Determining the Optimal Configuration for the Zone Routing Protocol

Zygmunt J. Haas, *Senior Member, IEEE,* and Marc R. Pearlman, *Student Member, IEEE*

***Abstract*-- The Zone Routing Protocol (ZRP) is a hybrid routing protocol that proactively maintains routes within a local region of the network (which we refer to as the routing zone). Knowledge of this routing zone topology is leveraged by the ZRP, to improve the efficiency of a reactive route query/reply mechanism. The ZRP can be configured for a particular network through adjustment of a single parameter, the routing zone radius.**

In this paper, we address the issue of configuring the ZRP to provide the best performance for a particular network, at any time. Previous work has demonstrated that an optimally configured ZRP operates at least as efficiently as traditional reactive flood-search or proactive distance vector / link state routing protocols (and in many cases, much more efficiently).

Adaptation of the ZRP to changing network conditions requires both an understanding of how the ZRP reacts to changes in network behavior and a mechanism to allow individual nodes to identify these changes, given only limited knowledge of the network behavior. In the first half of this paper, we demonstrate the effects of relative node velocity, node density, network span, and user data activity on the performance of the ZRP. We then introduce two different schemes ("min-searching" and "traffic adaptive") that allow individual nodes to identify and appropriately react to changes in network configuration, based *only* on information derived from the amount of received ZRP traffic. Through test-bed simulation, we demonstrate that these radius estimation techniques can allow the ZRP to operate within 2% of the control traffic resulting from perfect radius estimation.

***Index Terms*-- ad-hoc network, routing protocol, routing zone, ZRP, zone routing protocol, proactive routing, reactive routing, hybrid routing, bordercast**

I. INTRODUCTION

A. A Brief Overview of Ad-Hoc Networks

Ad-hoc networks are self-organizing wireless networks composed of mobile nodes and requiring no fixed infrastructure. The limitations on power consumption imposed by portable wireless radios result in a node transmission range that is typically small, relative to the span of the network. To provide communication throughout the entire network, each node is also designed to serve as a relay. The result is a distributed mutli-hop network with a time-varying topology.

Because ad-hoc networks do not rely on existing infrastructure and are self-organizing, they can be rapidly deployed to provide robust communication in a variety of hostile environments. This makes ad-hoc networks very appropriate for providing tactical communication for military, law enforcement and emergency response efforts. Ad-hoc networks can also play a role in civilian forums such as electronic classrooms, convention centers and construction sites. With such a broad scope of applications, it is not difficult to envision ad-hoc networks operating over a wide range of coverage areas, node densities and node velocities.

In order to provide decentralized and tetherless communication, ad-hoc networks need to overcome the limitations of portable wireless communication. The unguided wireless medium and surrounding physical environment significantly attenuate and distort radio transmissions, resulting in relatively unreliable communication channels[1]. Because signal quality rapidly degrades with distance, the effective transmission area of a node is limited, profoundly affecting the way that the wireless medium can be shared. The limited transmission range does benefit the system by allowing channels to be spatially reused. However, because interfering signals at a receiver may be considerably attenuated at a transmitting node, traditional carrier sensing techniques cannot be used to avoid collisions. This *hidden terminal problem* [16] can significantly reduce the amount of traffic carried by the system. Some of this lost capacity can be regained through special media access control mechanisms that allow receivers to control access to the channel (often in the form of hand-shaking packet exchanges and/or "busy tone" signaling). However, satisfactory throughput requires that nodes consume bandwidth judiciously. Further constraints are introduced by the communication devices themselves, which are equipped with lightweight batteries to support portability. In order to provide sufficient battery life, power must be conserved, limiting the transmission range, data rate, communication activity (both transmitting and receiving) and processing speed of these devices.

This work is supported by the US Air Force/Rome Labs, under the contract number C-7-2544 and a grant from Motorola Corporation, the Applied Research Laboratory.
The authors are with the Wireless Networks Laboratory (WNL), School of Electrical Engineering, Cornell University, Ithaca, NY 14853-3801, USA.

[1] Compared to wired channels operating with the same transmission power and bandwidth constraints.

The potential for large-scale ad-hoc networking applications calls for a special class of ad-hoc networks, which we refer to as *Reconfigurable Wireless Networks (RWNs).* RWNs may span over a wide geographic area and consist of many (perhaps hundreds or thousands) of nodes. The nodes can exist on top of a variety of platforms (i.e. pedestrians, tanks, planes, etc.) and exhibit a wide range of speeds and mobility patterns.

This wide range of RWN operating configurations poses a challenge for developing efficient routing protocols. On one hand, the effectiveness of a routing protocol increases as network topology information becomes more detailed and up-to-date. On the other hand, in a RWN, the topology may change quite often, requiring large and frequent exchanges of data among the network nodes. This is in contradiction with the fact that all updates in the wireless communication environment travel over the air and are costly in resources.

B. Routing Protocols -- A Short Survey

In the past, routing in multi-hop packet radio networks was based on shortest-path routing algorithms, such as Distributed Bellman-Ford (DBF) algorithm [1]. These algorithms suffer from very slow convergence (the "counting to infinity" problem). Besides, DBF-like algorithms incur large update message penalties. Protocols that attempted to cure some of the shortcomings of DBF, such as Destination-Sequenced Distance-Vector Routing (DSDV) [13], were proposed. However, synchronization and extra processing overhead are common in these protocols. Other protocols that rely on the information from the predecessor of the shortest path solve the slow convergence problem of DBF (e.g., [2] and [3]). However, the processing requirements of these protocols may be quite high, because of the way they process the update messages.

In *wired* networks, the problem of routing convergence has been addressed by link-state protocols, particularly the Open Shortest Path First (OSPF) protocol [9]. While link state protocols converge more rapidly than distance vector protocols, they do so at the expense of significantly more control traffic. For networks, like the RWN, which experience frequent changes in network topology, the increase in control traffic overhead can overwhelm the network's resources. The recently proposed Optimized Link State Protocol (OLSR) [7] utilizes a multicast-like mechanism (called "multi-point relay") to reduce the amount of traffic produced by the periodic topology updates. This has the potential for performing well on smaller ad-hoc networks. However, the underlying mechanisms of periodic and global topology updates does not appear to scale up to the larger, more dynamic RWNs.

Motivation to both improve protocol convergence and reduce traffic has led to the development of proactive path finding algorithms, which combine features of the distance vector and link state approaches. Each node constructs a minimum spanning tree based on knowledge of its neighbors' minimum spanning trees and the link costs to each neighbor. Realizations of the path finding algorithms, like the Wireless Routing Protocol (WRP) ([10] and [11]), are able to eliminate the "counting-to-infinity" problem and reduce the occurrence of temporary loops, often with less control traffic than traditional distance vector schemes. The main disadvantage of WRP is in the fact that routing nodes constantly maintain full routing information in each network node, which was obtained at relatively high cost in wireless resources.

Routing protocols that are based on a source initiated query-reply process have also been introduced. Such techniques typically rely on the *flooding* of queries to discover a destination. In the Temporally Ordered Routing Algorithm (TORA) [12] the resulting route replies are also flooded, in a controlled manner, to distribute routes in the form of directed acyclic graphs (DAGs) rooted at the destination. In contrast, protocols such as Dynamic Source Routing (DSR) [8] and Ad-Hoc On Demand Distance Vector (AODV) [14] unicast the route reply back to the querying source, along a path specified by a sequence of node addresses accumulated during the route query phase. In the case of DSR, the node addresses are accumulated in the query packet and are returned to the source, to be used for source routing. AODV, on the other hand, distributes the discovered routes in the form of next-hop information stored at each node in the route. The route accumulation process can be extended to address quality of service (QoS) routing requirements, by recording various quality measurements. For example, in Associativity Based Routing (ABR) [17], link stability is recorded for each link that a query packet traverses. The link stability measurements provide a basis for determining the "best" route to return back to the source.

The on-demand discovery of routes can result in much less traffic than standard distance vector or link state schemes, especially when innovative route maintenance schemes are employed. However, the reliance on packet flooding may still lead to considerable control traffic in the highly versatile RWN environment.

A routing algorithm for ad-hoc networks in which each node belongs to two networks: a physical and a virtual network, is presented in [15]. Routing is based on temporary addresses. A temporary address is a concatenation of the node's address on each one of the two networks. Upon physical migration, a node is required to acquire a new temporary address. In order to communicate with a node, a query phase is initiated by the source, in which the nodes that belong to the source's physical and virtual networks are polled about the address of the destination.

The virtual network routing is an interesting idea. However, the dynamic assignment of *unique* node addresses can be quite challenging in an ad-hoc network. In particular, duplicate addresses may arise, even if the address assignment is centrally controlled within each physical subnet (i.e. if a physical subnet becomes partitioned). Furthermore, the routing can be far from optimal, as it is based on hopping within virtual networks, which are determined by the sources and the destination addresses and not by the nodes' geographical locations.

Another interesting approach for routing in the ad-hoc network is the Landmark Hierarchy [18]. Each router in the network is treated as a *landmark* of level j, meaning that it is "visible" to all nodes that are within r_j hops ($r_j < r_{j+1}$). Landmarks are dynamically assigned such that a landmark of level j is always within r_j hops of at least one level j+I landmark. Furthermore, a network must have at least one highest-level landmark whose radius, r_H is at least as large as the network diameter (i.e. visible by all network nodes). Nodes are dynamically addressed based on a sequence of landmark IDs which represent its location in this landmark hierarchy.

Because a level j landmark may be associated with multiple level j+1 landmarks, rather than a single gateway / clusterhead, the landmark hierarchy is more robust than traditional area hierarchies. However, the landmark hierarchy still suffers from the basic inefficiencies typical of all hierarchical protocols. First, extra overhead is required to determine landmark assignment (this is similar to the overhead associated with gateway assignment in area hierarchies). This is especially problematic for RWNs, where frequent topology changes can result in a constant reorganization of the network-wide hierarchy. Secondly, the imposed hierarchy can prevent direct communication between two nodes capable of establishing a link. In many cases, this constraint leads to sub-optimal routes. Finally, the hierarchy can also result in network congestion, as a smaller number of higher level landmarks become the target of cross network traffic.

C. Reactive vs. Proactive Routing

The existing routing protocols can be classified either as *proactive* or as *reactive*. Proactive protocols attempt to continuously determine the network connectivity, so that the route is already available when a packet needs to be forwarded. The families of traditional distance-vector and link state protocols are an example of a proactive scheme. Reactive protocols, on the other hand, invoke a route determination procedure only on demand. Thus, when a route is needed, some sort of global search procedure is employed. The classical flood search algorithms are reactive protocols.

The advantage of the proactive schemes is that, once a route is needed, there is little delay until the route is determined. In reactive protocols, because route information may not be available at the time a route request is received, the delay to determine a route can be quite significant. Furthermore, the global search procedure of the reactive protocols requires significant control traffic. Because of this long delay and excessive control traffic, pure reactive routing protocols may not be applicable to real-time communication. However, purely proactive schemes are likewise not appropriate for the RWN environment, as they continuously use a large portion of the network capacity to keep the routing information current. Since nodes move quite fast in a RWN, and as the changes may be more frequent than the route requests, most of this routing information is never even used! This results further waste of the network capacity.

What is needed is a protocol that, on one hand, initiates the route-determination procedure on-demand, but at limited search cost. The Zone Routing Protocol (ZRP) provides efficient and fast discovery of routes, by integrating the two radically different classes of traditional routing protocols.

II. The Zone Routing Protocol

The *Zone Routing Protocol (ZRP)* [4], [5], [6], is an example of a hybrid reactive / proactive routing protocol. On one hand, it limits the scope of the proactive procedure only to the node's <u>local neighborhood.</u> The local routing information is referred to quite often in the operation of the ZRP, minimizing the waste associated with the purely proactive schemes. On the other hand, the search throughout the network, although it is global, is performed by efficiently querying selected nodes in the network, as opposed to querying all the network nodes.

The protocol identifies *multiple* loop-free routes to the destination, increasing reliability and performance. Routing is flat, rather than hierarchical, reducing organizational overhead, allowing optimal routes to be discovered and reducing the threat of network congestion. However, the most appealing feature of the protocol is that its behavior is adaptive, based on the current configuration of the network and the behavior of the users.

A. The Notion of a Routing Zone and Intrazone Routing

A *routing zone* (of radius ρ) is defined for *each* node and includes the nodes whose <u>minimum</u> distance <u>in hops</u> from the node in question is <u>at most</u> ρ hops. An example of a routing zone (for node *S*) of radius 2 hops is shown in Figure 1. For the purpose of illustration, we depict zones as circles around the node in question. However, one should keep in mind that the zone is not a description of physical distance, but rather nodal connectivity (hops).

Note that in this example, nodes A through K are within the routing zone of the *central node* S. Node *L* is outside *S*'s routing zone. *Peripheral nodes* are nodes whose minimum distance to the node in question is equal <u>exactly</u> to the zone radius. The remaining nodes are categorized as *interior nodes*. Thus, in Figure 1, nodes A-F are interior nodes, while G-K are peripheral nodes. Because each node maintains its own routing zone, the zones of neighboring nodes can heavily overlap.

Figure 1: A Routing Zone of Radius 2 Hops

For a routing zone of radius ρ, the number of routing zone nodes can be regulated through adjustments in each node's transmitter power. Subject to the local propagation conditions and receiver sensitivity, the transmission power determines the set of neighbor nodes, i.e. those nodes that are in direct communication with a node. To provide adequate network reachability, it is important that a node be connected to a sufficient number of neighbors. However, more is not necessarily, better. As the transmitters' coverage areas grow larger, so do the membership of the routing zones. This can result in an excessive amount of route update traffic. (Additionally, larger transmitter coverage leads to more neighbors and increased contention (locally) for the wireless channel.)

Each node is assumed to maintain routing information only to those nodes that are within its routing zone. Because the updates are only propagated locally, the amount of update traffic required to maintain a routing zone does not depend on the total number of network nodes (which can be quite large). We assume that a node learns its zone through some sort of a proactive scheme, which we refer to here as the *IntrAzone Routing Protocol (IARP).* In this paper, we use a split-horizon version of the Distance Vector algorithm. However, any other proactive scheme would do. While the performance of the ZRP depends on the choice of IARP implementation, our experience suggests that the tradeoffs are not strongly affected by the particular choice of the proactive scheme used.

B. Interzone Routing and the Zone Routing Protocol

The IARP maintains routes only for those nodes that are within the coverage of the routing zone. For RWNs, the coverage of a routing zone is relatively small compared to the size of the network. Thus, most destinations lie outside of a node's routing zone, and the desired routing information cannot be immediately provided by the IARP. As we shall see, the real benefit of the IARP is realized when the topology of a node's routing zone can be indirectly leveraged to satisfy the entire network's demand for routes.

The IntErzone Routing Protocol (IERP) is responsible for reactively discovering routes to destinations located beyond a node's routing zone. The IERP is distinguished from standard flooding-based query/response protocols by exploiting the structure of the routing zone. The routing zones increase the probability that a node can respond positively to a route query. This is beneficial for traffic that is destined for geographically close nodes. More importantly, knowledge of the routing zone topology allows a node to efficiently continue propagation of a query in the more likely case that destination cannot be found. This is achieved by a packet delivery service, called bordercasting, that allows a node to direct a message to its peripheral nodes. In its simplest form, bordercasting could be implemented through network layer unicasting or multicasting of messages to the peripheral nodes. This approach prevents non-peripheral nodes from accessing the bordercasted messages as they are relayed to the edge of the routing zone. As will be shown later, such access is central to the control of the route query process. As such, a more suitable implementation of bordercasting indirectly sends messages to peripheral nodes by forwarding them between adjacent nodes.

The IERP operates as follows: The source node first checks whether the destination is within its zone.[2] If so, the path to the destination is known and no further route discovery processing is required. If the destination is not within the source Routing Zone, the source bordercasts a *route request* (which we call simply a *request)* to all its peripheral nodes.[3] Now, in turn, all the peripheral nodes execute the same algorithm: check whether the destination is within their zone. If so, a *route reply* (which we call simply a *reply)* is sent back to the source indicating the route to the destination (more about this in a moment). If not, the peripheral node forwards the query to its peripheral nodes, which, in turn, execute the same procedure.

Figure 2: An Example of IERP Operation

An example of this *Route Discovery* procedure is demonstrated in Figure 2. The source node *S* needs to send

[2] Remember that a node knows the identity, distance to, and a route to all the nodes in its zone.

[3] Again, the identity of its zone peripheral nodes are known to the node in question.

a packet to the destination *D.* To find a route within the network, *S* first checks whether *D* is within its routing zone. If so, *S* knows the route to *D.* Otherwise, *S* bordercasts a query to its peripheral nodes; that is, S sends a query to nodes *C, G,* and *H.* Now, in turn, each one of these nodes, after verifying that *D* is not in its routing zone, forwards the query by bordercasting the query to its peripheral nodes. In particular, *H* sends the query to *B,* which recognizes *D* as being in its routing zone and responds to the query, indicating the forwarding path: *S-H-B-D.*

As indicated by this example, a route can be specified by a sequence of nodes that have received the successful IERP query thread. The manner in which this information is collected and distributed is specified by a *Route Accumulation* procedure. In the basic Route Accumulation, a node appends its ID to a received query packet. When a node finds the destination in its zone, the accumulated sequence of IDs specifies a route between querying source and destination. By reversing the accumulated route, a path is provided back to the query source. This information can be used to return the route reply, through source routing.

Given sufficient <u>short-term</u> storage at each node, a route can be temporarily stored by the queried nodes, in the form of next-hop routes back to the queried source. Rather than append its ID to an accumulated route, a node would write its ID, specifying it as the most recently queried node, and overwriting the ID of the previously queried node. A node receiving this query would know that the source could be reached through the most recently queried node, and record this information in a temporary routing table. In this case, route accumulation would occur during the route reply phase, rather than the route query phase, resulting in less IERP traffic.

A nice feature of this distributed route discovery process is that a single route query can return multiple route replies. The quality of these returned routes can be evaluated based on hop count (or any other path metric[4] accumulated during the propagation of the query. The best route can be selected based on the relative quality of the route (i.e. choose the route with the smallest hop count, or shortest accumulated delay).

The intuition behind the ZRP is that querying can be performed more efficiently by bordercasting queries to the periphery of a routing zone, rather than flooding the queries over the same area. However, problems can arise once the query leaves the initial routing zone. Because the routing zones heavily overlap, a node can be a member of many routing zones. It is very possible that the query will be forwarded to all the network nodes, effectively flooding the network. But a more disappointing result is that, due to fact that bordercasting involves sending the query along a path equal to the zone radius, the IERP will result in much more traffic than the flooding itself!

[4] Typical path metrics include hop count, delay, capacity, etc.

In order to understand the cause of the ZRP control traffic problem, it is important to stress one of the key features of the routing zone: A node's response to a route query contains information about that node's entire routing zone. From this perspective, excess route query traffic can be regarded as a result of overlapping query threads (i.e. overlapping queried routing zones). Thus, the design objective of query control mechanisms should be to reduce the amount of route query traffic by steering threads outward from the source's routing zone and away from each other (see Figure 3). This problem is addressed primarily through appropriate mechanisms of query detection and query termination.

Figure 3: Guiding the Search in Desirable Directions

When the ability to terminate threads is limited to peripheral nodes, threads are allowed to penetrate *into* previously covered areas, generating unnecessary control traffic. This excess traffic can be eliminated by extending the thread termination capability to the intermediate nodes that relay the thread. We refer to this approach as Early Termination (ET). Figure 4 illustrates the operation of the ET mechanism. Node S bordercasts a route query, with node C as one of the intended recipients. Intermediate node A passes along the query to B. Instead of delivering the query to node C, node B terminates the thread because a different thread of this query was previously detected. Intermediate nodes may terminate existing queries, but are restricted from issuing new queries. Otherwise, the ZRP would degenerate into a flooding protocol.

Figure 4: Early Termination (ET)

The ability to terminate an overlapping query thread depends on the ability of nodes to detect that a routing zone they belong to has been previously queried[5]. Clearly, the

[5] The ID of the node that bordercasted the first detected query thread is also recorded. In order to ensure full network coverage for that query,

central node in the routing zone (which processed the query) is aware that its zone has been queried. In order to notify the remaining routing zone nodes, without introducing additional control traffic, some form of "eavesdropping" needs to be implemented. The first level of Query Detection (QD1) allows the intermediate nodes, which transport queries to the edge of the routing zone, to detect these queries. In single channel networks, it may be possible for queries to be detected by any node within the range of a query-transmitting node. This extended query detection capability (QD2) can be implemented by using IP broadcasts to send route queries[6]. Figure 5 illustrates both levels of advanced query detection. In this example, node S bordercasts to two peripheral nodes, B and D. The intermediate nodes A and C are able to detect passing threads using QD1. If QD2 is implemented, node E will be able to "eavesdrop" on A's transmissions and record the query as well.

Figure 5: Query Detection (QD1/QD2)

The techniques just discussed improve the efficiency of the IERP by significantly reducing the cost of propagating a single query. Further improvements in IERP performance can be achieved by reducing the frequency of route queries, initiating a Route Discovery procedure only when there is a substantial change in the network topology. More specifically, active routes are cached by nodes: the communicating end nodes and intermediate nodes. Upon a change in the network topology, such that a link within an *active* path is broken, a <u>local</u> path repair procedure is initiated. The path repair procedure substitutes a broken link by a mini-path between the ends of the broken link. A path update is then generated and sent to the end points of the path. Path repair procedures tend to reduce the path optimality (e.g., increase the length for shortest path routing). Thus, after some number of repairs, the path end points may initiate a new Route Discovery procedure to replace the path with a new optimal one.

III. EVALUATION OF THE ZRP

We use the OPNET™ network simulator from MIL3, an event driven simulation package, to evaluate the performance of the ZRP over a variety of RWNs. Each RWN is characterized by the number of nodes (N), node density (δ) and relative node velocity (v). For each RWN scenario, the ZRP is evaluated over a range of routing zone radii, ranging from reactive routing ($\rho = 1$) to proactive routing ($\rho \to \infty$). Performance is gauged by measuring the control traffic generated by the ZRP. Our results demonstrate the dependence of the optimum ZRP routing zone radius on network configuration and node behavior.

We characterize the RWN based on the *nodal* perception of the network. For instance, average node density is expressed as the average number of *neighbors per node*, rather than a physical measure of *nodes per unit area*. It is the former measure that <u>directly</u> influences the behavior of routing protocols. The relationship between physical node density and perceived node density is actually a complicated function of factors like node transmission power, node activity, fading conditions, receiver sensitivity, etc. Likewise, relative node velocity, as perceived by the ZRP, can be expressed in terms of rate of new neighbor acquisition, rather than the physical measure of *distance traveled / unit time.* These two measures provide the information needed by a routing protocol to determine how connected the network is, and how often the connectivity changes. From now on, we refer to δ as the average number of neighbors and v as the rate of new neighbor acquisition.

Control traffic is viewed as the sum of the IARP route update packets and the transmissions of IERP request / reply / failure packets. While the neighbor discovery (HELLO) beacons could be considered control overhead, this additional traffic is independent of both mobile velocity and routing zone radius. Furthermore, the neighbor discovery process is not an exclusive component of the ZRP; various MAC protocols are also based on neighbor discovery. As such, the beacons do not contribute to the relative performance of the ZRP and are not accounted for in our analysis.

IARP route updates are triggered based on the change in connectivity with a neighbor. When a neighbor is discovered, the neighbor is sent a portion of the IARP routing table, and the routing zone is updated based on the new connection. When the neighbor is lost, the routing zone is updated about the lost connection. While a node is directly connected to a neighbor, we assume that the link cost remains constant. Thus, the rate of IARP traffic can be expressed as:

$$IARP\ traffic\ /\ node\ =\ v \cdot IARP\ updates\ /\ neighbor(\rho, \delta)$$

Note that the amount of IARP traffic per node does not depend on the total node population, N. This is due to the local scope of the routing zone.

The rate of IERP traffic can be regarded as the amount of traffic generated per query, multiplied by the rate at which queries are initiated. We note that the query control mechanisms described in Section II are able to significantly

future threads received from that bordercasting node are <u>not</u> automatically discarded.

[6] Alternatively, IP can unicast the queries if the MAC and IP layers are permitted to operate in promiscuous mode.

limit the occurrence of overlapping queries. The ability of these mechanisms to steer queries outwards makes the amount of traffic that a node receives (per query) essentially independent of the network size. Assuming that the network topology does not change significantly during the propagation of a route query, the amount of received traffic per query is also independent of relative node velocity.

The rate at which a node initiates queries depends on the user's demand for a route, as well as the discovered route's stability. The average initial demand and subsequent use of a route is usually application specific, and can be expressed by independent parameters: $\mathrm{R}_{\mathrm{initial\_query}}$ and $\mathrm{R}_{\mathrm{route\_usage}}$. Route stability, on the other hand, depends on the span of the network (reflecting average route length), as well as node density, node velocity and zone radius, which affect the stability of a routing zone and the stability of individual connections in the source route. We can express route stability in terms of average route lifetime, or its inverse, the average route failure rate ($\mathrm{R}_{\mathrm{route\_failure}}$).

The amount of IERP traffic per node can be expressed as:

$$IERP\ traffic\ /\ node$$

$$= \Xi(\rho,\delta)\cdot rate\ of\ received\ queries(...)$$

$$= \Xi(\rho,\delta)\cdot N\cdot(R_{initial\_query} + R_{subsequent\_queries}(...))$$

where:

– $\Xi(\rho,\delta) \equiv traffic\ /\ query\ /\ node(\rho,\delta)$

$$\mathrm{R}_{\mathrm{subsequent\_queries}} \approx \min(\mathrm{R}_{\mathrm{route\_failure}}(N,v,\rho,\delta),\ \mathrm{R}_{\mathrm{route\_usage}})$$

– $\mathrm{R}_{\mathrm{subsequent\_queries}}$, the rate at which routes need to be re-discovered, is approximated by:

This expression assumes that a route query is initiated upon the failure of one route, implying that a node only caches a single route, even though the query process returns multiple routes. With multiple cached routes, the rate of subsequent queries would decrease. This expression also does not take local route repair into consideration. Like multiple route caching, local route repair can reduce the rate that full-depth route queries are initiated. Analysis of the many possible policies for route caching and route maintenance is outside the scope of this paper. However, our research has shown that this upper bound provides a reasonably good indication of the basic relationships between network parameters and protocol performance.

The expression also reveals an interesting dependence on higher layer behavior. Route queries are triggered either by user demand for a unrecorded routes, or by use of a route that is no longer valid. When the route usage rate is high (relative to the route failure rate), a route can be used many times before a new route query needs to be initiated. Such a scenario might occur for applications such as streamed video or large file downloads, where there is a short packet interarrival delay. On the other hand, when route demand is more intermittent, as might be the case for web browsing or telemetry applications, a route failure is likely to occur between successive uses of a route. In this case, nearly all route accesses can lead to a new route query.

If routes are used more often than they fail, the querying rate of the network is determined by the network configuration. On the other hand, if routes fail more often than they are used, the querying rate becomes driven by the behavior of the user's application. The performance analysis will compare the overall performance of the ZRP under both conditions.

Each node, j, within the network moves at a constant speed, *v*, and is assigned an initial direction[7], θ, which is uniformly distributed between 0 and $2\cdot\pi$ [radians]. When a node reaches the edge of the simulation region, it is reflected back into the coverage area, by setting its direction to $-\theta$ (horizontal edges) or $\pi-\theta$ (vertical edges). The magnitude of the velocity is not altered.

Each simulation runs for duration of 125 seconds. No data is collected for the first 5 seconds of the simulation to avoid measurements during the transient period and to ensure that the initial intrazone route discovery process stabilizes.

For the purposes of our simulation, we have made a number of simplifying assumptions regarding the behavior of the lower network layers and channel. This simplified model helps to improve understanding of our routing protocol behavior by providing our performance measures with some immunity from lower layer effects.

In our model, neighbor discovery is based on the reception of HELLO beacons that are broadcast at the MAC layer. These short beacons (containing only source address) are transmitted at random intervals of mean $T_{beacon} = 0.25$ seconds. Neighbor connectivity is determined by the reception of the HELLO beacons. If a new beacon fails to arrive within $2\cdot T_{beacon}$ of the most recent beacon, a link failure is reported. Because the links are bi-directional, the need for a more complex HELLO $\rightarrow$ I-HEAR-YOU packet exchange is eliminated. Furthermore, we assume that neighbor discovery are given highest transmission priority and are not destroyed by collisions. This prevents the inaccurate reporting of link failures for the allowed $2\cdot T_{beacon}$ window.

The MAC protocol itself provides "ideal" scheduling of packet transmissions to avoid collisions. Although, such a scheme is not possible in practice (especially in a distributed ad-hoc environment), we have found that the control traffic performance of our routing protocol depends very little on the underlying MAC protocol. By reducing

[7] Direction is measured as an angle relative to the positive x-axis.

the complexity at the MAC layer, simulation of large, dense networks of highly mobile nodes becomes feasible. In addition, use of an "ideal" MAC allows us to isolate the delays associated with a particular MAC scheme (e.g., collision avoidance algorithms) from the delays related to the routing protocol.

Our assumption of a collision-free media access protocol means that the average SIR of a received packet is limited by the ambient background noise and receiver noise. For fixed transmitter and noise powers, we find that the BER is reasonably low within a distance, which we call d_{xmit}. Beyond d_{xmit}, the BER increases rapidly. This behavior results from a rapid decrease in received power as the separation distance is increased. We approximate this rapid increase in BER by the following simplified path loss model:

We interpret this behavior as follows: any packet can be received, error-free, within a radius of d_{xmit} from the transmitter, but is lost beyond d_{xmit}. Note that this implies bi-directional links. Since packet delivery is guaranteed to any destination in range of the source, we are able to further reduce the complexity of our model by eliminating packet retransmission at the data link level.

$$PL(d) = \begin{cases} 0[dB] & \text{for } d \le d_{xmit} \\ \infty[dB] & \text{for } d > d_{xmit} \end{cases}$$

Variable Simulation Parameters

Parameter	Symbol	Values	Default[8]
Zone Radius *[hops]*	ρ	1-8	--
Node Density *[neighbors/node]*	δ	3 – 9	6
Rel. node speed *[neighbors/sec]*	*V*	0.1 – 2.0	1.0
# of Nodes *[nodes]*	*N*	200 – 1000	500

IV. ZRP PERFORMANCE RESULTS

A. IARP Control Traffic

We begin by examining the performance of the proactive IARP. The IARP route update process consists of notifying a routing zone population about a change in zone topology, and exchanging routing tables with newly discovered neighbors. Both of these functions are $O(N_{zone})$, where N_{zone} is the zone population. Therefore, we expect that the amount of IARP traffic should be proportional to the node density (δ), and the routing zone area (ρ^2). This behavior is exhibited in Figure 6. Note that the amount of traffic triggered per neighbor does not depend on the velocity. However, the overall rate of IARP traffic does, and can be determined by multiplying this per-neighbor-traffic by the rate of new neighbor discovery (which we consider as velocity, v).

Figure 6

B. IERP Control Traffic: Traffic per Query

Having examined the IARP, we now focus on the behavior of the reactive IERP. Figure 7 demonstrates how the amount of query related traffic depends on the routing zone and density. Recall that when $\rho = 1$, the query process is effectively a flood search. In the case of flooding, we expect each network node to receive the query, and pass it along once to each of its neighbors. Therefore, each node should receive approximately δ packets per query since each node has δ neighbors[9]. For high densities, we find that this is the case. However, when there are fewer than six neighbors per node, we note that the average traffic is less than expected. The disparity becomes greater as δ becomes smaller. This behavior can be easily explained. In order for the network to be un-partitioned (with high probability), each node must be able to be *directly* reached by a sufficient number of neighbors. In the absence of partitioning, each flooded query would be received by all nodes. According to our data, six or more nodes can provide reasonable network reachability. Three neighbors, on the other hand, is not enough to provide sufficient reachability. In this case, many queries are not received, due to partitioning. The average number of packets per query is thus noticeably less than δ. For the remainder of our analysis, we will ignore densities of fewer than five neighbors.

Figure 7

[8] For the results in Section IV, if a simulation parameter is not explicitly stated, the default value is implied.

[9] More specifically, δ query packets plus the average number of reply packets that are forwarded back to the source.

As previous work has shown, the IERP traffic per query decreases with the zone radius. This is due to the combination of bordercasting and query control. Bordercasting allows queries to be directed to the edge of a routing zone, reducing unnecessary queries within a routing zone. Query detection and termination mechanisms complement bordercasting by avoiding redundant queries between routing zones. For a given zone radius, we observe that the amount of received traffic per query increases with the zone density. As routing zones become denser, the number of peripheral nodes (and bordercast messages) increases. Since these messages are sent out individually, it takes more time for nodes to detect a query within their zone. This increases the probability that a node will unknowingly pass along a redundant query.

Recall that a node will only propagate a query the first time that it is detected. Therefore the reception of packets per query is essentially determined by how the query propagates locally, and is not affected by the size of the network. However, the amount of data carried in each packet may depend on the size of the network. When route accumulation is performed in the query packet, the packets increase in length as they are forwarded. Larger networks would allow queries to travel further, thereby increasing the average packet length. On the other hand, if route accumulation is implemented through temporary caching of route information, then the query packet lengths will remain fixed, and independent of network size. (The length of route replies will increase with network length. However, the contribution of route replies to the overall IERP traffic is negligible.)

C. IERP Control Traffic: Querying Rate

Our understanding of the cost per query provides only partial insight into the entire cost of querying. Recall that the rate which queries are initiated depends on the stability of the discovered routes. Figures 8a – 8c demonstrate dependence of route stability on the zone radius, node density and node population. In the interest of conserving space, the linear relationship between the route failure rate and node velocity is not explicitly shown.[10] Previously, we showed that maintaining larger, denser routing zones required increasing amounts of IARP traffic. Some of this cost can now be justified by the improvement in route stability that they provide. An increase in node density expands the set of *intrazone* routes to a given destination, increasing the lifetime of a connection specified in an IERP source route. By increasing the zone radii, fewer connections need to be specified in the acquired loose source route. This provides fewer opportunities for route failure to occur

[10] Since route failure is triggered by node mobility, it follows that increasing the node velocity by a factor of α will cause the route failure to occur α times faster. The same reasoning can be applied to the triggering of IARP route updates.

Figure 8a

Figure 8b

Figure 8c

The following simple example illustrates the impact of the zone radius. Take two routes between nodes A and B. Route 1 consists of the following one hop connections: A-X-Y-B (i.e. based on a routing zone of radius 1 hop). Route 2 is constructed based on a routing zone of radius 3 hops, and is specified only by the loose connection A-B. In order for Route 2 to fail, all routes between A-B, less than or equal to 3 hops (of which Route 1 is only one such route) have to fail. Consequently, the looser connection is at least as reliable as the more specific sequence of shorter connections.

Keeping the zone radii and node density fixed, an increase in network population (N) is reflected by an increase in the network span. Assuming that nodes may communicate with any other node (rather than just those nodes that are geographically close), the average destination becomes more distant (in hops). This has the effect of increasing the length of an IERP source route, thereby reducing the route's reliability.

D. Total ZRP Traffic Performance

Our insight into the behavior of the IARP and IERP can now be applied to the overall ZRP performance. Figures 9a and 9b illustrate the effect of node density (δ) on the production of ZRP traffic. We've seen that an increase in node density results in more IARP route updates and more IERP packets per query. In cases where the rate of queries does not depend on route stability (i.e. R_{route_usage} < $R_{route\_failure}$), we expect the ZRP traffic to also increase with node density. On the other hand, it is not immediately clear how ZRP traffic will be affected when the query rate does depend on route stability, because higher node densities increase a route's reliability and therefore decrease the route query rate. Figure 9b demonstrates this reduction in route query rate is more than offset by the increase in overlapping query packet transmission. In general, it appears that the ZRP traffic is an increasing function of node density.

Figure 9a: R_{route_usage} << $R_{route\_failure}$

Figure 9b: R_{route_usage} >> $R_{route\_failure}$

Also of interest is how the optimal zone radius configuration (ρ_{opt}) depends on node density. Over the range of node densities examined, the optimal zone radius remained constant (ρ_{opt} = 4, for R_{route_usage} << $R_{route\_failure}$, and ρ_{opt} = 3 for R_{route_usage} >> $R_{route\_failure}$). However, we note that the effect of δ appears to be more significant for the IARP traffic than the IERP traffic. A large enough increase in δ would eventually make a routing zone prohibitively expensive to maintain, resulting in a decrease in ρ_{opt}.

The average node velocity (v) is a measure of the rate of network reconfiguration. Higher node velocities result in a linear increase in the IARP routing zone updates and IERP route failures. Figures 9c and 9d demonstrate that, as expected, the average rate of ZRP traffic increases with v. The effect of v on the optimal zone radius depends on the user demand for routes. When R_{route_usage} >> $R_{route\_failure}$, route discoveries are driven by route failures. Therefore, the overall ZRP traffic increases linearly with v, and ρ_{opt} remains constant. On the other hand, when R_{route_usage} << $R_{route\_failure}$, the route query rate is essentially independent of route stability, and node velocity. This means that larger v only serve to increase the load on the IARP, thereby favoring smaller zone radii.

Figure 9c: R_{route_usage} << $R_{route\_failure}$

Figure 9d: R_{route_usage} >> $R_{route\_failure}$

Figure 9e: R_{route_usage} << $R_{route\_failure}$

Figure 9f: R_{route_usage} >> $R_{route\_failure}$

Finally, we examine the dependence of the ZRP traffic on the node population (N). Because IARP route updates are a local event, they do not depend on the size of the network. By similar arguments, the amount of IERP traffic received by a node per query is also independent of N. The node population influences the ZRP through its effect on the rate of received route queries. The addition of a new network node places additional load on the network through the extra route queries that it initiates. All other factors remaining constant, an increase in N results in an increase in network span. When the rate of route queries is driven by route failure, we've shown that larger network span reduces route reliability, further increasing the query load by all other nodes. Figures 9e and 9f demonstrate this behavior. As expected, the amount of ZRP traffic increases with N. Furthermore, since this increase only occurs in the reactive part of the protocol, larger N favors a larger routing zone radius. The ability of larger routing zones to provide increased query efficiency and route reliability more than offsets their extra maintenance cost.

V. Adapting the ZRP Through Routing Zone Radius Configuration

Our analysis of the ZRP has demonstrated the many ways in which network characteristics and node behavior can affect the amount of overhead traffic. For any given scenario, it is desirable for the ZRP to operate as efficiently as possible. As we've shown, this can be achieved through proper selection of the routing zone radius.

In general, choosing the optimum routing zone radius requires an accurate model of the network and individual node behavior. Even with perfect knowledge of all network parameters, computation of the optimal routing zone radius is not straightforward. Through the dissection of the ZRP in Sections III and IV, we were able to illustrate how changes in a single parameter can affect the amount of ZRP traffic and the optimal zone radius. However, a change in a single parameter may not be significant enough to alter the optimal zone radius, and changes in multiple parameters may work against each other, making the resulting ZRP performance unclear.

Further research could focus on the derivation of a complete ZRP traffic function. However, the accuracy of the optimal zone radius computation would still be limited by the quality of the network model estimates. Some factors may be fairly easy to estimate on-line. For instance, node density and relative node velocity can be estimated by the IARP, based on the average of the number neighbors and the average rate of new neighbor acquisition. Likewise, network span can be inferred through analysis of the IERP route query traffic. Unfortunately there are many other influential factors, such as route selection criteria, route caching policies and data traffic behavior, that are not readily estimated based on information available to each user.

The previous discussion reveals the limitations of acquiring good estimates for network and node behavior, based on measurements of IARP and IERP traffic. Given our understanding the ZRP, we are able to propose and evaluate two new zone sizing schemes, "min-searching" and "traffic adaptive", which are designed to minimize the amount of control traffic *based directly on the control traffic measurements themselves.*

A. *Basic Estimation Through Min-Searching*

A simple approach to estimating the optimal zone radius is to periodically adjust the zone radius until a radius is found that appears to be minimize the amount of ZRP traffic. We refer to this type of scheme as *min-searching*. When the estimation process begins, the routing zone radius can either be incremented ($\Delta\rho = +1$) or decremented ($\Delta\rho = -1$) by one hop. The choice to initially increment or decrement can be arbitrary, or based on additional information provided by the triggering mechanism. During the k^{th} estimation interval, the amount of ZRP traffic, $Z(\rho(k))$, is measured. If the current amount of ZRP traffic is less than the previous amount $(Z(\rho(k)) < Z(\rho(k-1)))$, it is assumed that the ZRP traffic can be further reduced by continuing to increment / decrement the zone radius (i.e. $\rho(k+1) < \rho(k) + \Delta\rho$). Otherwise, the direction of radius change is reversed ($\Delta\rho = -\Delta\rho$) and the zone radius is altered accordingly. The process continues until a minimum is detected, based on the following condition: ($Z(\rho) < Z(\rho - 1)$ *and* $Z(\rho) > Z(\rho + 1)$).

Figure 10 illustrates the operation of the min-searching scheme. Initially (k=0), the routing zone radius is incremented by one hop. Until k=3, the amount measured traffic appears to decrease with an increase in ρ. This motivates the continued increase in ρ. At k=3, the amount of ZRP traffic exceeds that of the previous routing zone. Furthermore, the min search can be terminated at this point, because $\rho(2)$ meets the criteria for a minimum.

Figure 10: Min-Searching Example

Min-searching discovers a local minimum, provided that the network *behavior* does not change substantially during the min-search and that the measurements of ZRP traffic are exact. Clearly, this scheme becomes less effective as these conditions are relaxed. Estimates of the ZRP traffic can only be made more precise by increasing the duration of the estimation interval (T). However this comes at the expense of reducing the correlation of the network behavior between successive intervals. This trade-off has to be considered when implementing this kind of optimization.

Even under ideal conditions, the discovery conditions may not guarantee that a discovered minimum is global. If the detected minimum is not a global minimum, the min-search scheme could result in an extreme amount of ZRP control traffic. However, our experience has been that the IARP and IERP traffic are each convex functions of ρ (in the region of ρ > 0). Therefore, the total ZRP traffic, which is the sum of these two components, is also convex. Thus, the discovered local minimum is, in fact, a global minimum.

B. Analysis of the Estimation Triggering Mechanism

The efficiency of zone radius sizing scheme depends on the effectiveness of the triggering mechanism. An acceptable trigger should keep the likelihood of Type I (false alarm) and Type II (miss) errors to an acceptable level. In order to identify the optimal zone radius, ρ_{opt}, the min-searching scheme must evaluate the ZRP traffic at least two other radii[11], $\rho_{opt}+1$ and $\rho_{opt}-1$. In the case of a false alarm, the current radius is already optimal, so the cost of a Type I error is:

$$C_{Type\,I} = Z(\rho_{opt}+1) - Z(\rho_{opt}) + Z(\rho_{opt}+1) - Z(\rho_{opt})$$

When the trigger does not activate in response to a change in the optimal routing zone radius, a miss occurs. The cost of a miss between two min-searches is equal to the amount of excess ZRP traffic accumulated between the end of the first min search (at interval k_1) and the beginning of next min-search (at interval k_2):

$$C_{Type\,II} = \sum_{k=k_1}^{k_2}\left[Z(\rho(k)) - Z(\rho_{opt}(k))\right] = \sum_{k=k_1}^{k_2}\left[Z(\rho(k_1)) - Z(\rho_{opt}(k))\right]$$

$$= (k_2 - k_1)\cdot Z(\rho(k_1)) - \sum_{k=k_1}^{k_2} Z(\rho_{opt}(k))$$

If this excess cost = 0, then a miss did not occur during this period.

Even when the triggering mechanism does trigger correctly, some cost is incurred. This is because the min-search cannot determine the new optimal zone radius immediately. The cost of a correct trigger can be upper-bounded by: [12]

$$C_{correct} \le \begin{cases} \displaystyle\sum_{\rho=\rho(0)-1}^{\rho_{opt}+1}\left[Z(\rho) - Z(\rho_{opt})\right] & ,\rho(0) < \rho_{opt} \\ \displaystyle\sum_{\rho=\rho_{opt}-1}^{\rho(0)+1}\left[Z(\rho) - Z(\rho_{opt})\right] & ,\rho(0) > \rho_{opt} \end{cases}$$

In order to identify the optimal radius, the scheme must overshoot it by one hop: $\rho_{opt}+1$. Overshooting is necessary for all min-search schemes. The limit $\rho(0) \pm 1$ arises from situations where the min-search initially adjusts ρ in the wrong direction. While this problem cannot be entirely eliminated, traffic measurements can be used to make an informed estimate about the direction of the optimal zone radius. An increase in the amount of reactive IERP traffic or a decrease in the amount of proactive IARP traffic indicates that the optimal zone radius may be larger than the current zone radius. Likewise, a decrease in the IERP or an increase in IARP traffic would suggest a smaller optimal routing zone. In the case where both sets of traffic exhibit an increase or decrease, the choice of initial direction may not be clear. In this case, the direction could be chosen at random or by making a decision based on the ZRP component (IERP vs. IARP) that exhibits the more significant change.

A simple choice for a trigger would be a timer. Shorter timers would increase the likelihood of false alarms because of the time-correlation of the network behavior. Timers that are too long, on the other hand, increase the likelihood of a miss. In the min-searching scheme, false alarms are easily detected because the new optimum zone radius is the same as the previous optimum zone radius. Unfortunately, misses are not detectable. In the absence of additional information, the best timer strategy would be to start with a short timer, and gradually increase the time-out value until the measured probability of false alarm falls below an acceptable level.

Another choice for a trigger would be based on the amount of change in the IARP and IERP traffic. As previously discussed, the change in traffic can be used to estimate the proper initial direction of the min-search. The min-search could be triggered in cases where the IARP and IERP provide a strong indication that the optimal zone radius has changed. As with the timer, properly chosen thresholds for the changes in IERP and IARP traffic can help keep the false alarm probability just below an acceptable level, providing the best prevention against the undetectable Type II errors.

C. Adaptive Traffic Estimation

The discussion so far has focussed on the implementation of min-searching as a way to determine the optimal routing zone radius. Recall that this simple technique requires that all other factors contributing to the ZRP traffic remain constant during the execution of the min-search. However, the estimation interval needs to be long enough to provide accurate measurements of the ZRP traffic. It is quite conceivable that a sufficiently long estimation interval may be too long to provide adequate correlation beyond two consecutive intervals. If the behavior of the network is erratic (e.g. due to frequent partitioning and rejoining of the network), the min-search scheme may exhibit unacceptable performance.

In cases where long estimation intervals can reduce the accuracy of the multiple interval min-searching schemes, a more desirable approach would be one that adjusts the zone radius based *only* on measurements gathered from the

---

[11] The one exception would be ρ_{opt} = 1. In this case, only ρ_{opt} + 1 needs to be evaluated, because ρ_{opt} = 1 is an endpoint.

[12] Here we assume that the min-search scheme is slightly improved so that it doesn't evaluate the same zone radius twice during one search.

current estimate interval. Such a technique is possible if special properties of the ZRP traffic can be exploited.

When the zone radius is less than the optimal zone radius and the corresponding amount of ZRP traffic is significantly more than optimal, the ZRP traffic is dominated by the reactive IERP query traffic (i.e. reactive traffic / proactive traffic >> 1). Likewise, when the zone radius is larger than the optimal zone radius and the corresponding amount of ZRP is significantly more than optimal, the ZRP is dominated by the proactive IARP route updates (i.e. reactive traffic / proactive traffic << 1). These regions are depicted in Figure 11. The optimal routing zone radius lies in a region where neither the IARP nor the IERP is dominant.

Figure 11: Location of the Optimal Zone Radius

This behavior can be exploited to resize the routing zone in the following manner. Let $\Gamma(\rho)$ be the ratio of reactive (IERP) traffic to proactive (IARP) traffic, at radius ρ. Adjustment of ρ is made by comparing this ratio with a predetermined threshold,, Γ_{thresh}. If $\Gamma(\rho) > \Gamma_{thresh}$, increase the zone radius; if $\Gamma(\rho) < \Gamma_{thresh}$, decrease the zone radius. We refer to this scheme as *traffic adaptive*.

The traffic adaptive scheme can use $\Gamma(\rho)$ as the basis for its triggering mechanism as well. An appropriate trigger design calls for ρ to be resized when the ZRP enters the very reactive or very proactive operating regions. This condition can be realized by the introduction of hysteresis, H, such that $\Gamma(\rho) > H \cdot \Gamma_{thresh}$ or $\Gamma(\rho) < (1/H) \cdot \Gamma_{thresh}$.

Figure 12: A Hybrid Min-Search/Traffic Adaptive Zone Radius Estimator

Special consideration must be given to the case where $\rho_{opt} = 1$. Because $\Gamma(1) = \infty$, a zone radius of one hop would, by definition, be considered very reactive. Without any additional control, the traffic adaptive algorithm could be misled by the large value of $\Gamma(1)$, resulting in an oscillation between $\rho = 1$ and $\rho = 2$. This instability can be addressed by switching from the traffic adaptive scheme to min-searching when ρ reaches one hop. An implementation of this dual approach is illustrated in Figure 12. When the min-searching scheme increments ρ beyond two hops, control can be switched back to the traffic adaptive scheme.

While the zone sizing algorithm is operating in min-searching mode, the traffic adaptive scheme is given an opportunity to improve its assignments of hysteresis (H) and decision threshold (Γ_{thresh}). When the min-searching scheme identifies a minimum, the traffic adaptive scheme is alerted and can record the traffic ratio of that interval. Adjustments to the traffic adaptive parameters can be made based on statistical analysis of these ratios. For example, Γ_{thresh} can be adjusted based on the mean of the log-ratios, and H based on their standard deviation. Because the min-searching scheme may provide useful side information, it may be beneficial for the min-searching scheme to *occasionally* assume control, even when $\rho > 1$. Assuming that these parameter updates occur only occasionally, the cost of the min-search would not have a significant affect on the overall performance of the sizing scheme.

VI. ROUTING ZONE SIZING PERFORMANCE RESULTS

An evaluation of the zone radius sizing schemes requires an understanding of how the ZRP reacts to changes in network configuration. Based on our analysis of the ZRP in Sections III and IV, these relationships are fairly well understood.

In order to proceed, we require a model to express how network characteristics, such as relative velocity, network population, node density, etc. change over time. This time dependent behavior is very network specific. Issues such as geographic node distribution, radio propagation conditions and coordination of node movement, need to be considered and are quite varied from network to network. Because of these complex dependencies and diverse behavior, there are no widely accepted models for "standard" ad-hoc network behavior.

Keeping this limitation in mind, we base our simulations on a network model designed specifically to test the ability of the routing zone sizing schemes to adapt to a wide range of operating conditions.

The design of this test-bed network is based on the following criteria:

- The optimal zone radii should be allowed to vary over a reasonable range. Too little variability will not adequately reflect the tracking ability of these schemes. On the other hand, too much variability is not realistic time-varying behavior for most ad-hoc networks (even RWNs). Our test-bed network is designed to operate mostly within an approximate range of 4 hops to 7 hops.

- The model should be general enough to allow simulation for different levels of correlation between successive estimation intervals.
- The model should be of limited complexity, to allow easy implementation and analysis.

Through a mixture of analysis and simulation, we found that the following Gaussian distributions for node population (N) and relative node velocity (v) provide an acceptable range of zone radii (for a node density of $\delta = 6$ neighbors):

	Node population (N)	Relative node velocity (v)
Mean (μ)	1500 [nodes]	1.0 [neighbors/sec]
Std. Dev. (σ)	500 [nodes]	0.3 [neighbors/sec]

Both distributions are truncated at 0 to prevent negative values for population and velocity. In addition, the values of N are rounded to the nearest integer (since node population is an integer quantity). These distributions can be said to model the effects of partial network partitioning and short term coordination of mobility patterns, respectively.

The correlated values of N and v corresponding to each estimation interval k are generated according the following Markov processes:

$$N(k+1) - \mu_N = a \cdot \left(N(k) - \mu_N\right) + \sigma_N \sqrt{1-a^2} \cdot X_N(k)$$
$$v(k+1) - \mu_v = a \cdot \left(v(k) - \mu_v\right) + \sigma_v \sqrt{1-a^2} \cdot X_v(k)$$

where $X_n(k)$ and $X_v(k)$ are i.i.d zero-mean, unit variance Gaussian random variables. The value of a reflects the correlation between two successive sampling intervals. The schemes are simulated over a range of $0.0 \le a \le 0.99$.

Figure 13

Figure 13 shows the relative performance of the basic sizing schemes introduced in Section V. The zone radius estimates provided by a properly configured traffic adaptive scheme are able to keep the ZRP traffic levels between 1% - 2% of optimal. For comparison, the *best* min-search scheme produces between three to five times more extra ZRP than the best traffic adaptive scheme. This is not to say that min-search schemes do not perform well. Even a continuously triggered min-search produces less than 12% of the optimal ZRP control traffic. This suggests that the cost of a min-search may be low enough to justify the hybrid min-search / traffic adaptive scheme proposed in Section V.C.

Figure 14

The traffic adaptive scheme performs extremely well when configured with the appropriate decision threshold[13] (Γ_{thresh}) and hysteresis (H). Figure 14 shows the optimal decision threshold to be about 4 and the optimal hysteresis to be approximately 3. More importantly, we note that good traffic adaptive performance does not rely on precise estimates of Γ_{thresh} and H. The traffic adaptive scheme lies within approximately 5% of the optimal ZRP traffic for $2 \le \Gamma_{thresh} \le 7$, even with no hysteresis (H=1). This acceptable level estimation error makes the on-line adjustment of these parameters a practical approach.

Figure 15

So far, we have been gauging the cost of zone radius re-sizing in terms of extra ZRP control traffic. However, a change in the zone radius requires the routing zone nodes to be informed.[14] Since the estimation intervals are intended to be sufficiently long to obtain good estimates of average received traffic, these zone radius updates occur much less frequently than the neighbor discoveries that trigger IARP updates (perhaps by a factor of 10 to 100 times less frequently). Therefore, the amount of zone radius resizing

[13] Recall that the decision threshold is the ratio of reactive IERP traffic vs. proactive IARP traffic.

[14] Routing zone re-sizing consists of zone radius estimation and a broadcast-based protocol to update the routing zone nodes. This paper deals with the issue of zone radius estimation. The design of the update protocol is outside the scope of this paper.

traffic should be negligible when compared with the IARP traffic. Even so, it may still be desirable to minimize this zone radius update traffic without compromising the ZRP performance. Figure 15 illustrates the frequency of radius updates, relative to the rate of the true optimal radius updates. When the traffic adaptive scheme is implemented with no hysteresis, updates are performed about five times as often as necessary. As the amount of hysteresis increases, we find that the rate of radius updates decreases. For H=4.5, we find that the traffic adaptive scheme updates the zone radius for only 5% of the optimal radius updates. This implies a miss probability of near 95%, which would suggest that the radius sizing scheme should perform very badly. This, however, is not the case.

The traffic adaptive scheme is designed to keep the zone radius within a hybrid reactive-proactive region. Referring back the figures from section V, we can see that within this hybrid region, the amount of control traffic does not vary greatly. The traffic adaptive is able to successfully exploit this behavior, providing very good performance without having to track the optimal zone radius precisely.

Based on the excellent performance exhibited by the traffic adaptive scheme, we advocate its use as the primary zone radius estimator for configuring the ZRP. The min-searching schemes, though outperformed by traffic adaptive estimation, still provide acceptable suppression of ZRP traffic. This justifies the *occasional* use of min-searching as a way to fine-tune the operation of traffic adaptive estimation.

VII. Conclusions

The Zone Routing Protocol (ZRP) provides a flexible solution to the challenge of discovering and maintaining routes in the Reconfigurable Wireless Network (RWN) communication environment. The ZRP combines two radically different methods of routing into one protocol. Route discovery is based on a reactive route request / route reply scheme. This querying can be performed efficiently through the proactive maintenance of a local *routing zone* topology.

The amount of intrazone control traffic required to maintain a routing zone increases with the size of the routing zone. However, through a combination of bordercasting and query detection/termination, we are able to exploit the knowledge of the routing zone topology to reduce the amount of interzone route query traffic. This tradeoff between the costs of proactive and reactive components of the ZRP determines the optimal zone radius for a given network configuration. The span of the network does not affect the amount of intrazone traffic, but the amount of reactive route query traffic increases with network span, thereby favoring larger routing zones. Relative node velocity has the potential to increase both intrazone updates and interzone route queries. When route usage is frequent, node velocity has little affect on the optimal zone radius. However, occasional route usage makes the route query traffic independent of route failures and consequently, node velocity. In those situations, only proactive traffic increases with node velocity, and smaller zone radii provide more efficient ZRP operation. Finally, we've seen that the total ZRP traffic increases with node density. An increase in density has more of an impact on the proactive route updates. As a result, the optimal zone radius decreases with node density, although this dependence is fairly weak. In summary, dense networks consisting of a few relatively fast moving nodes favor reactive (small zone radius) configurations. On the other hand, a sparse network consisting of many slowly moving nodes would favor a more proactive (large zone radius) configuration.

For any particular network configuration, each node has an optimal routing zone radius. Given perfect knowledge of the network, it is possible to determine the best choice for the zone radius. In practice, much of this information cannot be measured, or even estimated, by the network nodes. In order for a node to estimate its optimal zone radius, it must make use of the information that is directly available. We have proposed and evaluated two classes of zone radius estimation algorithms, both of which attempt to minimize the amount of ZRP traffic based on direct measurements of the traffic. The first class, which we refer to as "min-searching", is based on the following observations about the ZRP behavior: proactive traffic increases with the zone radius, reactive traffic decreases with zone radius, and both proactive and reactive traffic are convex with respect to the zone radius. As their name implies, min-searching schemes apply these properties by searching for the optimal zone radius based on measurements of the ZRP traffic for a range of radii, until a minimum is found by inspection. Simulation results on a test-bed RWN demonstrate that efficiently triggered min-search schemes can keep the ZRP traffic to within 7% of the minimum traffic.

The second class of zone radius estimators exploits the tendency for the optimal zone radius to lie in a region where the amounts of reactive and proactive control traffic are comparable. This approach is termed "traffic adaptive", because it adapts the routing zone, based only on the current measurements of ZRP traffic. Our test-bed simulation demonstrated that the traffic-adaptive scheme outperforms min-searching, maintaining the ZRP traffic within 1% - 2% of the optimal configuration.

Our results demonstrate that the proposed route estimation techniques, applied in conjunction with a simple radius update protocol, allow the ZRP to perform more efficiently than traditional routing protocols, without the need for centralized control or knowledge of the network operating conditions.

ACKNOWLEDGEMENTS

The authors thank the following organizations and individuals for their continuous support of this work:

Dr. Roger Peterson (Motorola), Dr. Michael Kotzin (Motorola), and Mr. Siamik Tabrizi (Air Force Research Laboratory)

REFERENCES

[1] D. Bertsekas and R. Gallager, *Data Networks,* Second Edition, Prentice Hall, Inc., 1992.

[2] C. Cheng, R. Reley, S.P.R. Kumar, and J.J. Garcia-Luna-Aceves, "A Loop-Free Extended Bellman-Ford Routing Protocol without Bouncing Effect," *ACM Computer Communications Review,* 19(4), 1989, pp.224-236.

[3] J.J. Garcia-Luna-Aceves, "Loop-Free Routing Using Diffusing Computations," *IEEE/ACM Transactions on Networking,* vol.1, no.1, Feb. 1993, pp.130-141.

[4] Z.J. Haas and M.R. Pearlman, "The Performance of a New Routing Protocol for the Reconfigurable Wireless Networks," ICC'98, Jun. 8-11, 1998

[5] Z.J. Haas and M.R. Pearlman, "Evaluation of the Ad-Hoc Connectivity with the Reconfigurable Wireless Networks," Virginia Tech's Eighth Symposium on Wireless Personal Communications, Jun. 10-12, 1998

[6] Z.J. Haas and M.R. Pearlman, "The Performance of Query Control Schemes for the Zone Routing Protocol," SIGCOMM'98, Sept. 2 –4, 1998

[7] P. Jacquet, P. Muhlethaler, and A. Qayyum, "Optimized Link State Routing Protocol," *IETF MANET*, Internet Draft, Nov. 1998.

[8] D.B. Johnson and D.A. Maltz, "Dynamic Source Routing in Ad-Hoc Wireless Networking," in *Mobile Computing,* T. Imielinski and H. Korth, editors, Kluwer Academic Publishing, 1996.

[9] J. Moy, "OSPF Version 2", *RFC 2178,* Mar. 1997.

[10] S. Murthy and J.J. Garcia-Luna-Aceves, "A Routing Protocol for Packet Radio Networks," Proc. of ACM Mobile Computing and Networking Conference, *MOBICOM'95,* Nov. 14-15, 1995.

[11] S. Murthy and J.J. Garcia-Luna-Aceves, "An Efficient Routing Protocol for Wireless Networks," *MONET,* vol.1, no.2, pp.183-197, Oct. 1996.

[12] V.D. Park and M.S. Corson, "A Highly Adaptive Distributed Routing Algorithm for Mobile Wireless Networks," *IEEE INFOCOM '97*, Kobe, Japan, 1997

[13] C. E. Perkins and P. Bhagwat, "Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers," *ACM SIGCOMM,* vol.24, no.4, Oct. 1994, pp.234-244.

[14] C. E. Perkins and E.M. Royer, "Ad Hoc On-Demand Distance Vector Routing," *IEEE WMCSA'99*, New Orleans, LA, Feb. 1999.

[15] J. Sharony, ``A Mobile Radio Network Architecture with Dynamically Changing Topology Using Virtual Subnets,'' *MONET,* vol.1, no.1, pp.75-86. Also in Proceedings of *ICC/SUPERCOM'96,* Dallas, TX, Jun. 23-27, 1996, pp.807-812.

[16] F. A. Tobagi and L. Kleinrock, "Packet Switching in Radio Channels: Part 2 - The Hidden Terminal Problem in Carrier Sense Multiple-Access and the Busy Tone Solution, *IEEE Trans. on Comm.,* vol. 23, no.12, pp.1417-1433, Dec. 1985.

[17] C-K Toh, "Associativity-Based Routing for Ad-Hoc Mobile Networks," Wireless Personal Communications Journal, Special Issue on Mobile Networking and Computing Systems, Kluwer Academic Publishers, vol. 4, no. 2, pp. 103-139, Mar. 1997.

[18] P.F. Tsuchiya, "The Landmark Hierarchy: A New Hierarchy for Routing in Very Large Networks," *ACM Computer Communications Review*, vol. 18, no. 4, pp.35-42, 1988