Network Working Group
Request For Comments: 1863
Category: Experimental

D. Haskin
Bay Networks, Inc.

October 1995

A BGP/IDRP Route Server alternative to a full mesh routing

Status of this Memo

This memo defines an Experimental Protocol for the Internet community. This memo does not specify an Internet standard of any kind. Discussion and suggestions for improvement are requested. Distribution of this memo is unlimited.

Abstract

This document describes the use and detailed design of Route Servers for dissemination of routing information among BGP/IDRP speaking routers.

The intention of the proposed technique is to reduce overhead and management complexity of maintaining numerous direct BGP/IDRP sessions which otherwise might be required or desired among routers within a single routing domain as well as among routers in different domains that are connected to a common switched fabric (e.g. an ATM cloud).

1. Overview

Current deployments of Exterior Routing protocols, such as the Border Gateway Protocol [BGP4] and the adaptation of the ISO Inter-Domain Routing Protocol [IDRP], require that all BGP/IDRP routers, which participate in inter-domain routing (border routers) and belong to the same routing domain, establish a full mesh connectivity with each other for purpose of exchanging routing information acquired from other routing domains. In large routing domains the number of intra-domain connections that needs to be maintained by each border route can be significant.

In addition, it may be desired for a border router to establish routing sessions with all border routers in other domains which are reachable via a shared communication media. We refer to routers that are directly reachable via a shared media as adjacent routers. Such direct peering allows a router to acquire "first hand" information about destinations which are directly reachable through adjacent routers and select the optimum direct paths to these destinations. Establishment of BGP/IDRP sessions among all adjacent border routers would result in a full mesh routing connectivity. Unfortunately for

a switched media as ATM, SMDS or Frame Relay network which may inter-connect a large number of routers, due to the number of connections that would be needed to maintain a full mesh direct peering between the routers, makes this approach impractical.

In order to alleviate the "full mesh" problem, this paper proposes to use IDRP/BGP Route Servers which would relay external routes with all of their attributes between client routers. The clients would maintain IDRP/BGP sessions only with the assigned route servers (sessions with more than one server would be needed if redundancy is desired). All routes that are received from a client router would be propagated to other clients by the Route Server. Since all external routes and their attributes are relayed unmodified between the client routers, the client routers would acquire the same routing information as they would via direct peering. We refer to such arrangement as virtual peering. Virtual peering allows client routers independently apply selection criteria to the acquired external routes according to their local policies as they would if a direct peering were established.

The routing approach described in this paper assumes that border routers possess a mechanism to resolve the media access address of the next hop router for any route acquired from a virtual peer.

It is fair to note that the approach presented in this paper only reduces the number of routing connection each border router needs to maintain. It does not reduce the volume of routing information that needs to maintained at each border router.

Besides addressing the "full mesh" problems, the proposal attempts to achieve the following goals:

- to minimize BGP/IDRP changes that need to be implemented in client routers in order to inter-operate with route servers;

- to provide for redundancy of distribution of routing information to route server clients;

- to minimize the amount of routing updates that have to be sent to route server clients;

- to provide load distribution between route servers;

- to avoid an excessive complexity of the interactions between Route Servers themselves.

2. Terms And Acronyms

The following terms and acronyms are used in this paper:

Routing Domain	-	a collection of routers with the same set of routing policies. For IPv4 it can be identified with an Autonomous System Number, for IPv6 it can be identified with a Routing Domain Identifier.
Border Router (BR)	-	a router that acquires external routes, i.e. routes to internet points outside its routing domain.
Route Server (RS)	-	a process that collects routing information from border routers and distributes this information to 'client routers'.
RS Client (RC)	-	a router than peers with an RS in order to acquire routing information. A server's client can be a router or another route server.
RS Cluster (RSC)	-	two or more of route servers that share the same subset of clients. A RS Cluster provides redundancy of routing information to its clients, i.e. routing information is provided to all RS Cluster clients as long as there is at least one functional route server in the RS Cluster.
RCID	-	Cluster ID

3. RS Model

In the proposed scheme a Route Server (RS) does not apply any selection criteria to the routes received from border routers for the purpose of distributing these routes to its clients. All routes acquired from border routers or other Route Servers are relayed to the client border routers.

There can be two classes of Route Servers: Route Servers that relay external routes between routers in a single routing domain and Route Servers that relay external routes between border routers in different routing domains. The former are Intra-Domain Route Servers and the latter are Inter-Domain Route Servers.

In the RS model proposed in this document there is no routing exchange between Intra-Domain Route Servers and Inter-Domain Route

Servers. Routes that cross a domain boundary must always pass through a border router of such a domain which may apply administrative filters to such routes.

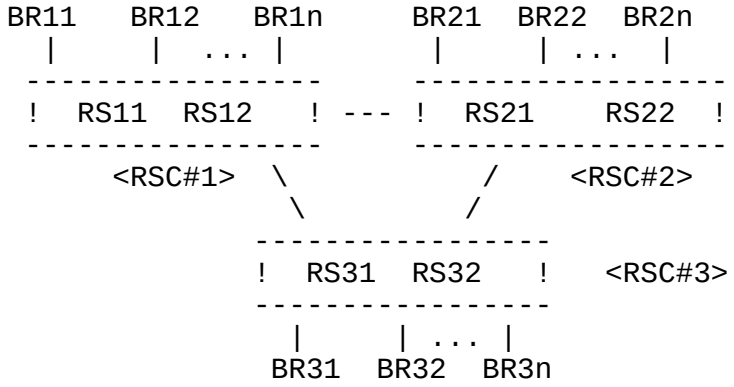
Operations of Intra-Domain Route Servers and Inter-Domain Route Servers are identical.

One or more Route Servers form an RS Cluster (RSC). For redundancy's sake two or more RSs can be configured to operate in an RS Cluster. All route servers in an RSC share the same clients, i.e. cluster clients establish connections to all route servers in such an RSC for the purpose of exchanging routing information. Each cluster is assigned an unique RSC Identifier (RCID) represented by a 2-octet unsigned integer.

Clusters which provide virtual connectivity between their clients would be normally exchanging routing information among themselves so that all external routes are propagated to all participating clients.

Though a Route Server Client (RC) can be associated with multiple RSC, it seems that there is no real advantage of doing so except for a short transition period to provide a graceful re-assignment from one RSC to another or, if for some reason, there are multiple RS groups that don't exchange routing information with each other.

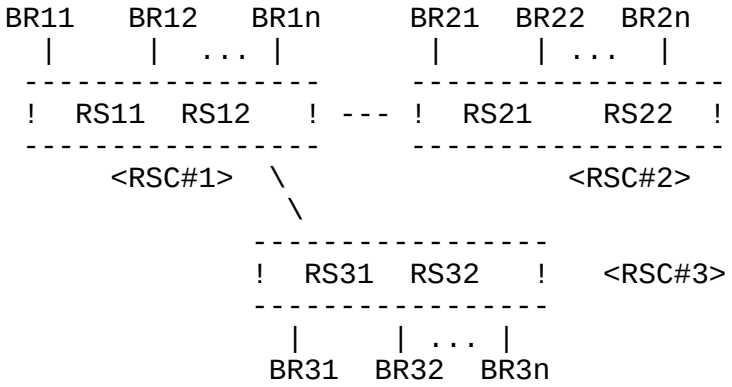
The inter-cluster route exchange can be accomplished by forming a full mesh routing adjacency between clusters. In this approach, illustrated in the diagram below, each RS in each RSC would maintain a routing connection with every RS in other RS clusters. Only routes that are acquired from border routers are propagated to RSs in other RS clusters.



Another way to propagate routing information between clusters would be to form a cluster hierarchy in which an RS in one cluster maintains sessions only with RSs in designated clusters. In this

approach an RS must advertise all acquired routes to an RS in another cluster except the routes that are acquired from that cluster. Nevertheless, it allows for minimizing the number of routing sessions which can be highly desirable in some network. It is important for the hierarchical scheme that the inter-cluster route exchange links form a tree, i.e. there is only one route propagation path between any two clusters, otherwise routing loops may result. For detection and pruning of routing loops in a hierarchical cluster topology, it is advisable to include the "RCID Path" attribute (see 4.3.4) in all routing updates sent between route servers. This attribute lists IDs of all clusters in the route propagation path. When a duplicate ID is detected in this attribute an offending route needs to be discarded.

The diagram below which illustrates the hierarchical approach is created from the diagram above by removing the route exchange link between clusters 2 and 3.



It seems that the only disadvantage of the hierarchical model, is the management headache of avoiding routing loops and redundant information flow by insuring that inter-cluster links always form a tree. But more study is needed to fully evaluate the comparative merits of the full-mesh and hierarchical models.

Since RSs in the same cluster maintain routing sessions with the same set of clients, it may seem that there is no need to exchange routing information between RSs in the same cluster. Nevertheless, such a route exchange may help to maintain identical routing databases in the servers during client acquisition periods and when a partial failure may affect some routing sessions.

Route servers in the same RS cluster exchange control messages in attempt to subdivide the responsibilities of providing routing information to their clients. In order to simplify the RS design, the RS messaging is implemented on top of exterior protocol which is

used by route servers for the routing information exchange.

4. Operation

4.1 ADVERTISER Path Attribute

Route servers act as concentrators for routes acquired by border routers so that the border routers need to maintain routing connections with only one or two designated route servers. Route Servers distribute routing information that is provided to them by the border routers to all their client.

If routing information were relayed to RS clients in UPDATE messages with only those path attribute that are currently defined in the BGP-4/IDRP specification, the RS clients would not be able to associate external routes they receive with the border routers which submitted that routes to route servers. Such an association is necessary for making a correct route selection decision. Therefore, the new path attribute, ADVERTISER, is defined.

The ADVERTISER is an optional non-transitive attribute that defines the identifying address of the border router which originally submitted the route to a router server in order for it to be relayed to other RS clients. Type Code of the ADVERTISER attribute is 255. This attribute must be included in every UPDATE message that is relayed by route servers and must be recognized by RS clients.

4.2 Route Client Operation

An RS client establishes an BGP/IDRP connection to every route server in the RS cluster to which the route client is assigned.

RS clients must be able to recognize the ADVERTISER path attribute that is included in all UPDATE messages received from route servers. Routes received in UPDATE messages from route servers are processed as if they were received directly from the border routers specified in the ADVERTISER attributes of the respective updates.

If an RS client receives a route from a Intra-Domain Route Server, is assumed that the border router identified in the ADVERTISER attribute is located in the receiving client’s own routing domain.

If an RS client receives a route from a Inter-Domain Route Server, the locality of the border router identified in the ADVERTISER attribute can be determined from the BGP’s AS_PATH attribute or IDRP’s RD_PATH attribute respectively.

If no ADVERTISER attribute was included in an UPDATE message from a route server it is assumed that the route server itself is the advertiser of the corresponding route.

If the NEXT_HOP path attribute of an UPDATE message lists an address of the receiving router itself, the route that is carried in such an update message must be declared unreachable.

In addition, it is highly desirable, albeit not required, to slightly modify the "standard" BGP/IDRP operation when acquiring routes from RSs:

when a route is received from an RS and a route with the completely identical attributes has been previously acquired from another RS in the same cluster, the previously acquired route should be replaced with the newly acquired route. Such a route replacement should not trigger any route advertisement action on behalf of the route.

RSs are designed to operate in such a way that eliminates the need to keep multiple copies of the same route by RS clients and minimizes the possibility of a route flap when the BGP/IDRP connection to one of the redundant route servers is lost.

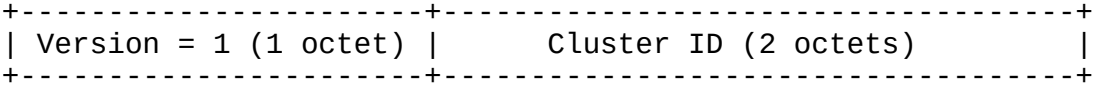
It is attempted to subdivide the route dissemination load between route servers such that only one RS provides routing updates to a given client. But since, for avoiding an excessive complexity, the reconciliation algorithm does not eliminate completely the possibility of races, it is still possible that a client may receive updates from more than one route server. Therefore, the client's ability to discard duplicate routes may reduce the need for a bigger routing database.

4.3 Route Server Operation

A Route Server maintains BGP-4/IDRP sessions with its clients according to the respective BGP-4/IDRP specification with exception of protocol modifications outlined in this document.

UPDATE messages sent by route servers have the same format and semantics as it respective BGP-4/IDRP counterparts but also carry the ADVERTISER path attribute which specifies the BGP Identifier of the border router that submitted the route advertised in the UPDATE message. In addition, if the hierarchical model is deployed to interconnect Route Server clusters, it is advisable to include the "RCID Path" attribute in all routing updates sent between route servers as described in 4.3.4.

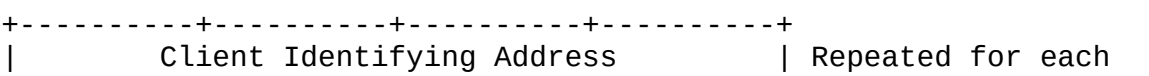
When route servers exchange OPEN messages they include the Route Server protocol version (current version is 1) as well as Cluster IDs of their respective clusters in an Optional Parameter of the OPEN message. The value of Parameter Type for this parameter is 255. The length of the parameter data is 3 octets. The format of parameter data is shown below:



Also, route servers that belong to the same cluster send to each other LIST messages with lists of clients to which they're providing routing information. In the LIST message an RS specifies the Router Identifier of each client to which that RS is providing routing updates. Since LIST messages are relatively small there is no need to add a processing complexity of generating incremental updates when a list changes; instead the complete list is sent when RSs need to be informed of the changes. The format of the LIST message is presented in 4.3.1.

4.3.1 LIST Message Format

The LIST message contains the fixed BGP/IDRP header that is followed with the fields shown below. The type code in the fixed header of the LIST message is 255.



The number of Client Identifying Address" fields is not encoded explicitly, but can be calculated as:

(<LIST message Length> - <Header Length>) / <Address Length>,

where <LIST message Length> is the value encoded in the fixed BGP/IDRP header, <Header Length> is the length of that header, and <Address Length> is 4 for IPv4 and 16 for IPv6.

4.3.2 External Route Acquisition And Advertisement

A route server acquires external routes from RS clients that are also border routers. A RS also may acquire external routes from other RSs. Route servers relay all acquired routes unaltered to their clients. No route selection is performed for purpose of route re-advertisement to RS clients.

While route servers receive and store routing data from all their client, Routing Servers in the same cluster coordinate their route advertisement in the attempt to ensure that only one RS provides routing updates to a given client. If an RS fails, other Route Servers in the cluster take over the responsibility of providing routing updates to the clients that were previously served by the failed RS. A route flap that can result from such switch-over can be eliminated by the configuring client's "Hold Time" of their BGP-4/IDRP sessions with the route servers to be larger than the switch-over time. The switch-over time is determined by the Hold Time of BGP-4/IDRP sessions between the route servers in the cluster and the period that is needed for that route servers to reconcile their route advertisement responsibilities. The reconciliation protocol is described in 4.3.3.

The BGP-4/IDRP operations of route servers differs from the "standard" operation in the following ways:

- when receiving routes from another RS, the RS Client mode of operation is assumed, i.e., when a route with completely identical attributes has been previously acquired from an RS belonging to the same cluster as the RS that advertises the new route, the previously acquired route should be discarded and the newly acquired route should be accepted. Such a route replacement should not trigger any route advertisement action on behalf of the route.

- all acquired routes are advertised to a client router except routes which were acquired from that client (no route echoing);

- if the hierarchical model of inter-cluster route exchange is used, all acquired routes are advertised to an RS in another RSC except routes that are acquired from that RSC. In the full-mesh model, only routes which are acquired from border routers are advertised to route servers in other clusters;

- if route servers in the same RS cluster are configured to exchange routing information, only external routes that are acquired from border routers are advertised to route servers in the local cluster;

- the ADVERTISER path attribute is included in every UPDATE messages that is generated by RS. This attribute must specify the identifying address of the border router from which information provided in UPDATE has been acquired. All other routing attributes should be relayed to RS's peers unaltered.

- when a route advertised by to an RS by a client becomes unreachable such a route needs to be declared unreachable to all other clients. In order to withdraw a route, the route server sends an UPDATE for that route to each client (except the client that this route was originally acquired) with the NEXT_HOP path attribute set to the address of the client to which this UPDATE is sent to. The the ADVERTISER path attribute with the identifying address of the border router that originally advertised the withdrawn route must be also included in such an update message.

- if the hierarchical model is deployed to interconnect Route Server clusters, it is advisable to include the RCID_PATH attribute in all routing updates sent between route servers as described in 4.3.4. The RCID_PATH attribute is never included in UPDATE messages sent to border routers.

4.3.3 Intra-Cluster Coordination

In order to coordinate route advertisement activities, route servers which are members of the same RS cluster establish and maintain BGP/IDRP connections between themselves forming a full-mesh connectivity. Normally, there is no need for more than two-three route servers in one cluster.

Route servers belonging to the same cluster send to each other LIST messages with lists of clients to which they’re providing routing information; let’s call such clients "informed clients".

Each RS maintains a separate "informed client" list for each RS in the local cluster including itself. All such lists are linked in an ascending order that is determined by the number of clients in each list; the order among the lists with the same number of clients is determined by comparing the identifying addresses of the corresponding RSs -- an RS in such a "same number of clients" subset is positioned after all RSs with the lower address.

An RS can be in one of two RS coordination states: ’Initiation’ and ’Active’.

4.3.3.1 Initiation State

This is the initial state of route server that is entered upon RS startup. When the Initiation state is entered the ’InitiationTimer’ is started. The Initiation state transits to the Active state upon expiration of the ’InitiationTimer’ or as soon as all configured BGP/IDRP connections to other route servers in the local RS Cluster are established and LIST messages from that route servers are

received.

In the Initiation state an RS:

- tries to establish connections with other RSs in the local and remote clusters.
- accepts BGP/IDRP connections from client routers.
- receives and process BGP/IDRP updates but doesn't send any routing updates.
- stores "informed client" lists received from other RSs in the local cluster - a newly received list replaces the existing list for the same RS. If a LIST message is received from the route server in another RS cluster, it should be silently ignored.
- initializes an empty "informed client" list for its own clients.
- as soon as a BGP/IDRP connection to an RS in the same RS Cluster is established, transmits an empty LIST message to such an RS.

4.3.3.2 Active State

This state is entered upon expiration of the 'InitiationTimer' or as soon as all configured BGP/IDRP connections to other route servers in the local RS Cluster are established and LIST messages from that route servers are received.

In the Active state an RS:

- continues attempts to establish connections with other route servers in the local and remote clusters;
- accepts new BGP/IDRP connections;
- transmits a LIST message to an RS in the local cluster as soon as an BGP/IDRP session with the RS is established and then whenever the local "informed client" list changes;
- receives and process BGP/IDRP updates;
- receives and processes "informed client" lists as described below:

  a) If a LIST message is received from the route server in another RS cluster, it should be silently ignored.

b) If a LIST message is received from a route server that belongs to the same RS Cluster, the differences between the old and the new list are determined and the old "informed client" list for that RS is replaced by the list from the new message. For each client that was in the old list but not in the new list it is checked whether the server has an established BGP/IDRP connection to that client and the client is not in any of the other "informed client" lists. If both conditions are met, the processing described for a new client takes place (see 4.3.3.3).

- for each new BGP/IDRP client (including connections established in Initiation state), decides if that client should become an "informed client", i.e. whether routing updates are to be sent to the client or that client has been already taken care by another RS in the local cluster. The decision process is described in the next section.

4.3.3.3 New Client Processing

Whenever an RS acquires a new BGP/IDRP peer it scans through all "informed client" lists in order to determine if this peer has already been receiving routing updates from another RS in the local RS cluster. If the identifying address of the peer is found in one of the list, no routing updates are sent to that peer.

If the peer's Router Id is not found, the route server initiates a 'DelayTimer' timer for that peer and the decision is postponed until that timer expires. The delay value is calculated as followed:

the RS determines the relative position of its own "informed client" list in the linked list of all "informed client" lists. If such position is expressed with a number, say N, in the 1 to "maximum number of lists" range, then the delay value is set to (N-1)*<DelayGranularity>.

Upon expiration of the DelayTimer, the "informed client" lists are scanned once again to see if the corresponding peer has already been receiving routing updates from another RS in the local RS cluster. If the Router Id of the peer is found in one of the lists as a result of receiving a new LIST message, no routing updates are sent to that peer. Otherwise, the peer's Router ID is entered in the "informed client" list that belongs to the RS, the transmission of the updated LIST message is immediately scheduled, and routing updates are sent to the client.

The rational for the delay is to minimize races in the decision as which RS among route servers in the same RSC is going to provide

routing information to a given client. The RS with least number of "informed clients" would have a shortest delay and is the most probable to win the race. This helps to equalize the number of "informed clients" between RSc in a cluster.

After an BGP/IDRP peer is placed in the "informed client" list, it is only removed from the list when the BGP/IDRP connection to this peer is lost. While an RS client is in the list it is accurately updated with all routing changes.

4.3.3.5 Inter-RS Connection Failure

If a route server loses a routing session with a route server in the same cluster, it must consider taking the responsibilities of route advertisement to the clients that are in the "informed client" list of the remote route server of the failed session.

For each such client it is checked whether the server has an established BGP/IDRP connection to that client and the client is not in any of the "informed client" lists of active RS. If both conditions are true, the processing described for a new client takes place (see 4.3.3.3).

After advertisement responsibilities are reconciled the "informed client" list associated with the failed session should be discarded.

4.3.4 RCID_PATH Attribute

The RCID_PATH is an optional non-transitive attribute that is composed of a sequence of RS Cluster Identifiers (RCID) that identifies the RS Cluster through which routing information carried in the UPDATE message has passed. Type Code of the RCID_PATH attribute is 254. The attribute value field contains one or more RS Cluster Identifiers, each encoded as a 2-octets long field.

When a route server propagates a route which has been learned from nother Route Server’s UPDATE message, the following is performed with respect to the the RCID_PATH attribute:

- if the destination of the route is not a route server, the RCID_PATH Attribute is excluded from the UPDATE message sent to that client.

- if the destination of the route is another route server that is located in the advertising server’s own RS cluster, the RCID_PATH attribute is sent unmodified.

- if the destination of the route is a route server in a different RS cluster, the advertising route server shall verify that the RCID of the destination speaker's cluster is not present in the RCID_PATH attribute associated with route. If it does, the route shall not be advertised and an event indicating that a route loop was detected should be logged, otherwise the advertising router shall prepend its own RCID to the RCID sequence in the RCID_PATH attribute (put it in the leftmost position).

When a route server propagates a route which has been learned from a border router to another route server then:

- if the destination of the route is a route server that is located in the advertising router's own RS cluster, an empty RCID_PATH attribute shall be included in the UPDATE message (an empty RCID_PATH attribute is one whose length field contains the value zero).

- if the destination of the route is a route server in a different RS cluster, the advertising route server shall include its own RCID in the RCID_PATH attribute. In this case, the RCID of advertising route server will be the only entry in the RCID_PATH attribute.

4.3.5 NOTIFICATION Error Codes

In addition to the error codes defined in the BGP-4/IDRP specification, the following error can be indicated in a NOTIFICATION message that is sent by a route server:

255 LIST Message Error

The following error subcodes can be associated with the LIST Message Error:

1 - Bad Address. This subcode indicates that a Client Identifying Address in the received LIST message does not represent a valid network layer address of a router interface.

The following additional UPDATE error subcodes are also defined:

255 - Invalid ADVERTISER Attribute. This subcode indicates that a value of the ADVERTISER Attribute does not represent a valid network layer address of a router interface.

4.3.7 Timers

The InitiationTimer value of 5 minutes is suggested.

In order to avoid route flaps during an RS switch-over, a value of DelayGranularity should be such so the maximum possible value of the DelayTimer (see 4.3.3.3) combined with the Hold Time of inter-RS connections would be shorter than two-third of the smallest Hold Time interval of all BGP/IDRP connections between the route servers and their clients (including RSs in other clusters). So in a cluster with three RSs and the respective Hold Times of 30 and 90 seconds the DelayGranularity of 15 seconds would be a recommended value.

For the same reason it is recommended that the Hold Time of BGP/IDRP connections between route servers in the same cluster is set to one-third of the smallest Hold Time of all BGP/IDRP connections between the route servers and their clients (including RSs in other clusters). So, if the smallest Hold Time of BGP/IDRP sessions with clients is 90 seconds, the recommended value of the Hold Time of BGP/IDRP connections between route servers in that cluster would be 30 seconds.

5. Route Server Discovery

This document does not propose any mechanism for the dynamic RS discovery by RS clients or/and by other route servers. It is assumed that at minimum a manual configuration will be provided in participating routers to achieve the needed connectivity.

7. Security Considerations

Security issues are not discussed in this document.

8. Acknowledgment

Some design concepts presented in this paper benefited from discussions with Tony Li (cisco Systems).

Author likes to thank John Krawczyk (Bay Networks) and Susan Harris (Merit) for their review and valuable comments.

Also, author would like to thank Yakov Rekhter (IBM) for the review of the earlier version of this document and constructive comments.

Special thanks to Ray Chang (Bay Networks) whose experience in implementing the concepts presented in this document helped to refine the route server design.

9. References

[BGP4] Rekhter, Y., and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, T.J. Watson Research Center, IBM Corp., cisco Systems, March 1995.

[IDRP] Rekhter, Y., and P. Traina, "IDRP for IPv6", Work In Progress.

10. Author's Address

Dimitry Haskin
Bay Networks, Inc.
2 Federal Street
Billerica, MA 01821

EMail: dhaskin@baynetworks.com