request with specific QoS requirement(s). Alternate route candidates are restricted to be among those concatenated paths meeting the form of a combination of two shortest path segments, one from the source directly to an intermediate node, and the other from that intermediate node directly to the destination. Among all such concatenated paths, the one with minimal cost is selected provided that the given QoS requirement(s) are met. Directional restricted search is employed to further reduce communication overhead. This alternate routing mechanism, however, can far or less suffers from information inaccuracy because, in dynamic networks, the QoS properties of a path may deviate from that a node locally stores as time evolves. Table 1 gives an overall comparison of protocols designed to support QoS routing in wireless ad hoc networks.

Delay-Constrained Unicast Routing Protocol

In this section we devise an On-Demand Delay-Constrained Unicast Routing Protocol (ODRP) for wireless ad hoc networks. The design of ODRP focuses on the operations at the network layer and assumes the capabilities of determining resource availability on neighboring links and the availability of resource reservation functions at nodes, as discussed earlier.

Routing Information

For ODRP to work correctly, each node is required to maintain a distance vector consisting of $|V|-1$ entries, one for every other node in the network. The entry for v at node u ($u \neq v$) 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 (best effort) distance vector routing protocol in the network. ODRP utilizes the vectors stored at different nodes to guide route-searching packets to propagate in the promising direction and avoid pure flooding. No QoS information related to such best-effort paths is assumed to avoid using outdated information. This assumption can decouple the design of ODRP from the characteristics of the underlying unicast routing protocol.

Procedures

ODRP is designed to effectively reduce the communication overhead consumed in acquiring a low-cost delay-constrained path while achieving high route acquisition success probability. To achieve this objective, ODRP employs the following strategies in its route-searching operations: hybrid routing, directional search, and link-delay-based scheduling of (control) packet forwarding. The hybrid routing strategy works to first probe the feasibility of the min-hop path connecting the source-destination pair of an arriving QoS request. This path is returned if feasible; otherwise, a destination-initiated route-searching process via restricted flooding is enforced. Directional search is employed to restrict the search range of the route-searching process. In this way, the rate at which expensive (restricted) flooding operations are enforced can be greatly reduced. Link-delay-based scheduling of control packet forwarding is for an intermediate node to decide on when it retransmits a Route Request (RREQ) packet it receives. In ODRP, this retransmission is scheduled at a speed proportional to the delay of the link over which the packet was received. If multiple RREQs are received, the one carrying the least delay value is chosen. If the scheduling functions

Protocols	Strategies	Topology management		Metrics	Communication overhead for route discovery	Route acquisition latency
		Routing information kept at nodes	Storage overhead at each node			
Location-aided Routing [4]	Source routing	Global state	$O(\|V\|)$	Bandwidth/ delay	Zero	Low
CEDAR [5]	Restricted flooding + localized source routing	Partial network state	Partial network state	Bandwidth	Moderate	High
Min-hop routing [3]	Shortest path routing	Distance vector	$O(\|V\|)$	Bandwidth	$O(\sqrt{\|V\|})$	Low
Bandwidth routing [6, 7]	Flooding	Local state	$O(m)$	Bandwidth	O(|V|)	High
TBP [8]	Multiple paths routing	Distance vector	$O(km\|V\|)$	Bandwidth/ delay	$O(T\sqrt{\|V\|})$	Moderate
Alternate routing [9]	Shortest path + alternate routing	Distance vector	$O(\|V\|)$	Bandwidth/ delay	Moderate	Moderate

m: Average number of one-hop neighbors of a node.
T: Number of tickets issued in a route searching process using TBP.
$|V|$: Number of nodes in the network.
$|E|$: Number of links in the network.
k: Number of concerned metrics.

Table 1. *Comparisons of protocols supporting QoS routing in wireless ad hoc networks.*

used at different nodes are the same, a low delay path can be acquired as a result of such collaborative operations at intermediate nodes. Note that a blind flooding process does not necessarily lead to a delay-constrained path (when available) due to the potentially chaotic propagations of RREQs in such a process.

The operations of ODRP accordingly can be divided into the following two phases. The first phase is to probe the feasibility of the min-hop path. If it is not feasible, a second phase is enforced to perform destination-initiated route discovery.

Phase I: Probing the Feasibility of the Min-Hop Path — Source s initiates a process to acquire a constrained path upon receiving a request for a route to a destination d subject to a delay constraint Δ. For this purpose, s sends a probe packet directly to d along the min-hop path connecting them and starts a timer. This probe packet gathers the accumulated delay information when propagating along the path. Destination d operates as follows upon receiving the probe packet. If the min-hop path connecting the s-d pair meets the delay requirement, a decision made by using the information that the probe message gathers, then a constrained path has been identified. Destination d then sends an ACK packet back to the source along the reverse path and accordingly creates an entry and reserves resources at each intermediate node. When source s receives this ACK, it can send data packets along the path. The success of this process returns a least-cost constrained path for the request.

Phase II: Destination-Initiated Route Discovery for a Delay-Constrained Path — The second phase starts if the min-hop path is not feasible, and operations in this phase work as follows. Destination d initiates a path discovery process by flooding a RREQ packet, which contains:

- Identifiers of source s and destination d
- A session ID and a sequence number, as copied from the probe message d received
- Delay constraint Δ
- The accumulated delay on the path probed so far, which is initially zero
- Searching scope limit $D = MinHop(d,s) + H$, where $MinHop(d,s)$ represents the min-hop distance from d to s and H is a small positive integer
- A flag F being set, which indicates the destination-initiated property of the routing process

An intermediate node x (x(s,d) executes the following operations upon receiving such an RREQ. It first determines whether it should accept the incoming RREQ or not. The preconditions under which x accepts such a RREQ are as follows:

1 $MinHop(x,s) + MinHop(x,d) \le \Delta$.
2 $Delay\ (p) \le d$.
3 Node x receives the RREQ for the first time or the RREQ carrying a lower path delay value than those x received earlier.

Here p represents the downstream partial path discovered thus far. Condition (1) is to check whether node x is located inside a predetermined searching range. Figure 1 illustrates how ODRP creates a (reduced) searching zone by

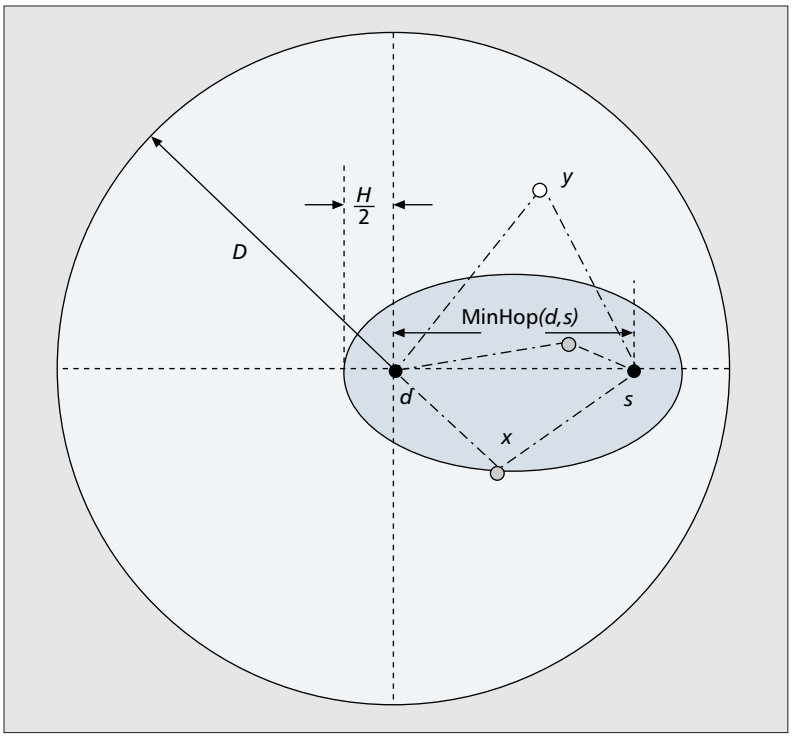

Figure 1. *Illustrating the searching range created by the destination-initiated directional search process in ODRP for a QoS route from source* s *to destination* d *subject to a path length constraint* D. *Here* D = MinHop(d,s) + H. *Here,* MinHop(d,s) *represents the min-hop distance from* d *to* s *and* H *is a small positive integer. The shaded ellipse area is the intended searching zone. Note that the distance in this figure is measured in hop count.*

using the distance vector information stored at nodes in the network. This is the so-called directional search strategy. Condition (2) is to ensure nonviolation of the given delay constraint. Condition (3) is to ensure that an intermediate node forwards only the RREQ carrying the least delay value among those that it receives. Once node x accepts an RREQ, it schedules its own rebroadcasting of the packet with a deferring proportional to the delay value of the outgoing link over which the packet was received. When x forwards the packet, it updates the delay value on the subpath discovered thus far. Node x records its next hop as the neighboring node from which it received the RREQ carrying the least delay value and ignores any further duplicate incoming RREQs belonging to the request. As a result, each node in the network (more exactly, those inside the predetermined forwarding zone) forwards the RREQ at most once.

When s receives an RREQ carrying a path delay value $\leq \Delta$, a constrained path is identified. It then can send data packets along this path or wait for a certain time to collect more (qualified) RREQs and then select the one carrying the least-cost value. The first sent data packet will reserve resources at intermediate nodes along the path. If timed out without receiving a corresponding acknowledgment (ACK) or an RREQ taking a constrained path, s temporally sends data packets as best effort traffic along the min-hop path or performs QoS renegotiation.

The above ODRP design has the following desirable features. First, path discovery through restricted flooding is enforced only when the min-hop path does not meet the delay requirement, which helps to reduce communication overhead. Second, the route searching process is directed and restricted by a predetermined searching range limit, which creates an ellipse searching zone. This restriction can further reduce communication overhead at little penalty in route acquisition probability if H is properly set. The routing process in the second phase of ODRP degenerates to network-wide flooding if H is set to infinity.

Path Maintenance

Upon detecting a link break, the upstream node of the broken link sends a Route Error (RERR) packet to the source of the session and the downstream node sends a Release message downstream to the destination to remove the entries at intermediate nodes along the path and release the resources reserved earlier. A practical method for a downstream node to detect a link break can be by not receiving Hello or data packets from the upstream node of a link for a certain time. Upon receiving an RERR, the source enforces a route rediscovery process if it still has data packets to send to the destination node of the session.

Tuning Latencies at Nodes

Here, we discuss an implementation issue related to the route discovery using ODRP. To forward an RREQ, an intermediate node schedules a local latency as $T = (1 + \gamma) \cdot delay(e)$, where e is the link based on which the latency is scheduled. Note that T is calculated using average link delay value. γ is small and positive. Such a larger setting of latency is to absorb possible uncertainty in the actual delay experienced in packet forwarding with a certain penalty in route acquisition latency.

In the earlier description of ODRP, we did not consider the possible uncertainty in actual node traversal time a packet may experience when traversing an intermediate node. This uncertainty is mainly attributed to MAC access latency and possibly queuing delay since others such as propagation, transmission, and processing delay are in general deterministic for a given radio channel, which can then be expressed in total as a constant C. Uncertainty in queuing delay is expected to be small in a system wherein control packets have priority over data packets in queuing, and appropriate scheduling discipline is employed for effective queue management. It can thus be ignored.

To capture the best rebroadcasting time, a node sends an RREQ to its MAC layer at an instant when the *remaining* scheduled latency is equal to $T_{MACaccess}$s + C, where $T_{MACaccess}$ represents the average medium access latency that the node locally measures. In this way the total latency the RREQ experiences at the current node is expected to be T. For when the uncertainty in medium access latency is nonnegligible, we present here a simple method for absorbing such uncertainty or at least minimizing its negative effect on routing performance. Each node inserts a piece of information into the RREQ to send, which specifies the difference between the

time the packet actually experiences and the average case value. The latency with which a receiving node schedules its retransmission of the packet should subtract this (either negative or positive) difference. As a result, it is expected that the uncertainty in latency does not accumulate in the propagation of the packet. If a prioritized access mechanism is used for high-priority packets in medium accessing, the uncertainty in MAC latency is expected to be negligible.

Simulation Results

In this section we conduct simulations to evaluate the performance of the devised protocol. The following performance metrics are measured: success ratio, average control message overhead, and average path cost. Success ratio is defined as the total number of routed connection requests over the total number of arriving requests. Average control message overhead per request is defined as the total number of control messages sent over the total number of requests. Average path cost is defined as the total cost of established paths over the total number of routed requests.

Networks are generated such that 40 nodes are randomly and uniformly distributed within a 15×15 m^2 area. Nodes are static. The transmission radius of each node is 3.5 m. A link is added between two nodes if they are within transmission range of each other. The source and destination of each request are randomly selected. Each link has unit cost and is associated with a delay value uniformly distributed in the range of [1, 50] ms. In our simulation an ideal MAC protocol is assumed, which can resolve the issues of hidden/exposed nodes and guarantee delivery. Nodes can enable promiscuous receiving mode. Sending an RREQ or RREP message over a link is counted as a control message.

The following routing algorithms were evaluated in our simulations: Flood-1 [10], Flood-2, ODRP, the alternate Delay-Constrained Routing (DLR) algorithm [9], and shortest path routing with respect to hop count (SP). The Flood-1 algorithm floods a route searching message across the entire network for discovering a delay-constrained route. A route-searching message proceeds only if the accumulated delay on the path that it takes does not violate the delay bound. In [10], Shin et al. showed that, if certain scheduling policies are used at nodes and control messages are set to the appropriate priority, the route-searching messages travel at speeds according to the link delay. Hence, the message traveling along the least delay path arrives first. In this case, an intermediate node needs only to forward the first received message to each of its neighbors except the one from which the message was received. This algorithm finds a feasible route when one exists. The total number of control messages sent in such a path searching process, in the worst case, is $O(|E|)$. Flood-1 was simulated in [8, 9] as a benchmark for comparison. Flood-2 realizes the second phase of ODRP but using a source-initiated *network-wide* flooding process for route discovery by setting H to infinity. Flood-2 thus reduces the communication overhead to $O(|V|)$. Neither of the Flood algorithms assumes any global state information at nodes; therefore, no search scope limit can be enforced.

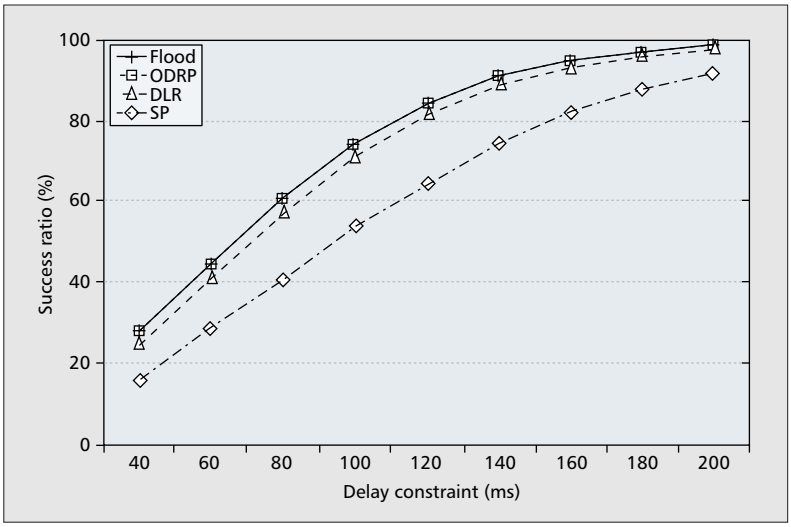

Figure 2. *Comparisons of success ratio in acquiring a delay-constrained path vs. delay bound. Here Flood represents both Flood-1 and Flood-2 (same in Fig. 4).*

The SP algorithm works as follows. The source s sends a probe packet along the min-hop path to the intended destination to check its feasibility. If this path is feasible, the destination sends an ACK back to source s; otherwise, the intermediate node at which the delay bound is violated sends a negative ACK (NAK) back to s.

In implementing each of the protocols, local query is enforced to perform local connection establishment where the destination is within the transmission range of the source in order to reduce communication overhead. For each measured value presented, each of the studied protocols was run repeatedly with new random requests on different randomly-created graphs until a 5 percent confidence interval with a 95 percent confidence level was achieved. In the implementation of DLR, when a constrained path is identified, a Reserve packet is sent along the path to reserve resources at intermediate nodes for the connection, and the destination sends an ACK to notify the source of successful connection establishment.

Figure 2 shows that the *success ratio* due to each of the simulated protocols increases with relaxation of the delay constraint. The Flood-based algorithms have the best success ratio since they return a feasible path when one exists. The success ratio of ODRP is very close to the optimal value. An additional experiment was also conducted to determine the appropriate value of parameter H. Through extensive simulations, we set H to two in all other experiments. In the experiments little gain with respect to success ratio was observed by increasing H further. DLR performs worse than the optimal case due to its restriction on the selection of alternate path candidates. SP has the lowest success ratio.

Figure 3 compares the *average control message overhead* per request. We can see that protocols from highest to lowest in terms of message overhead are Flood-1, Flood-2, ODRP, DLR,

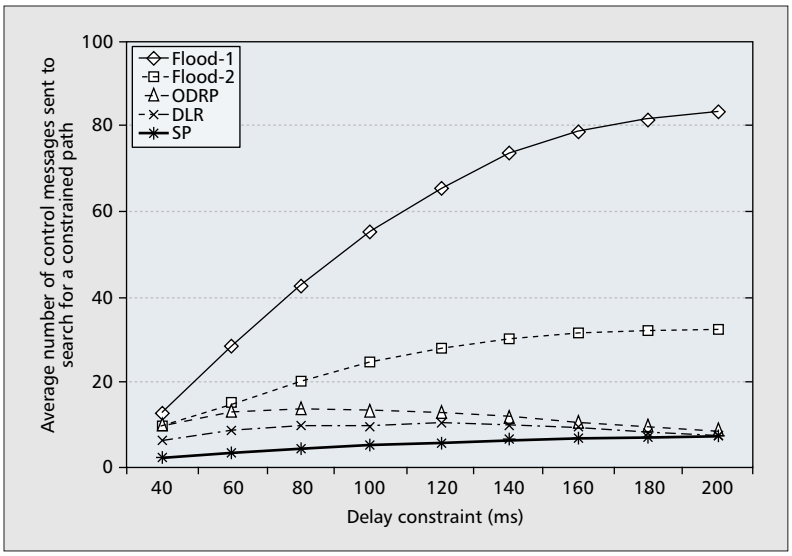

Figure 3. *Comparisons of average control overhead per request using different protocols vs. delay bound.*

and SP. SP performs the best because it checks a single path for a routing decision. The reason the communication overhead resulting in ODRP drops after the delay constraint exceeds a certain threshold is as follows. With the relaxation of delay bound, the probability that a min-hop path satisfies the delay requirement of a request increases and its feasibility avoids the enforcement of restricted flooding for route discovery. Note that although the per-request overhead for route acquisition by DLR is slightly smaller than that by ODRP, the overhead for disseminating the state information required by DLR is much larger than for ODRP. This is because DLR requires the underlying protocol to proactively gather and disseminate additional QoS-related path information and keep it up to date continuously.

Figure 4 compares the *average path cost* performance of the simulated protocols. In our implementations of ODRP and Flood, the constrained path with minimal cost is selected provided that multiple feasible choices are available. Figure 4 shows that protocols with the worst cost performance to the best are Flood, ODRP, DLR, and SP. SP achieves the best performance by rejecting all those requests whose respective least cost path is infeasible. The better cost performance of DLR is because its implementation requires more QoS-related path information to be kept at nodes, with which DLR can easily find and then rank alternate feasible paths for the one with the lowest cost among them. By decoupling QoS routing from the characteristics of the underlying protocol, ODRP aims to find a good QoS path with limited information at nodes.

In summary, the performance of a QoS routing protocol can be greatly affected by the routing information available at nodes. More detailed state information can lead to much better routing performance while being more sensitive to information inaccuracy caused by network dynamics. Compared to Flood-2, ODRP achieves near-optimal success ratio, improved cost performance, and reduced communication overhead. Compared to DLR, ODRP is superior with respect to success ratio and inferior with respect to path cost and communication overhead. Therefore, the design devised in ODRP achieves a good trade-off between success ratio and communication overhead as well as path cost by storing limited state information at nodes to support efficient QoS routing.

Conclusions and Future Directions

QoS routing is an essential component of a QoS architecture. In this article we first present a review of existing work addressing the issue of route selection subject to QoS constraint(s) for wireless ad hoc networks and discuss in detail their merits and deficiencies. We have devised an on-demand delay-constrained unicast routing protocol, ODRP. ODRP employs various strategies to effectively reduce its resultant communication overhead to acquire cost-effective delay-constrained paths. Simulations results demonstrate that the design in ODRP meets our design goals. Moreover, the hybrid routing and directional search strategies in ODRP can also be used to support cost-effective bandwidth-constrained routing in wireless ad hoc networks.

QoS routing in wireless ad hoc networks is challenging. Although some work has been carried out to address this critical issue, research in this area is far from exhaustive. Here, we provide some topics that fall into the area of QoS routing and deserve further investigation.

Scalability. This is also an overruling concern in designing ad hoc networks. The performance of an algorithm for selecting routes subject to QoS constraint(s) primarily depends on the way state information is disseminated, gathered, and maintained, as well as the accuracy of the information. In general, fundamental trade-offs in efficient QoS routing exist between communication overhead for disseminating state information and maintaining this information as well as its inaccuracy, and between overhead for acquiring QoS routes and performance in terms of success probability as well as global resources utilization. Moreover, designing a routing protocol should also consider constraints posed by the wireless ad hoc network environment.

Integration with other network components. MAC protocols can greatly affect the design of a routing protocol and its performance in wireless ad hoc networks. Some existing QoS routing protocols focused on route selection subject to a bandwidth requirement over time-slotted ad hoc networks [3, 6, 7]. QoS routing can benefit a lot from efficient support of resource management and allocation at the MAC layer.

QoS multicast routing. Most existing multicast routing protocols for ad hoc networks were designed to support best effort services. How to provide multicasting subject to QoS constraint(s) in wireless ad hoc networks needs further study.

QoS routing with power control. Power control can effectively reduce interference between neighboring nodes and increase network throughput. In networks wherein nodes have

transmission power adjusting capabilities, a low-power route prefers routes consisting of more short-haul hops. However, this decision on route selection can lead to paths with high packet delivery latency. Therefore, a good trade-off is needed between supporting QoS routing and power control. Furthermore, the selection of QoS routes should also take into consideration the residual energy at nodes in order to balance the energy depletion among nodes and therefore prolong the network operational time.

In summary, to provide efficient QoS routing over wireless ad hoc networks, problems such as scalability, power control, energy drain balancing, and resource management including designing efficient QoS-aware MAC protocols and signaling need to be further investigated.

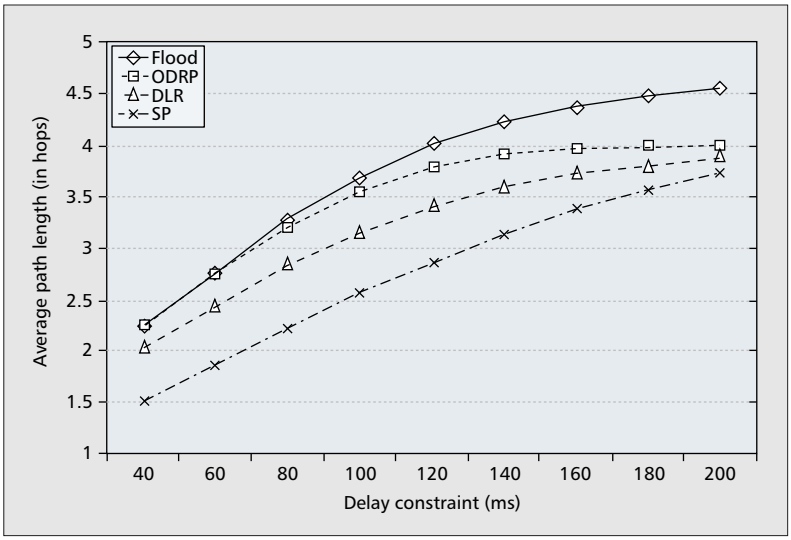

Figure 4. *Comparisons of average path cost by different protocols versus delay bound.*

REFERENCES

[1] IEEE 802.11e/D6.0, "Draft Amendment to Standard for Information Technology — Telecommunications and Information Exchange between Systems — LAN/MAN Specific Requirements — Part 11: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Medium Access Control (MAC) Quality of Service (QoS) Enhancements," Nov. 2003.

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

[3] C. R. Lin and J.-S. Liu, "QoS Routing in Ad Hoc Wireless Networks," *IEEE JSAC*, vol. 17, no. 8, Aug. 1999, pp. 1426–38.

[4] S. H. Shah and K. Nahrstedt, "Predictive Location-based QoS Routing in Mobile Ad Hoc Networks," *Proc. IEEE ICC '02*, Apr. 2002, pp. 1022–27.

[5] P. Sinha, R. Sivakumar, and V. Bharghavan, "CEDAR: A Core-extraction Distributed Ad Hoc Routing Algorithm," *IEEE JSAC*, vol. 17, no. 8, Aug. 1999, pp. 1454–65.

[6] C. Zhu and M. S. Corson, "QoS Routing for Mobile Ad Hoc Networks," *Proc. IEEE INFOCOM '02*, June 2002, pp. 958–67.

[7] D. Kim, C.-H. Min, and S. Kim, "On-demand SIR and Bandwidth-guaranteed Routing with Transmit Power Assignment in Ad Hoc Mobile Networks," *IEEE Trans. Vehic. Tech.*, vol. 53, no. 4, July 2004, pp. 1215–23.

[8] S. Chen and K. Nahrstedt, "Distributed Quality-of-service Routing In Ad Hoc Networks," *IEEE JSAC*, vol. 17, no. 8, Aug. 1999, pp. 1488–505.

[9] B. Zhang and H. T. Mouftah, "QoS Routing Through Alternate Paths in Wireless Ad Hoc Networks," *Int'l. J. Commun. Sys.*, vol. 17, no. 3, Mar. 2004, pp. 233–52.

[10] K. G. Shin and C.-C. Chou, "A Distributed Route-selection Scheme for Establishing Real-time Channel," *Proc. SPIE HPN '95*, Sept. 1995, pp. 319–29.

BIOGRAPHY

BAOXIAN ZHANG [M] (bxzhang@site.uottawa.ca) received his B.S., M.S., and Ph.D. degrees in electrical engineering from Northern Jiaotong University, Beijing, China, in 1994, 1997, and 2000, respectively. From January 2001 to August 2002 he was working with the Department of Electrical and Computer Engineering, Queen's University, Kingston, Ontario, Canada, as a postdoctoral fellow. He is currently with the School of Information Technology and Engineering (SITE), University of Ottawa, Canada. He has published over 40 refereed technical papers in international journals and conference proceedings. His research interests include routing, QoS, wireless ad hoc and sensor networks, survivable optical networks, and performance evaluation.

HUSSEIN MOUFTAH [F'90] (mouftah@site.uottawa.ca) joined the School of Information Technology and Engineering (SITE) of the University of Ottawa in September 2002 as a Canada Research Chair (Tier 1) Professor in Optical Networks. He was with the Department of Electrical and Computer Engineering at Queen's University (1979–2002), and prior to his departure was a full professor and Department Associate Head. He has three years of industrial experience, mainly at Bell Northern Research of Ottawa, now Nortel Networks (1977–1979). He also spent three sabbatical years at Nortel Networks (1986–1987, 1993–1994, and 2000–2001), always conducting research in the area of broadband packet switching networks, mobile wireless networks, and QoS over the optical Internet. He served as Editor-in-Chief of *IEEE Communications Magazine* (1995–1997), IEEE Communications Society Director of Magazines (1998–1999), and Chair of the Awards Committee (2002–2003). He has been a Distinguished Speaker of the IEEE Communications Society since 2000. He is the author or coauthor of five books, 22 book chapters, more than 700 technical papers, and eight patents in this area. He is the recipient of the 1989 Engineering Medal for Research and Development of the Association of Professional Engineers of Ontario (PEO), and the Ontario Distinguished Researcher Award of the Ontario Innovation Trust. He is the joint holder of the Best Paper Award for a paper presented at ICC 2005 Optical Networking Symposium, SPECTS 2002, and the Outstanding Paper Award for papers presented at IEEE HPSR 2002 and IEEE ISMVL 1985. He was a joint holder of a Honorable Mention for the Frederick W. Ellersick Price Paper Award for Best Paper in *IEEE Communications Magazine* in 1993. He was the recipient of the IEEE Canada (Region 7) Outstanding Service Award (1995). He also received the 2004 IEEE Communications Society Edwin Howard Armstrong Achievement Award and the 2004 George S. Glinski Award for Excellence in Research of the Faculty of Engineering, University of Ottawa. He is a Fellow of the Canadian Academy of Engineering (2003) and the Engineering Institute of Canada (2005).