Zone Routing Protocol (ZRP)

Nicklas Beijar
Networking Laboratory, Helsinki University of Technology
P.O. Box 3000, FIN-02015 HUT, Finland
Nicklas.Beijar@hut.fi

Abstract

Routing protocols for mobile ad-hoc networks have to face the challenge of frequently changing topology, low transmission power and asymmetric links. Both proactive and reactive routing protocols prove to be inefficient under these circumstances. The Zone Routing Protocol (ZRP) combines the advantages of the proactive and reactive approaches by maintaining an up-to-date topological map of a zone centered on each node. Within the zone, routes are immediately available. For destinations outside the zone, ZRP employs a route discovery procedure, which can benefit from the local routing information of the zones.

This paper presents the Zone Routing Protocol. First, we discuss the problem of routing in ad-hoc networks and the motivation of ZRP. We describe the architecture of ZRP, which consists of three sub-protocols. We describe the routing process and illustrate it with an example. Further, we describe the query control mechanisms, which are used to reduce the traffic amount in the route discovery procedure. ZRP does not define the actual implementation of the protocol components. Therefore, we present the guidelines for implementation, and example implementations provided in the draft specifications. We discuss the problem of routing in networks with unidirectional links, and the proposal for a solution to it. The overhead of the routing protocol is important in the power and bandwidth limited ad-hoc networks. We discuss the factors influencing on the traffic amount based on measurements performed in a number of papers. We describe the significant issue of choosing an optimal zone radius, and two algorithms for automatic selection of the radius. Finally, we draw some conclusions about the performance of the protocol. The paper is based on literature research.

Keywords: Zone Routing Protocol, ZRP, IARP, IERP, BRP, Ad-hoc network, Routing, MANET

1 Introduction

Ad-hoc networks are mobile wireless networks that have no fixed infrastructure. There are no fixed routers – instead each node acts as a router and forwards traffic from other nodes. Ad-hoc networks were first mainly used for military applications. Since then, they have become increasingly more popular within the computing industry. Applications include emergency search-and-rescue operations, deployment of sensors, conferences, exhibitions, virtual classrooms and operations in environments where construction of infrastructure is difficult or expensive. Ad-hoc networks can be rapidly deployed because of the lack of infrastructure. [2] [16]

A MANET (Mobile Ad-hoc Network) is a type of ad-hoc network with rapidly changing topology. These networks typically have a large span and connect hundreds to thousands of nodes [16]. Correspondingly, the term Reconfigurable Wireless Networks (RWN) refers to large ad-hoc networks that can be rapidly deployed without infrastructure and where the nodes are highly mobile [14]. In this paper, we concentrate on routing in large ad-hoc networks with high mobility.

Since the nodes in a MANET are highly mobile, the topology changes frequently and the nodes are dynamically connected in an arbitrary manner. The rate of change depends on the velocity of the nodes. Moreover, the devices are small and the available transmission power is limited. Consequently, the radio coverage of a node is small. The low transmission power limits the number of neighbor nodes, which further increases the rate of change in the topology as the node moves. Because of interference and fading due to high operating frequency in an urban environment, the links are unreliable. Ad-hoc networks are further characterized by low bandwidth links. Because of differences in transmission capacity, some of the links may be unidirectional. As a result of link instability and node mobility, the topology changes frequently and routing is difficult.

1.1 Routing in ad-hoc networks

A number of routing protocols have been suggested for ad-hoc networks [2]. These protocols can be classified into two main categories: proactive (table-driven) and reactive (source-initiated or demand-driven).

Proactive routing protocols attempt to keep an up-to-date topological map of the entire network. With this map, the route is known and immediately available when a packet needs to be sent. The approach is similar to the one used in wired IP networks, for example in OSPF [3]. [1] [2]

Proactive protocols are traditionally classified as either distance-vector or link-state protocols. The former are based on the distributed Bellman-Ford (DBP) algorithm, which is known for slow convergence because of the “counting-to-infinity” problem. To address the problem, the Destination-Sequenced Distance-Vector routing (DSDV) [4] protocol was proposed for ad-hoc networks. On the other hand, link-state protocols, as represented by OSPF [3], have become standard in wired IP networks. They converge more rapidly, but require significantly more control traffic. Since ad-hoc networks are bandwidth limited and their topology changes often, an Optimized Link-State Protocol (OLSR) [5] has been proposed. While being suitable for small networks, some scalability problems can be seen on larger networks. The need to improve convergence and reduce traffic has led to algorithms that combine features of distance-vector and link-state schemes. Such a protocol is the wireless routing protocol (WRP) [6], which eliminates the counting-to-infinity problem and avoids temporary loop without increasing the amount of control traffic. [1] [10]

In contrast to proactive routing, reactive routing does not attempt to continuously determine the network connectivity. Instead, a route determination procedure is invoked on demand when a packet needs to be forwarded. The technique relies on queries that are flooded throughout the network. [1]

Reactive route determination is used in the Temporally Ordered Routing Algorithm (TORA) [7], the Dynamic Source Routing (DSR) [8] and the Ad-hoc On-demand Distance Vector (AODV) [9] protocols. In DSR and AODV, a reply is sent back to the query source along the reverse path that the query traveled. The main difference is that DSR performs source routing with the addresses obtained from the query packet, while AODV uses next-hop information stored in the nodes of the route. In contrast to these protocols, TORA creates directed acyclic graphs rooted at the destination by flooding the route replies in a controlled manner. [1] [2] [10]

1.2 Comparison of proactive and reactive routing

Both proactive and reactive routing have specific advantages and disadvantages that make them suitable for certain types of scenarios. Since proactive routing maintains information that is immediately available, the delay before sending a packet is minimal. On the contrary, reactive protocols must first determine the route, which may result in considerable delay if the information is not available in caches. [1]

Moreover, the reactive route search procedure may involve significant control traffic due to global flooding. This, together with the long setup delay, may make pure reactive routing less suitable for real-time traffic. However, the traffic amount can be reduced by employing route maintenance schemes. [10]

Purely proactive schemes use a large portion of the bandwidth to keep routing information up-to-date. Because of fast node mobility, the route updates may be more frequent than the route requests, and most of the routing information is never used. Some of the scarce bandwidth is thus wasted. [1] [10]

2 The Zone Routing Protocol

2.1 Motivation

As seen, proactive routing uses excess bandwidth to maintain routing information, while reactive routing involves long route request delays. Reactive routing also inefficiently floods the entire network for route determination. The Zone Routing Protocol (ZRP) [11]–[13] aims to address the problems by combining the best properties of both approaches. ZRP can be classed as a hybrid reactive/proactive routing protocol. [10]

In an ad-hoc network, it can be assumed that the largest part of the traffic is directed to nearby nodes. Therefore, ZRP reduces the proactive scope to a zone centered on each node. In a limited zone, the maintenance of routing information is easier. Further, the amount of routing information that is never used is minimized. Still, nodes farther away can be reached with reactive routing. Since all nodes proactively store local routing information, route requests can be more efficiently performed without querying all the network nodes. [10]

Despite the use of zones, ZRP has a flat view over the network. In this way, the organizational overhead related to hierarchical protocols can be avoided. Hierarchical routing protocols depend on the strategic assignment of gateways or landmarks, so that every node can access all levels, especially the top level. Nodes belonging to different subnets must send their communication to a subnet that is common to both nodes. This may congest parts of the network. ZRP can be categorized as a flat protocol because the zones overlap. Hence, optimal routes can be detected and network congestion can be reduced. [15]

Further, the behavior of ZRP is adaptive. The behavior depends on the current configuration of the network and the behavior of the users. [10]

2.2 Architecture

The Zone Routing Protocol, as its name implies, is based on the concept of zones. A routing zone is defined for each node separately, and the zones of neighboring nodes overlap. The routing zone has a radius ρ expressed in hops. The zone thus includes the nodes, whose distance from the node in question is at most ρ hops. An example routing zone is shown in Figure 1, where the routing zone of S includes the nodes A–I, but not K. In the illustrations, the radius is marked as a circle around the node in question. It should however be noted that the zone is defined in hops, not as a physical distance. [10]

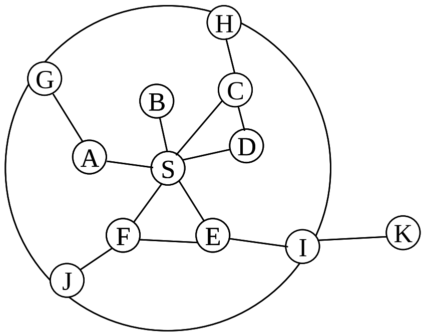

Figure 1: Example routing zone with ρ=2

The nodes of a zone are divided into peripheral nodes and interior nodes. Peripheral nodes are nodes whose minimum distance to the central node is exactly equal to the zone radius ρ. The nodes whose minimum distance is less than ρ are interior nodes. In Figure 1, the nodes A–F are interior nodes, the nodes G–J are peripheral nodes and the node K is outside the routing zone. Note that node H can be reached by two paths, one with length 2 and one with length 3 hops. The node is however within the zone, since the shortest path is less than or equal to the zone radius. [10] [11]

The number of nodes in the routing zone can be regulated by adjusting the transmission power of the nodes. Lowering the power reduces the number of nodes within direct reach and vice versa. The number of neighboring nodes should be sufficient to provide adequate reachability and redundancy. On the other hand, a too large coverage results in many zone members and the update traffic becomes excessive. Further, large transmission coverage adds to the probability of local contention. [10]

ZRP refers to the locally proactive routing component as the IntrA-zone Routing Protocol (IARP). The globally reactive routing component is named IntEr-zone Routing Protocol (IERP). IERP and IARP are not specific routing protocols. Instead, IARP is a family of limited-depth, proactive link-state routing protocols. IARP maintains routing information for nodes that are within the routing zone of the node. Correspondingly, IERP is a family of reactive routing protocols that offer enhanced route discovery and route maintenance services based on local connectivity monitored by IARP. [11] [12]

The fact that the topology of the local zone of each node is known can be used to reduce traffic when global route discovery is needed. Instead of broadcasting packets, ZRP uses a concept called *bordercasting*. Bordercasting utilizes the topology information provided by IARP to direct query request to the border of the zone. The bordercast packet delivery service is provided by the Bordercast Resolution Protocol (BRP). BRP uses a map of an extended routing zone to construct bordercast trees for the query packets. Alternatively, it uses source routing based on the normal routing zone. By employing *query control* mechanisms, route requests can be directed away from areas of the network that already have been covered. [13]

In order to detect new neighbor nodes and link failures, the ZRP relies on a Neighbor Discovery Protocol (NDP) provided by the Media Access Control (MAC) layer. NDP transmits "HELLO" beacons at regular intervals. Upon receiving a beacon, the neighbor table is updated. Neighbors, for which no beacon has been received within a specified time, are removed from the table. If the MAC layer does not include a NDP, the functionality must be provided by IARP. [14]

The relationship between the components is illustrated in Figure 2. Route updates are triggered by NDP, which notifies IARP when the neighbor table is updated. IERP uses the routing table of IARP to respond to route queries. IERP forwards queries with BRP. BRP uses the routing table of IARP to guide route queries away from the query source. [15]

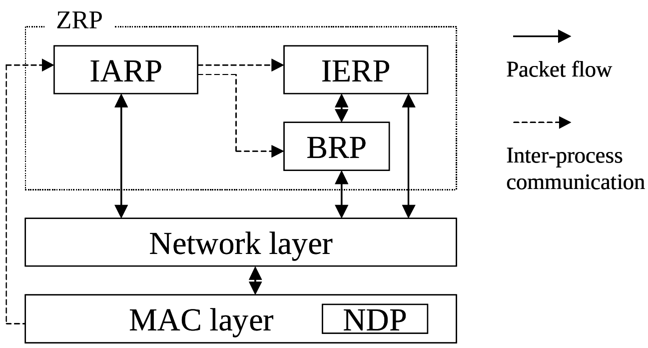

Figure 2: ZRP architecture

2.3 Routing

A node that has a packet to send first checks whether the destination is within its local zone using information provided by IARP. In that case, the packet can be routed proactively. Reactive routing is used if the destination is outside the zone. [13]

The reactive routing process is divided into two phases: the *route request* phase and the *route reply* phase. In the route request, the source sends a route request packet to its peripheral nodes using BRP. If the receiver of a route request packet knows the destination, it responds by sending a route reply back to the source. Otherwise, it continues the process by bordercasting the packet. In this way, the route request spreads throughout the network. If a node receives several copies of the same route request, these are considered as redundant and are discarded [12], [13]

The reply is sent by any node that can provide a route to the destination. To be able to send the reply back to the source node, routing information must be accumulated when the request is sent through the network. The information is recorded either in the route request packet, or as next-hop addresses in the nodes along the path. In the first case, the nodes forwarding a route request packet append their address and relevant node/link metrics to the packet. When the packet reaches the destination, the sequence of addresses is reversed and copied to the route reply packet. The sequence is used to forward the reply back to the source. In the second case, the forwarding nodes records routing information as next-hop addresses, which are used when the reply is sent to the source. This approach can save transmission resources, as the request and reply packets are smaller. [12]

The source can receive the complete source route to the destination. Alternatively, the nodes along the path to the destination record the next-hop address in their routing table. [12]

In the bordercasting process, the bordercasting node sends a route request packet to each of its peripheral nodes. This type of one-to-many transmission can be implemented as multicast to reduce resource usage. One approach is to let the source compute the multicast tree and attach routing instructions to the packet. This is called Root-Directed Bordercasting (RDB). Another approach is to reconstruct the tree at each node, whereas the routing instructions can be omitted. This requires that every interior node knows the topology seen by the bordercasting node. Thus, the nodes must maintain an extended routing zone with radius $2\rho-1$ hops. Note that in this case the peripheral nodes where the request is sent are still at the distance ρ. This approach is named Distributed Bordercasting (DB). [13] [15]

The zone radius is an important property for the performance of ZRP. If a zone radius of one hop is used, routing is purely reactive and bordercasting degenerates into flood searching. If the radius approaches infinity, routing is reactive. The selection of radius is a tradeoff between the routing efficiency of proactive routing and the increasing traffic for maintaining the view of the zone. [12]

2.4 Route maintenance

Route maintenance is especially important in ad-hoc networks, where links are broken and established as nodes move relatively to each other with limited radio coverage. In purely reactive routing protocols, routes containing broken links fail and a new route discovery or route repair must be performed. Until the new route is available, packets are dropped or delayed. [12]

In ZRP, the knowledge of the local topology can be used for route maintenance. Link failures and sub-optimal route segments within one zone can be bypassed. Incoming packets can be directed around the broken link through an active multi-hop path. Similarly, the topology can be used to shorten routes, for example, when two nodes have moved within each other's radio coverage. For source-routed packets, a relaying node can determine the closest route to the destination that is also a neighbor. Sometimes, a multi-hop segment can be replaced by a single hop. If next-hop forwarding is used, the nodes can make locally optimal decisions by selecting a shorter path. [12]

2.5 Example

Consider the network in Figure 3. The node S has a packet to send to node X. The zone radius is $\rho=2$. The node uses the routing table provided by IARP to check whether the destination is within its zone. Since it is not found, a route request is issued using IERP. The request is bordercast to the peripheral nodes (gray in the picture). Each of these searches their routing table for the destination.

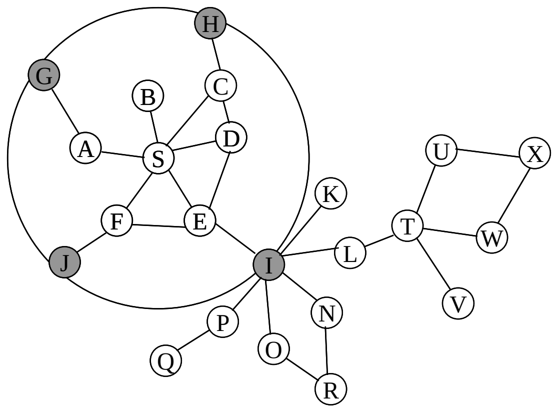

Figure 3: The routing zone of node S

Node I does not find the destination in its routing table. Consequently, it broadcasts the request to its peripheral nodes, shown in gray in Figure 4. Due to query control mechanisms, the request is not passed back to nodes D, F and S.

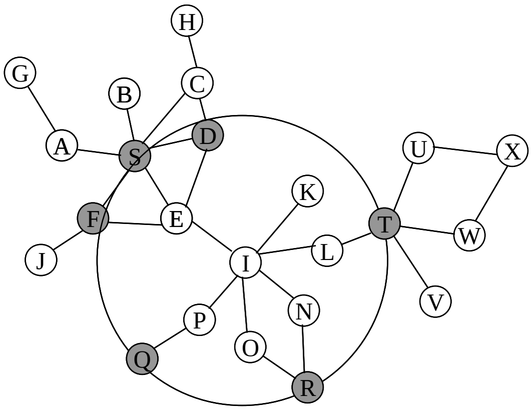

Figure 4: The routing zone of node I

Finally, the route request is received by node T, which can find the destination in its routing zone, shown in Figure 5. Node T appends the path from itself to node X to the path in the route request. A route reply, containing the reversed path is generated and sent back to the source node. If multiple paths to the destination were available, the source would receive several replies.

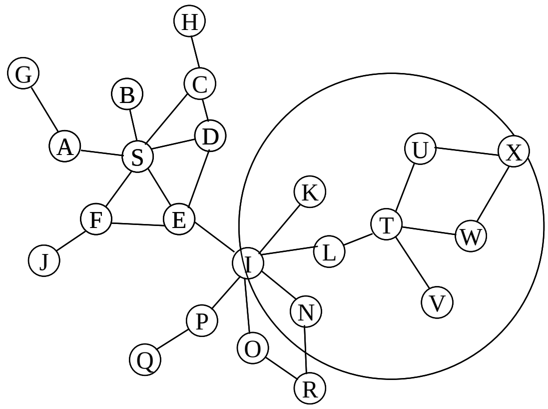

Figure 5: The routing zone of node T

3 Query-control mechanisms

Bordercasting can be more efficient than flooding, since route request packets are only sent to the peripheral nodes, and thus only on the corresponding links. Further efficiency can be gained by utilizing multicast techniques. In that case, only one packet is sent on a link although several peripheral nodes can reside behind this link. [15]

However, since the routing zones of neighboring nodes overlap, each node may forward route requests several times, which results in more traffic than in flooding. When a node bordercasts a query, the complete routing zone is effectively covered. Any further query messages entering the zone are redundant and result in wasted transmission capacity. The excess traffic is a result from queries returning to covered zones instead of covered nodes as in traditional flooding. [15]

To solve this problem, ZRP needs query-control mechanisms, which can direct queries away from covered zones and terminate query packets before they are delivered to peripheral nodes in regions of the network already covered by the query. ZRP uses three types of query-control mechanisms: query detection, early termination and random query-processing delay. Query detection caches the queries relayed by the nodes. With early termination, this information is used to prune bordercasting to nodes already covered by the query. [15]

3.1 Query detection

When a bordercast is issued, only the bordercasting node is aware that the routing zone is covered by the query. When the peripheral nodes continue the query process by bordercasting to their peripheral nodes, the query may be relayed through the same nodes again. To illustrate with an example, the node S in Figure 6 bordercasts a query to its peripheral nodes F–J. As the node J continues by bordercasting to the nodes C, S and E, the query is again relayed by nodes D and E. The query issued by node J to nodes C, S and E is redundant, since these nodes have been covered by the previous query.

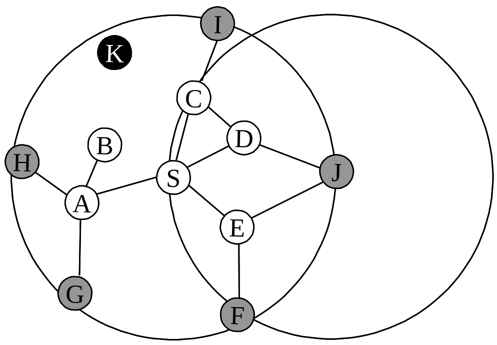

Figure 6: Query detection example

To be able to prevent queries from reappearing in covered regions, the nodes must detect local query relaying activity. BRP provides two query detection methods: QD1 and QD2. Firstly, the nodes that relay the query are able to detect the query (QD1). Secondly, in single-channel networks, it is possible to listen to the traffic by other nodes within the radio coverage (QD2). Hence, it is possible to detect queries relayed by other nodes in the zone. QD2 can be implemented by using IP broadcasts to send route queries. Alternative, unicast can be used if the MAC and IP layers operate in promiscuous mode. [15]

In the above example, all nodes except node B relay the query of S. They are thus able to use QD1. Node B does not belong to the bordercast tree, but it is able to overhear the relayed query using QD2. However, node K does not overhear the message, and is therefore unaware that the zone of node S is covered.

A query detection table is used to cache the detected queries. For each entry, the cache contains the address of the source node and the query ID. The address–ID pair is sufficient to uniquely identify all queries in the network. The cache may also contain other information depending on the query detection scheme. Especially the address of the node that most recently bordercasted a query is important. [15]

3.2 Early termination

With Early Termination (ET), a node can prevent a route request from entering already covered regions. Early termination combines information obtained through query detection with the knowledge of the local topology to prune branches leading to peripheral nodes inside covered regions. These regions consist of the interior nodes of nodes that already have bordercast the query. A node can also prune a peripheral node if it has already relayed a query to that node. [15]

Early termination requires topology information extending outside the routing zone of the node. The information is required to reconstruct the bordercast tree of other nodes within the routing zone. The extended routing zone has a radius of $2\rho-1$. Alternatively, in the case of root-directed bordercast (RDB), the topology of the standard routing zone and information about cached bordercast trees can be used. [15]

In the previous example, node E can use the information in its query detection table to prune the query that the node J sends to its peripheral node F. Node E has an extended routing zone with radius $2\rho-1=3$, shown as a dashed circle in Figure 7.

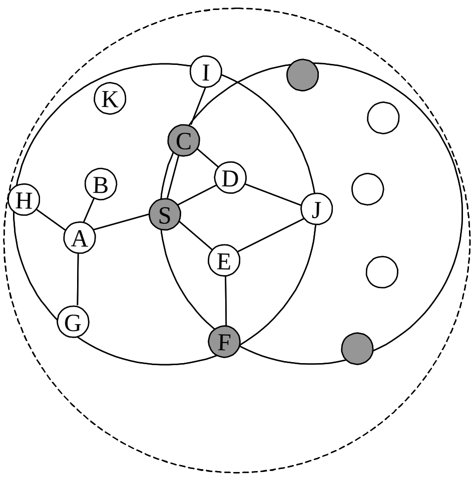

Figure 7: The extended routing zone of E

3.3 Random query-processing delay

When a node issues a node request, it takes some time for the query to be relayed along the bordercast tree and to be detected through the query detection mechanisms. During this time, another node may propagate the same request. This can be a problem when several nearby nodes receive and re-broadcast a request at roughly the same time. [15]

To reduce the probability of receiving the same request from several nodes, a Random Query-Processing Delay (RQPD) can be employed. Each bordercasting node waits a random time before the construction of the bordercast tree and the early termination. During this time, the waiting node can detect queries from other bordercasting nodes and prune the bordercast tree. To avoid additional route discovery delay, the delay can be combined with the pre-transmission jitter used by many route discovery protocols. [15]

Assume that in Figure 7 the nodes C and S both receive a query. Node C schedules a bordercast to its peripheral node E, and node S to its peripheral node F. Without RQPD, both nodes would issue the broadcast simultaneously, and thereafter detect the message of the neighbor node. With RQPD, the node C may detect the query sent by node S during the delay, and prune the branch leading to E.

3.4 Caching

The paper [10] further proposes caching as a technique for reducing control traffic. The nodes cache active routes, and by using this cache, the frequency of route discovery procedures can be reduced. Changes in network topology, such as broken links, are compensated by local path repair procedures. A new path then substitutes the path between the ends of the broken link and a path update message is sent to the endpoints of the path. Since the repair reduces the efficiency of the routes, the endpoints may initiate a new route discovery procedure after a number of repairs.

4 Protocol implementations

Although the ZRP specification drafts [11], [12] do not define the actual IERP and IARP protocols, they provide example IERP and IARP implementations. These are described in terms of packet format, data structures, state machine and pseudo code implementation. Additionally, the documents provide guidelines for converting an existing protocol into an IERP or an IARP.

4.1 Guidelines

A traditional proactive link-state protocol can be modified to be suitable as an IARP by limiting the scope of the routing zone to ρ hops. The limitation is implemented with a time-to-live (TTL) field in the link state update packet. The field has the value ρ*-1* when the packet leaves the source, and the packet is discarded when the value reaches zero. When nodes update their link state tables, sources that are farther than ρ*-1* hops away are discarded. If link state table transfers are performed with new neighbors, sources farther than ρ*-1* hops are excluded. In this way, redundant link state transmissions are reduced to neighbors closer to the link source, and transmission of out-of-zone link state to neighbors farther away is prevented. [11]

Correspondingly, a reactive routing protocol can be converted into an IERP. The protocol must be able to import IARP routes into its routing table and it must be able to support lookups with the IARP routing table. There should not be any local proactive route updates and neighbor advertisements, since this functionality is performed by IARP. Instead of broadcasting route request packets, the protocol should bordercast the route requests with BRP. Flood control and other forms of redundant query termination must be disabled since this is handled by BRP. Nevertheless, route requests can be discarded based on other criteria, such as successful route discovery, exceeded QoS metrics and expired TTL. Also jittering of route requests is provided by BRP. The protocol may use advanced route maintenance techniques, such as on-line route repair and route shortening. [12]

The IARP should support link state metrics that are consistent with the metrics of IERP. This is required for the IERP to be able to import IARP routes for supporting enhanced route maintenance. [11]

4.2 Example IARP

The specification [11] describes a timer based link state protocol as an implementation example for IARP. A node periodically transmits link state information to the nodes in its routing zone. The link state update packet contains a list of entries consisting of the destination addresses, destination subnet mask and a number of metrics of different type. The scope of the updates is controlled by a time-to-live field initialized to ρ*-1*.

The protocol uses a routing table, where each entry contains the destination address, subnet mask, route list and route metric list. The link state table contains the link source address, zone radius, link state ID, timestamp and a list of destination address, destination subnet mask and link metrics.

4.3 Example IERP

The specification [12] provides a simple implementation of an IERP using source routing. The protocol does not include advanced features such as diversity injection, expanding ring search and route metric collection. These can be added if desired.

When a node has a packet to send and there is no route to the destination, a route request packet is bordercasted using BRP. When a node receives a route request for a destination that it has no route to, it appends its IP address and the link metrics to the request and forwards the request with BRP. On the other hand, if the node has a route to the requested destination, it appends the route to the route in the request, and creates a reply packet with the route. The route reply is forwarded back to the query source along the reversed accumulated route.

IARP notifies IERP when a change in the routing zone is detected, so that IERP can perform route repair and optimization. For each IERP route affected by the change, an alternative path through the routing zone is identified. The new path minimizes the distance to the destination, and can thus bypass failed links and sub-optimal segments.

The packet format is similar for route requests and route replies, with an identifier indicating the type. The packet contains a list of IP addresses built along the path: the query source, a number of intermediate nodes, and the destination. A pointer identifies the next node in the list to forward the packet to. A query ID is used to uniquely identify the request for limiting the propagation. IERP uses a routing table similar to the IARP routing table.

4.4 BRP

The draft [13] describes the operation of BRP. In this case, the document does not specify how precisely an implementation must follow the provided definitions.

The BRP is responsible for forwarding IERP route queries to the peripheral nodes of the bordercasting node. To save network resources, a multicast tree is used. Although the receivers of a bordercast packet are the peripheral nodes, the BRP deliver the query to the IERP at every hop.

The protocol keeps track over the nodes that have been covered by the query. When a node receives a query packet, it marks the interior nodes of the previous bordercaster as covered by reconstructing its bordercast tree. If the receiving node is a peripheral node of the previous bordercaster, then this node becomes a new bordercaster and its interior nodes are marked as covered. Before the query is delivered to higher layers, the state is stored in a cache, so that the query can be properly forwarded when it returns from the higher layer.

When BRP receives a new query to bordercast, it marks the node as the bordercaster and marks the interior nodes as covered, and the query is delivered to the peripheral nodes. On the other hand, when a previously received query returns from higher layers, the protocol determines which branches to be pruned based on the map of covered nodes. The query is delivered to the remaining peripheral nodes, and these nodes are marked as covered. By maintaining a map of covered nodes, BRP can terminate the delivery if it receives the query from another direction.

The BRP packet contains the query source and destination addresses, the query ID and previous bordercaster address. The route request is transported as an encapsulated packet. BRP utilizes the routing table and link state table of IARP. In addition, it uses a cache of detected queries, containing the query source, the query ID, the BRP cache ID and the previous bordercaster. The query coverage map contains a graph for every combination of query source and query ID.

5 Unidirectional links

Most routing protocols assume that the links are bi-directional. However, due to differences in power capabilities, the transmission range of the nodes may differ. Consider a scenario where a node A is communicating with a node B, whose transmission range is smaller than the range of node A. In that case, node A is able to send to node B, but not able to receive from B because of the limited power capability. With the neighbor detection protocol suggested for ZRP, node B sees A as a neighbor, but A cannot see B. A unidirectional link is thus created between the nodes.

ZRP provides local support for unidirectional links. The support is provided by IARP and works only if both the link source and destination are in the same zone. [11]

5.1 Extensions for unidirectional routing

The paper [17] proposes an extension to ZRP for networks with unidirectional links. It provides modified IARP, IERP and query control mechanisms, which work with unidirectional links. It also proposes a mechanism for recursive enhancement of queries for unidirectional links with cycles larger than the zone radius. Bi-directional links are seen as a pair of unidirectional links.

In the modified IARP, nodes regularly send information about their inbound neighbors. The hop-count of these advertisements is limited to the zone radius. The information is used to compute the outbound tree, which is a shortest path three from the central node to the nodes from which advertisements were heard.

The modified IERP is significantly more complex than the basic IERP, and a full description exceeds the limits of this paper. Through bordercasting, the forward path is built up as a list of nodes. Only the border nodes are added to the list, which is sufficient since the topology of the intermediate nodes is known. The reverse path is traversed when the reply is sent back to the originator. Unidirectional links with inclusive cycles smaller than the zone size can be bypassed since the nodes know the topology of their zone.

The query enhancement mechanism in IERP is used for computing routes consisting of unidirectional links with inclusive cycles larger than the zone size. The mechanism computes a set of alternative destinations that are known to have paths to the requested destination if a route is not discovered. The original sender can request this set with an enhanced query, which can be repeated a limited number of times to further enhance the query.

Every node participating in the bordercasting process checks its inbound tree to see if it knows alternative nodes with a path to the desired destination. If such nodes are found, it sends a query enhancement message to the sender. If the query source does not receive a route response message within a defined timeout, it checks for received query enhancement message. If a query enhancement message has been received, a new enhanced query message is sent to the alternative destinations.

The modified IERP uses five messages, as briefly presented in Table 1. IERP utilizes two trees: the bordercast tree and the two-way tree. The bordercast tree is used for sending bordercast messages (i.e. route queries) to a set of nodes. It is a sub-graph of the outbound tree of a node. The two-way tree is used to find alternative destinations, which have routes to the desired destination.

Table 1: Messages of IERP with unidirectional links

Message	Description
Route Query Request (RQRQ)	Basic query for locating a destination.
Query Enhancement Request (QERQ)	Message for requesting nodes to respond if they know alternative nodes with a path to the original destination.
Enhance Route Request (ERRQ)	Similar to QERQ, but can only be enhanced a limited number of times.
Query Response (QR)	Response to a RQRQ or QERQ when a border node knows a path to the destination.
Query Enhancement Response (QER)	Generated if the border node does not have a path to any of the queried destinations but it knows a node with a path to some of the queried destinations.

6 Analysis

The key idea of ZRP is to utilize the features of both proactive and reactive routing. With proactive routing inside a limited zone, the connection establishment time can be reduced. Reactive routing reduces the amount of control traffic by discovering the path on demand for destinations outside the routing zone. The most dominant parameter influencing on the efficiency of ZRP is the zone radius. A few papers [1], [10], [15], [18] have been written that analyze the protocol performance and amount of control traffic as a function of the zone radius.

6.1 Traffic measurements

ZRP control traffic under different query control mechanisms was measured in [15]. The results show that the IARP traffic grows with the number of nodes in the zone, which is proportional to the "area" of the zone, ρ^2. Therefore, the cost of maintaining an extended routing zone (in DB) is high compared to the use of only a normal routing zone (in RDB). Both RDB and DB showed a similar number of packets in IERP route discovery. However, RDB has a higher bit load, since the packets must contain the bordercast tree map.

According to [15], the effects of the query control mechanisms were significant in multiple-channel networks. In multiple-channel networks, a routing zone of radius $\rho=2$ reduces query traffic with 50% compared to flooding ($\rho=1$), whereas the same improvement in single-channel networks were only 15%. If RQPD is employed, the traffic is further reduced by 10%. The improvement rate slows down with increasing radius.

Since the amount of control traffic depends on both node mobility and route query rate, the call-to-mobility ratio (CMR) is useful to characterize the relative traffic amounts. For large values of CMR, where mobility is relatively low, the traffic amount can be reduced with a larger radius. The cost of maintaining proactive information is low relatively to the route discovery traffic. The opposite behavior is seen for low CMR values.

In [18], ZRP was tested in a small network with a few nodes and low traffic amounts. IARP overhead increased rapidly with increasing zone radius. Increasing velocity did not affect the IARP traffic, but the IERP overhead increased due to route repairs. Link stability increased in larger zones, since BRP utilizes local topology information to route around failed links.

6.2 Determining the routing zone radius

With the correct zone size, it is possible to reduce the control traffic to a minimum. Each network configuration has an optimal zone radius value. To determine the optimal value, it is necessary to understand how different factors influence on the traffic amount. According to simulations performed in [10], the main factors are the zone radius ρ, network size N, node density δ (average number of neighbors per node) and average node velocity v (affecting route stability). Of these, only the zone radius is a configurable parameter.

Because of proactive route maintenance, the amount of control traffic from IARP increases with increasing zone radius. Since IARP route updates are a local event, the network size does not affect the amount of proactive traffic. The amount of IERP traffic received by a node is independent of N as well. Instead, an increase in the network size increases the number of route queries. Thus, the amount of reactive route query traffic increases with increasing network size. Therefore, larger zone sizes are favored in large networks. Larger zones provide more efficient queries, which compensates for the higher IARP maintenance costs.

The amount of control traffic largely depends on the relationship between node velocity and route usage. Higher velocity causes a linear increase in IARP routing updates and IERP route failures. If the route usage rate is considerably higher than the route failure rate, route discoveries are driven by route failures, and the traffic amount increases linearly with the node velocity. In contrast, if route usage is smaller than the route failure rate, the route query rate is independent of route stability and node velocity. In this case, the load on IARP increases with the node velocity, and a small routing zone is preferential.

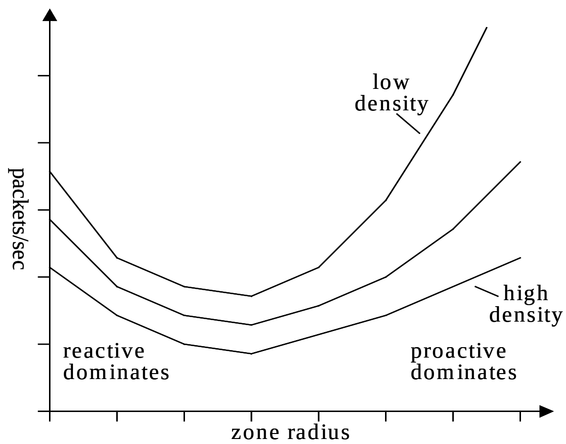

Figure 8: ZRP traffic per node (schematic)

The optimal radius seems to be independent of the node density in most cases. Yet, a large increase in the node density increases the cost of IARP routing zone maintenance, which decreases the optimal routing zone radius.

6.3 Zone sizing schemes

As seen, the optimal routing zone radius depends on a number of factors, which varies for different networks and also varies within a network as a function of time. Even with perfect knowledge of all parameters, computation of the optimal radius is complicated. Even though it is possible to estimate the node density, relative node velocity, network size, the performance also depends on other factors, such as route selection criteria, route caching policies and data traffic behavior. Therefore, the paper [10] proposes two zone sizing schemes.

The “min searching” scheme searches for a local minimum of the total ZRP traffic. The routing zone radius is either incremented or decremented in steps of one. The process is repeated in the same direction as long as the new measured traffic amount is smaller than the previous one. The found minimum is maintained until the process restarts later. The paper also suggests an automatic method for determining the time before the process is restarted. The local minimum that is found is also a global minimum since both the IARP and IERP traffic are convex functions of the zone radius. The problem with this technique is that the estimation interval must be long enough to provide accurate measurements, but a long interval may not provide adequate correlation between consecutive intervals.

The other scheme is based on the relationship between IARP and IERP traffic. When the zone radius is less than the optimal and the ZRP traffic is more than optimal, the traffic is dominated by IERP queries. If the zone radius is larger than the optimal and the traffic is more than optimal, most of the traffic is IARP route updates (see Figure 8). This property is used in the “traffic adaptive” scheme. The ratio of IERP and IARP traffic is compared with a threshold. The zone size is increased if IERP/IARP is larger than the threshold and reduced if less. A hysteresis value is used to improve stability. In this scheme only data collected from one measurement interval is used, which improves performance in frequently changing networks.

An oscillation problem may appear in the “traffic adaptive” scheme if the zone size is small. It is caused by the fact that a zone of radius one is purely reactive. To solve this problem, both the above schemes can be combined. The min searching scheme is then used when the radius is small (one or two hops) and the adaptive scheme is used otherwise.

7 Conclusions

ZRP combines two completely different routing methods into one protocol. Within the routing zone, the proactive component IARP maintains up-to-date routing tables. Routes outside the routing zone are discovered with the reactive component IERP using route requests and replies. By combining bordercasting, query detection and early termination, it is possible to reduce the amount of route query traffic. Since the actual implementation of IARP and IERP is not defined, the performance can be further improved by adapting other routing protocols as ZRP components. ZRP can be regarded as a routing framework rather than as an independent protocol [19].

ZRP reduces the traffic amount compared to pure proactive or reactive routing. Routes to nodes within the zone are immediately available. ZRP is able to identify multiple routes to a destination, which provides increased reliability and performance. It ensures that the routes are free from loops. It is a flat protocol, which reduces congestion and overhead usually related to hierarchical protocols. [10]

The zone routing protocol is targeted for large networks. It differs from cluster based routing protocols because the zones overlap. Because proactive updates are propagated only locally, the amount of control traffic does not depend on network size. The reactive routing is more efficient than flooding since local topology information can be used. Enlarging the zone size reduces the amount of reactive traffic. [19] [10]

The protocol performance can be optimized by adjusting a single parameter, the zone radius. The parameter controls the tradeoff between the cost of the proactive and reactive components, which both are convex functions of the zone radius. The optimal zone radius depends on a number of factors, including node velocity, node density and network span. Since these parameters changes, also the zone radius must be adjusted for optimal performance. Two methods for dynamically adjusting the zone radius have been examined in [10]. The “min searching” scheme keeps the traffic within 7% of the minimum traffic. The “traffic adaptive” scheme performs even better with traffic less than 1-2% than the optimal.

The ZRP is defined in three separate Internet drafts: IARP in [11], IERP in [12] and BRP in [13]. ZRP is one of the protocols that are currently under evaluation and standardization by the IETF MANET working group. Since ZRP is more like a routing framework, it does not directly compete with other routing protocols. [20]

Most evaluations and comparisons (e.g. [21] and [22]) of protocols for ad-hoc networks skip ZRP. The reason is usually that ZRP is aimed for larger networks than the test comprises, or that ZRP is not an independent protocol but rather a routing framework. Further, any evaluation of the ZRP version with support for unidirectional links could not be found. Nevertheless, tests made in [10] verify that ZRP with proper configuration of radius performs more efficiently than traditional routing protocols without need for centralized control. It is especially well adapted to large networks and diverse mobility patterns [23].

Based on the evaluations studied in this paper, we can conclude that ZRP performs better than any single proactive or reactive protocol. This is especially true if we take into account that almost any pure proactive and reactive protocol can be adapted as an IARP or IERP component of ZRP. However, the cost of ZRP is increasing complexity, and in the cases where ZRP performs only slightly better than the pure protocol components, one can speculate whether the cost of added complexity outweigh the performance improvement. Furthermore, new protocols that are neither proactive nor reactive, as well as protocols utilizing geographical information may outperform the ZRP.

References

[1] Haas, Zygmunt J., Pearlman, Marc R.: Providing Ad-hoc connectivity With Reconfigurable Wireless Networks, Ithaca, New York

[2] Elizabeth, Royer, Chai-Keong, Toh: A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks, April 1999, IEEE Personal Communications

[3] Moy, J.: OSPF Version 2, IETF RFC 2178, July 1997

[4] Perkins, C. E., Bhagwat, P.: Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers, October 1994, Computer Communications, pp. 234-244

[5] Clausen, T., Jacquet, P., Laouiti, A., Minet, P., Muhlethaler, P., Qayyum, A., Viennot, L.: "Optimized Link State Routing Protocol", September 2001, IETF Internet Draft, draft-ietf-manet-olsr-06.txt

[6] Murthy, S., Garcia-Luna-Aceves, J. J.: An Efficient Routing Protocol for Wireless Networks, October 1996, MONET, Vol 1, No 2, pp. 183-197

[7] Park, V. D., Corson, M. S.: A Highly Adaptive Distributed Routing Algorithm for Mobile Wireless Networks, April 1997, Proc. IEEE INFOCOM '97, Kobe, Japan

[8] Johnson, D. B., Maltz, D. A.: Dynamic Source Routing in Ad-Hoc Wireless Networking, Mobile Computing, T. Imielinski and H. Korth, Eds. Norwell, MA: Kluwer, 1996, pp. 153-181

[9] Perkins, C. E., Royer, E. M.: Ad-hoc On-Demand Distance Vector Routing, February 1999, Proc. 2nd IEEE Workshop on Mobile Computer Systems and Applications, pp. 90-100

[10] Pearlman, Marc R., Haas, Zygmunt J.: Determining the Optimal Configuration for the Zone Routing Protocol, August 1999, IEEE Journal on Selected Areas in Communications, Vol. 17, No. 8

[11] Haas, Zygmunt J., Pearlman, Marc R., Samar, P.: Intrazone Routing Protocol (IARP), June 2001, IETF Internet Draft, draft-ietf-manet-iarp-01.txt

[12] Haas, Zygmunt J., Pearlman, Marc R., Samar, P.: Interzone Routing Protocol (IERP), June 2001, IETF Internet Draft, draft-ietf-manet-ierp-01.txt

[13] Haas, Zygmunt J., Pearlman, Marc R., Samar, P.: The Bordercast Resolution Protocol (BRP) for Ad Hoc Networks, June 2001, IETF Internet Draft, draft-ietf-manet-brp-01.txt

[14] Haas, Zygmunt J., Pearlman, Marc R.: Providing Ad-hoc Connectivity With Reconfigurable Wireless Networks, Ihaca, New York, http://www.ee.cornell.edu/~haas/wnl.html

[15] Haas, Zygmunt J., Pearlman, Marc R.: The Performance of Query Control Schemes for the Zone Routing Protocol, August 2001, IEEE/ACM Transactions on Networking, Vol. 9, No. 4

[16] Haas, Zygmunt J., Jing Deng, Ben, Liang, Panagiotis Papadimitrator, S., Sajama, Wireless Ad Hoc Networks, Ithaca, New York, 2001

[17] Prasun, Sinha, Srikanth, Krishnamurthy, Son, Dao: Scalable Unidirectional Routing with Zone Routing Protocol (ZRP) Extensions for Mobile Ad-Hoc Networks

[18] Shea, Kevin, Ives, Robert W., Tummala, Murali: Mobile Ad Hoc Network Routing Protocol Analysis and Its Application to a Programmable Modular Communication System

[19] Samir, Das, Robert, Castañeda, Jiangtao Yan: Simulation-based performance evaluation of routing protocols for mobile ad hoc networks, Mobile Networks and Applications 5 (2000) 179-189, http://www.cc.jyu.fi/~ngerman/HANNIBAL/NNIB/Papers/TP8.pdf

[20] Joseph, Macker, Scott, Corson: Mobile Ad-hoc Networks (MANET), IETF Working Group Charter,

http://www.ietf.org/html.charters/manet-charter.html

[21] Frank, McSherry, Gerome, Miklau, Don, Patterson, Steve, Swanson: The Performance of Ad Hoc Networking Protocols in Highly Mobile Environments, Spring 2000, Washington

[22] Samir, Das, Robert, Castañeda, Jiangtao Yan, Rimli, Sengupta: Comparative Performance Evaluation of Routing Protocols for Mobile, Ad hoc Networks, http://www.cs.washington.edu/homes/rimli/papers/manet.pdf

[23] DeHoust, Matt: Routing in Mobile Ad Hoc Netwoks, April 2000, Individual Project Report