Yakov Rekhter
Cisco Systems
Joel Halpern
Institutional Venture Partners

May 1998

NHRP for Destinations off the NBMA Subnetwork

draft-ietf-ion-r2r-nhrp-03.txt

1. Status of this Memo

2. Abstract

The NBMA Next Hop Resolution Protocol (NHRP) [1] specifies a mechanism that allows a source station (e.g., a host or a router) on an NBMA subnetwork to find the NBMA subnetwork address of a destination station when the destination station is connected to the NBMA subnetwork. For the case where the destination station is off the NBMA subnetwork the mechanism described in [1] allows a node to determine the NBMA subnetwork address of an egress router from the NBMA subnetwork that is ``nearest'' to the destination station. If used to locate an egress router wherein the destination station is directly behind the egress router, the currently documented NHRP behaviors are sufficient. However, as documented elsewhere [2], there are cases where if used between routers for generalized transit, NHRP can produce loops.

This document describes extensions to the NBMA Next Hop Resolution Protocol (NHRP) [1] that allow a node to acquire and maintain the information about the egress router without constraining the destination(s) to be directly connected to the egress router.

3. CONVENTIONS

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [3].

4. NHRP Target Information

The mechanism described in this document allows a node to find an egress router for either a single destination, or a set of destinations (where the set is expressed as a single address prefix). Since a single destination is just a special case of a set of destinations, for the rest of the document we will always talk about a set of destinations, and will refer to this set as an ``NHRP target''.

The NHRP target is carried in the NHRP Request, Reply, and Purge messages as an address prefix (using the Prefix Length field of the NHRP Client Information Extension). In order to ensure correctness, a target may be replaced by an identical target with a longer prefix length. This replacement may be done at an intermediate or responding NHS. Other than this increase of prefix length, no NHS shall modify the NHRP target information in an NHRP message.

In general a router may maintain in its Forwarding Information Base (FIB) routes whose Network Layer Reachability Information (NLRI) that exhibits a subset relation. Such routes are called overlapping routes. To expand upon this, entries in a FIB are often related, with one entry being a prefix of another entry. The longer prefix therefore covers a set of routes that are a subset of the shorter prefix. To provide correct forwarding in the presence of such overlapping (or nested) routes this document constrains an NHRP target by requiring that all the destinations covered by the target must form a subset of the NLRI of at least one route in the Forwarding Information Base (FIB) of the router that either originates, or propagates an NHRP Request. That is, there must be at least one route in the FIB which is a prefix of (or equal to) the target of the request. For the rest of the document we'll refer to this as the ``first NHRP target constraint''. A station can originate an NHRP Request, and a router can propagate an NHRP Request only if the NHRP target of the Request does not violate the first

NHRP target constraint.

If a received NHRP request does not meet this ``first NHRP target constraint'' when received, the receiving router has two choices. It may answer the request, defining itself as the egress. This is compatible with the base NHRP specification, and preserves the ``first NHRP target constraint''. Alternatively, the router may lengthen the received prefix until the first constraint is met. The prefix is lengthened until the target falls within (or becomes equal to) a FIB entry.

A route (from a local FIB) whose NLRI forms a minimal superset of all the destinations covered by the NHRP target is called an ``NHRP forwarding route''. This is the longest FIB entry that covers the entire target. Observe that by definition the set of destinations covered by an NHRP target always exhibits a subset relation to the set of destinations covered by the NHRP forwarding route associated with the target.

This document further constrains origination/propagation of NHRP Requests by prohibiting the NHRP target (carried by a Request) to form a superset of the destinations covered by any of the routes in the local FIB. Remembering that there are nested FIB entries, this constraint says that there must not be a FIB entry which is itself a subset of the target of the NHRP request. If there were, there would be some destinations within the request which would be forwarded differently then others, preventing a single answer from being correct. The constraint applies both to the station that originates an NHRP Request and to the routers that propagate the Request. For the rest of the document we'll refer to this constraint as the ``second NHRP target constraint''. A station can originate an NHRP Request, and a router can propagate an NHRP Request only if the NHRP target of the Request does not violate the second NHRP target constraint. The second NHRP target constraint guarantees that forwarding to all the destinations covered by the NHRP target would be accomplished via a single (common) route, and this route would be the NHRP forwarding route for the target.

Again, if a received NHRP request does not meet the ``second NHRP target constraint'', the router may either respond to the request, providing its own NBMA address, or it may lengthen the prefix in the request so as to meet the second constraint.

5. NHRP Requester and Terminator Processing

The issue being addressed with the behaviors being mandated in this document is to ensure that sufficient information is present and processed to avoid NHRP shortcuts causing packet forwarding loops.

In order to do this, the requester and responder of the request must undertake certain work, and any "border routers" in the forwarding path must also perform certain additional work beyond checking the target consistency with the FIB during request processing. This border work suffices to detect any changes that would cause the path selection to have failed the target constraints.

The work performed by the requester and responder consists of two kinds of work. One set is requester only work, and is required in order to determine where the protocol boundaries are. The other set is the route monitoring work.

5.1. NHRP IGP information

The primary cause of NHRP forwarding loops is the loss of information at a routing protocol boundary. Normally, such boundaries are detected by the router at the boundary. However, it is possible for IGP boundaries to overlap. Therefore, NHRP requesting Routers MUST include the NHRP IGP Information extension (as defined in section 9). This extension indicates what IGP the originator of the request uses. A requesting router must always include this extension, since it is not possible to tell a priori whether the eventual resolution of the request will be a host or a router.

Because the entire BGP domain is consider one routing domain, the extension also contains an indication as to whether the originator was a BGP speaker.

5.2. NHRP Requestor and Responder monitoring

NHRP requestors and responders are required to monitor routing to maintain correct shortcut information.

Once a router that originates an NHRP Request acquires the shortcut next hop information, it is essential for the router to be able to detect any changes that would affect the correctness of this information. The following measures are intended to provide the correctness.

Both ends of a shortcut have to monitor the status of the route that

was associated with the shortcut (the NHRP forwarding route). If the status changes at the router that generated the NHRP Reply, this router should send a Purge message, so that the NHRP Requester would issue another NHRP. If the status changes at the Requester, the Requester must issue another NHRP. This ensures that when both ends of a shortcut are up, any changes in routing that impact forwarding to any of the destinations in the NHRP target would result in a revalidation (via NHRP) of the shortcut. Note that in addition to sending purges/reverifies in response to routing changes which directly effect the NHRP target, there is one other case.

A router MUST perform the appropriate purge/reverification process if it receives routing updates that cause an issued NHRP request to violate either of the target constraints defined earlier. This is possible at an NHRP originator, and is more likely at border devices.

Once a shortcut is established, the Requester needs to have some mechanism(s) to ensure that the other end of the shortcut is alive. Among the possible mechanisms are: (a) indications from the Data Link layer, (b) presence of traffic in the reverse direction that comes with the Link Layer address of the other end, (c) keepalives sent by the other end. This is intended to suppress black holes, when the next hop router in the shortcut (the router that generated Reply) goes down.

A requester should establish a shortcut only after the requester determines that the information provided by NHRP is fairly stable. This is necessary in order to avoid initiating shortcuts that are based on transients in the routing information, and thus would need to be revalidated almost immediately anyway. Thus, a router may wait to use NHRP information if the underlying routing information has recently changed. If the routing protocol being used has a notion of stability, it should be used. Information in a transient or holddown state SHOULD NOT be used, and requests which need to be processed based on such information SHOULD be discarded.

6. Border Processing of NHRP Request

Processing of an NHRP Request is covered by two sets of rules: the first set for IGP related processing, and the second set for BGP related processing. The rules for IGP processing relate to determining where the IGP borders are (in particular in the case of overlapping IGPs), and then for what must happen at said borders.

6.1. Border Determination

When a router receives a request, and determines that it is not the NBMA exit router, it must perform a series of checks before forwarding the request.

When a router receives such a Request, the router uses the NHRP target and the NHRP IGP information to check whether (a) the first and the second NHRP target constraints are satisfied, (b) the router it is in the same routing domain as the originator of the Request, and if yes, then whether (c) it is a border router for that domain.

When the NHRP target is checked against the forwarding database, a determination must be made as to whether either of the target constraints has been violated. If they are violated, then the router MAY either

- o Extend the prefix so as to meet the constraints.
- o reply to the request indicating that it is the destination
- o return an error indicating which constraint was violated.

If the NHRP forwarding route indicates a next hop that is not on the same NBMA as the interface on which the Request was received, the router sends back an NHRP Reply and terminates the query.

If a router receives a request without IGP information, then it was originated within this domain by a host. If the router is an AS Border Router (i.e. running BGP), and if the forwarding path exits the AS, then it must behave as a border router for this request. Otherwise, for requests without IGP information, the router is not a border router.

For requests with IGP information, the router compares the forwarding information against the IGP in the request. If the forwarding entry indicates that the next hop is to exit the AS (an AS Border Router), then check the BGP behaviors below.

When the IGP the next hop was learned from is the same IGP as indicated in the request, then the NHS simply forwards the request. [Of course, as per NHRP, it is free to respond indicating it is the termination of the shortcut, for example when the Router/NHS is a firewall.]

When the IGP the next hop was learned from is different from that listed in the NHRP request, then this NHS is a border router for this request.

6.2. Border Behavior

In all cases, a border router has two choices. It MAY terminate and respond to the request, responding with its IP and NBMA address.

Alternatively, it MAY perform border propagation.

6.2.1. Reorigination

Upon receiving an NHRP request for which the NHS is a border router, if it chooses to propagate the request, it MUST originate a new NHRP request. This request will have a locally generated request identifier, and the same NHRP target information as in the received request. The NHRP IGP Information will be the correct indication for the outgoing interface, with BGP indication if the received request had the BGP indication, or if this transition crosses the AS border. All other extensions are copied from the incoming request to the new request.

6.2.2. Response Propagation

When an NHRP response is received for a propagated request, the information is copies from the received request, and passed on in a new NHRP response, responding to the originally received request. The prefix length in the received response is copied to the new response. All extensions except the NHRP IGP Information are copied to the new response.

In addition, the border router saves state about this information exchange. The saved state includes the NHRP target from the response, with the NHRP prefix length that resulted from the exchange. It also includes the both the original requester, and the identity of the responder. These are used to generate appropriate reverification and purges whenever routing changes in a way that could effect the resolution.

6.3. Border Information

Sometimes the routing protocol will have provided the border router with enough information to generate a response to an incoming NHRP request. In particular, the border router may have information about IP prefix to NBMA address bindings. If such information is present, it may be used by a border router to produce an NHRP response without actually propagating the request. In such a case, that information must be monitored for stability to maintain the correctness of the shortcut.

7. BGP Operation

While the NHRP mechanism described above is mostly constrained to the routers within a single routing domain, the same mechanisms can be used for shortcuts that span multiple domains. In doing so, one wants to produce as little additional overhead in the BGP space as possible.

Therefore, we will treat the space over which BGP runs as a single routing domain. Care must be taken to propagate information across the individual AS without error, and to indicate that one has properly entered the BGP space.

Additional complexity in handling multi-domain shortcuts arise if routing information gets aggregated at the border routers (which certainly happens in practice). Since BGP is the major protocol that is used to exchange routing information across multiple routing domains, we'll restrict our proposal to the case where the routing information exchange across domains' boundaries is controlled by BGP.

If both the source and the destination domains are on a common NBMA network, and the path between these two domains is also fully within the same NBMA network, then we have only three routing domains to deal with: source routing domain, BGP routing domain, and destination routing domain. If the destination domain is not on the same NBMA as the source domain, then we need to deal only with two domains - the source and the BGP. Note that we treat all routers that participate in a single (common) instance of BGP as a single BGP routing domain, even if these routers participate in different intra-domain routing protocols, or in different instances of the same intra-domain routing protocol. There are three aspects to consider.

(a) how a border router in the domain that the originator of the Request is in handles the Request (crossing IGP/BGP boundary),

(b) how the Request is handled across the BGP domain, and finally

(c) how a border router in the domain where the NHRP target is in handles the Request (crossing BGP/IGP boundary).

7.1. Handling NHRP Request at the source domain border router

When a border router receives an NHRP Request originated from within its own (IGP) routing domain, the border router determines the NHRP forwarding route for the NHRP target carried by the Request. If the router already has the shortcut information for the forwarding route, then the router uses this information to construct a Reply to the source of the NHRP Request. Otherwise, the router originates its own NHRP Request. The Request contains exactly the same NHRP target, as was carried by the original Request; The NHRP IGP Information will indicate that the request was generated by BGP, and will indicate the IGP of the BGP AS being entered. While it is assumed that a BGP transit AS will generally use only one IGP, the IGP information (and border processing) is included to allow all cases. The newly originated Request is sent to the next hop of the NHRP forwarding route. Once the border router receives a Reply to its own Request, the border router uses the next hop information from the Reply to construct its own Reply to the source of the original NHRP Request.

If the border router later on receives a Purge message for the NHRP forwarding route, the border router treats this event as if there was a local change in the NHRP forwarding route (even if the there was no changes in the route).

This is exactly the same behavior as all other border cases, and is described here for completeness.

7.2. Handling NHRP Request within the BGP domain

Routers within an AS will check the IGP, and perform appropriate processing based on the IGP match. In general, this will result in normal forwarding of the NHRP request.

Therefore, the significant cases occur at the BGP speaking routers. There are two conditions to check for, early exit of the NBMA, and reachability aggregation. Both of these conditions apply to Autonomous systems that do not contain the NHRP target.

7.2.1. NBMA exit

The BGP router in deciding where to send the NHRP request will determine what the correct exit from the autonomous system is. It will determine if that exit is within the NBMA. If it is not within the NBMA, then the router MUST respond to the NHRP request, indicating its own IP and NBMA addresses as the correct termination of the shortcut. This is because the actual NBMA border device is not in a position to monitor the topology properly.

BGP routers within an NBMA which are supporting R2R NHRP SHOULD be configured to know where the NBMA border is. In the absence of such configuration, requests from other router SHOULD be terminated at the BGP router, since it can not tell what will be crossing the border. A BGP router supporting R2R NHRP may be configured to assume that all of its neighbors are within the NBMA, and therefore not perform such early termination.

7.2.2. Reachability Aggregation

BGP routers aggregate reachability. If the router aggregates reachability that includes the NHRP target, only this router has the visibility to some of the topology changes that can affect the correctness of the route. Therefore, this router is a border router for this NHRP request.

It must originate a new request, place the correct information in the request, receive the response, and generate the correct response towards the requester. This aggregating router must also monitor routing in case of changes which affect the request.

If the router later on receives a Purge message for the NHRP forwarding route, the router treats this event as if there was a change in the NHRP forwarding route (even if the there was no changes in the route).

It should be noted that this conditions applies if the router COULD aggregate relevant routing information, even if it currently does not.

7.3. Handling NHRP Request at the destination domain border router

When a border router receives an NHRP Request from a BGP speaker, and the border router determines that all the destinations covered by the NHRP target of the Request are within the (IGP) domain of that border router, the border router determines the NHRP forwarding route for the NHRP target carried by the Request. The newly formed Request contains exactly the same NHRP target as the received Request; the NHRP IGP Information indicates the IGP this router is using to select the route to the destination. The newly originated Request is sent to the next hop of the NHRP forwarding route. Once the border router receives a Reply to its own Request, the border router uses the next hop information from the Reply to construct its own Reply to the source of the original NHRP Request.

If the border router later on receives a Purge message for the NHRP forwarding route, the border router treats this event as if there was a change in the NHRP forwarding route (even if the there was no changes in the route).

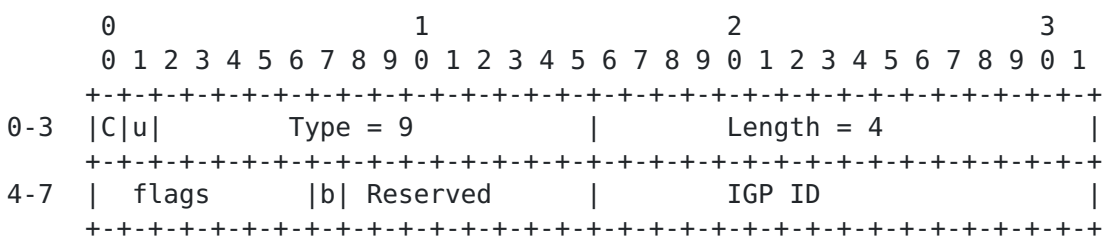
8. More state, less messages

It should be possible to reduce the number of Purge messages and subsequent NHRP messages (caused by the Purge messages) by maintaining more state on the border routers at the source and destination domains, and the BGP routers that perform aggregation along the path from the source to the destination.

Specifically, on these routers it would be necessary to keep the information about all the NHRP targets for which the routers maintain the shortcut information. This way when such a router determines that the NHRP forwarding route (for which the router maintains the shortcut information) changes due to some local routing changes, the router could check whether these local changes impact forwarding to the destinations covered by the NHRP targets. For the targets that are impacted by the changes the router would send Purge messages.

Note that this mechanism (maintaining NHRP targets) precludes the use of Address Prefix Extension - the shortcut will be determined only for the destinations covered by the NHRP target (so, if the target is a single IP address, then the shortcut would be determined only for this address).

9. NHRP IGP Information Extension Format



C "Compulsory." If clear, and the NHS does not recognize the type code, the extension maybe safely be ignored. For the IGP Information extension, this bit is clear.

u Unused and must be set to zero

Type The extension type code. For the IGP Information extension, this is 9.

Length the length in octets of the value. For this extension, this is 4.

flags Other than the "b" flag, these are reserved, SHALL be set to 0 on transmission, and SHALL be ignored on reception.

b This flag indicates whether the request (or a predecessor thereof) was originated by a BGP speaker. Set (to 1) to indicate that the BGP speaker has operated on this. Clear (to 0) if not.

IGP ID This field indicates the IGP used by the request originator. The currently defined values are:

1 = RIP
2 = RIPv2
3 = OSPF
4 = Dual IS-IS

10. IANA Considerations

This document defines an enumerated field for identifying IGPs in router-to-router NHRP requests. Since there may be additional IGPs in use, a procedure is needed for allocating additional values. The IANA shall allocate values for this field as needed. Specifically, when requested a value shall be allocated for an IGP for any layer 3 protocol for which there is a clear and stable definition of the protocol. An RFC is the best example of such stability. Vendor published specifications are also acceptable. The IANA should avoid issuing two values for the same protocol. However, it is not incumbent upon the IANA to determine if two similar protocols are actually the same.

11. Open issues

The mechanisms described in this document assume that certain routers along a path taken by an NHRP Request would be required to maintain state associated with the NHRP forwarding route associated with the NHRP target carried by the Request. However, it is quite clear that the router(s) may also lose this state. Further study of the impact of losing the state is needed before advancing the use of NHRP for establishing shortcuts among routers beyond Proposed Standard.

The mechanisms described in this document may result in a situation where a router would be required to maintain NHRP peering with potentially a fairly large number of other routers. Further study is needed to understand the implications of this on the scalability of the approach where NHRP is used to establish shortcuts among routers.

This document doesn't have a proof that the mechanisms described here result in loop-free steady state forwarding when NHRP is used to establish shortcuts among routers, however, a counterexample has not yet been found. Further analysis should be done as part of advancing beyond Proposed Standard.

12. Security Considerations

Security is provided in the base NHRP protocol, using hop-by-hop authentication. There is no change to the fundamental security capabilities provided therein when these extensions are used. It should be noted that the assumption of transitive trust that is the basis of such security may well be significantly weaker in an inter-domain environment, and administrators of border routers should take this into consideration. The hop-by-hop security model is used by NHRP originally because there is no end-to-end security association between the requesting and responding NHRP entities. In this environment there is the additional facet that intermediate NHS are modifying the prefix length field of the CIE, thus changing the end-to-end information.

13. References

[1] J. Luciani, D. Katz, D. Piscitello, B. Cole, N. Doraswamy., "NBMA Next Hop Resolution Protocol", RFC-2332, USC/Information Sciences Institute, April 1998.

[2] D. Cansever., "NHRP Protocol Applicability Statement", RFC-2333, USC/Information Sciences Institute, April 1998

[3] S. Bradner., "Key words for use in RFCs to Indicate Requirement Levels", RFC-2119, USC/Information Sciences Institute, March 1997.

14. Acknowledgements

The authors wish to Thank Curtis Villamizer for his contributions emphasizing both the importance of the looping cases, and some examples of when loops can occur.

15. Author Information

Joel M. Halpern
Institutional Venture Partners
3000 Sand Hill Road
Menlo Park, CA
Phone: (650) 926-5633
email: joel@mcquillan.com

Yakov Rekhter
cisco Systems, Inc.
170 Tasman Dr.
San Jose, CA 95134
Phone: (914) 528-0090
email: yakov@cisco.com