Internet Engineering Task Force (IETF)
Request for Comments: 7048
Updates: 4861
Category: Standards Track
ISSN: 2070-1721
January 2014

E. Nordmark
Arista Networks
I. Gashinsky
Yahoo!

Neighbor Unreachability Detection Is Too Impatient

Abstract

IPv6 Neighbor Discovery includes Neighbor Unreachability Detection. That function is very useful when a host has an alternative neighbor -- for instance, when there are multiple default routers -- since it allows the host to switch to the alternative neighbor in a short time. By default, this time is 3 seconds after the node starts probing. However, if there are no alternative neighbors, this timeout behavior is far too impatient. This document specifies relaxed rules for Neighbor Discovery retransmissions that allow an implementation to choose different timeout behavior based on whether or not there are alternative neighbors. This document updates RFC 4861.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction 2
2. Definition of Terms 3
3. Protocol Updates 3
4. Example Algorithm 6
5. Acknowledgements 7
6. Security Considerations 8
7. References 8
   7.1. Normative References 8
   7.2. Informative References 8

1. Introduction

IPv6 Neighbor Discovery [RFC4861] includes Neighbor Unreachability Detection (NUD), which detects when a neighbor is no longer reachable. The timeouts specified for NUD are very short (by default, three transmissions spaced one second apart). These short timeouts can be appropriate when there are alternative neighbors to which the packets can be sent -- for example, if a host has multiple default routers in its Default Router List or if the host has a Neighbor Cache Entry (NCE) created by a Redirect message. In those cases, when NUD fails, the host will try the alternative neighbor by redoing the next-hop selection. That implies picking the next router in the Default Router List or discarding the NCE created by a Redirect message, respectively.

The timeouts specified in [RFC4861] were chosen to be short in order to optimize scenarios where alternative neighbors are available.

However, when there is no alternative neighbor, there are several benefits to making NUD probe for a longer time. One benefit is to make NUD more robust against transient failures, such as spanning tree reconvergence and other layer 2 issues that can take many seconds to resolve. Marking the NCE as unreachable, in that case, causes additional multicast on the network. Assuming there are IP packets to send, the lack of an NCE will result in multicast Neighbor Solicitations being sent (to the solicited-node multicast address) every second instead of the unicast Neighbor Solicitations that NUD sends.

As a result, IPv6 Neighbor Discovery is operationally more brittle than the IPv4 Address Resolution Protocol (ARP). For IPv4, there is no mandatory time limit on the retransmission behavior for ARP [RFC0826], which allows implementors to pick more robust schemes.

The following constant values in [RFC4861] seem to have been made part of IPv6 conformance testing: MAX_MULTICAST_SOLICIT, MAX_UNICAST_SOLICIT, and RETRANS_TIMER. While such strict conformance testing seems consistent with [RFC4861], it means that the standard needs to be updated to allow IPv6 Neighbor Discovery to be as robust as ARP.

This document updates RFC 4861 to relax the retransmission rules.

Additional motivations for making IPv6 Neighbor Discovery more robust in the face of degenerate conditions are covered in [RFC6583].

2. Definition of Terms

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 [RFC2119].

3. Protocol Updates

Discarding the NCE after three packets spaced one second apart is only needed when an alternative neighbor is available, such as an additional default router or discarding an NCE created by a Redirect.

If an implementation transmits more than MAX_UNICAST_SOLICIT/ MAX_MULTICAST_SOLICIT packets, then it SHOULD use the exponential backoff of the retransmit timer. This is to avoid any significant load due to a steady background level of retransmissions from implementations that retransmit a large number of Neighbor Solicitations (NS) before discarding the NCE.

Even if there is no alternative neighbor, the protocol needs to be able to handle the case when the link-layer address of the neighbor/ target has changed by switching to multicast Neighbor Solicitations at some point in time.

In order to capture all the cases above, this document introduces a new UNREACHABLE state in the conceptual model described in [RFC4861]. An NCE in the UNREACHABLE state retains the link-layer address, and IPv6 packets continue to be sent to that link-layer address. But in the UNREACHABLE state, the NUD Neighbor Solicitations are multicast (to the solicited-node multicast address), using a timeout that follows an exponential backoff.

In the places where [RFC4861] says to discard/delete the NCE after N probes (Sections 7.3 and 7.3.3, and Appendix C), this document instead specifies a transition to the UNREACHABLE state.

If the Neighbor Cache Entry was created by a Redirect message, a node MAY delete the NCE instead of changing its state to UNREACHABLE. In any case, the node SHOULD NOT use an NCE created by a Redirect to send packets if that NCE is in the UNREACHABLE state. Packets should be sent following the next-hop selection algorithm in [RFC4861], Section 5.2, which disregards NCEs that are not reachable.

Section 6.3.6 of [RFC4861] indicates that default routers that are "known to be reachable" are preferred. For the purposes of that section, if the NCE for the router is in the UNREACHABLE state, it is not known to be reachable. Thus, the particular text in Section 6.3.6 that says "in any state other than INCOMPLETE" needs to be extended to say "in any state other than INCOMPLETE or UNREACHABLE".

Apart from the use of multicast NS instead of unicast NS, and the exponential backoff of the timer, the UNREACHABLE state works the same as the current PROBE state.

A node MAY garbage collect a Neighbor Cache Entry at any time as specified in [RFC4861]. This freedom to garbage collect does not change with the introduction of the UNREACHABLE state in the conceptual model. An implementation MAY prefer garbage collecting UNREACHABLE NCEs over other NCEs.

There is a non-obvious extension to the state-machine description in Appendix C of [RFC4861] in the case for "NA, Solicited=1, Override=0. Different link-layer address than cached". There we need to add "UNREACHABLE" to the current list of "STALE, PROBE, Or DELAY". That is, the NCE would be unchanged. Note that there is no corresponding change necessary to the text in [RFC4861], Section 7.2.5, since it is phrased using "Otherwise" instead of explicitly listing the three states.

The other state transitions described in Appendix C handle the introduction of the UNREACHABLE state without any change, since they are described using "not INCOMPLETE".

There is also the more obvious change already described above. [RFC4861] has this:

State	Event	Action	New state
PROBE	Retransmit timeout, N or more retransmissions.	Discard entry	-

That needs to be replaced by:

State	Event	Action	New state
PROBE	Retransmit timeout, N retransmissions.	Increase timeout Send multicast NS	UNREACHABLE
UNREACHABLE	Retransmit timeout	Increase timeout Send multicast NS	UNREACHABLE

The exponential backoff SHOULD be clamped at some reasonable maximum retransmit timeout, such as 60 seconds (see MAX_RETRANS_TIMER below). If there is no IPv6 packet sent using the UNREACHABLE NCE, then it is RECOMMENDED to stop the retransmits of the multicast NS until either the NCE is garbage collected or there are IPv6 packets sent using the NCE. The multicast NS and associated exponential backoff can be applied on the condition of continued use of the NCE to send IPv6 packets to the recorded link-layer address.

A node can unicast the first few Neighbor Solicitation messages even while in the UNREACHABLE state, but it MUST switch to multicast Neighbor Solicitations within 60 seconds of the initial retransmission to be able to handle a link-layer address change for the target. The example below shows such behavior.

4. Example Algorithm

This section is NOT normative but specifies a simple implementation that conforms with this document. The implementation is described using operator-configurable values that allow it to be configured to be compatible with the retransmission behavior in [RFC4861]. The operator can configure the values for MAX_UNICAST_SOLICIT, MAX_MULTICAST_SOLICIT, RETRANS_TIMER, and the new BACKOFF_MULTIPLE, MAX_RETRANS_TIMER, and MARK_UNREACHABLE. This allows the implementation to be as simple as:

next_retrans = ($BACKOFF_MULTIPLE ^ $solicit_retrans_num) * $RetransTimer * $JitterFactor where solicit_retrans_num is zero for the first transmission, and JitterFactor is a random value between MIN_RANDOM_FACTOR and MAX_RANDOM_FACTOR [RFC4861] to avoid any synchronization of transmissions from different hosts.

After MARK_UNREACHABLE transmissions, the implementation would mark the NCE UNREACHABLE and as a result explore alternate next hops. After MAX_UNICAST_SOLICIT, the implementation would switch to multicast NUD probes.

The behavior of this example algorithm is to have 5 attempts, with time spacing of 0 (initial request), 1 second later, 3 seconds after the first retransmission, then 9, then 27, and switch to UNREACHABLE after the first three transmissions. Thus, relative to the time of the first transmissions, the retransmissions would occur at 1 second, 4 seconds, 13 seconds, and finally 40 seconds. At 4 seconds from the first transmission, the NCE would be marked UNREACHABLE. That behavior corresponds to:

MAX_UNICAST_SOLICIT=5

RETRANS_TIMER=1 (default)

MAX_RETRANS_TIMER=60

BACKOFF_MULTIPLE=3

MARK_UNREACHABLE=3

After 3 retransmissions, the implementation would mark the NCE UNREACHABLE. That results in trying an alternative neighbor, such as another default router, or ignoring an NCE created by a Redirect as specified in [RFC4861]. With the above values, that would occur after 4 seconds following the first transmission compared to the

2 seconds using the fixed scheme in [RFC4861]. That additional delay is small compared to the default ReachableTime of 30,000 milliseconds.

After 5 transmissions, i.e., 40 seconds after the initial transmission, the example behavior is to switch to multicast NUD probes. In the language of the state machine in [RFC4861], that corresponds to the action "Discard entry". Thus, any attempts to send future packets would result in sending multicast NS packets. An implementation MAY retain the backoff value as it switches to multicast NUD probes. The potential downside of deferring switching to multicast is that it would take longer for NUD to handle a change in a link-layer address, i.e., the case when a host or a router changes its link-layer address while keeping the same IPv6 address. However, [RFC4861] says that a node MAY send unsolicited NS to handle that case, which is rather infrequent in operational networks. In any case, the implementation needs to follow the "SHOULD" in Section 3 to switch to multicast solutions within 60 seconds after the initial transmission.

If BACKOFF_MULTIPLE=1, MARK_UNREACHABLE=3, and MAX_UNICAST_SOLICIT=3, you would get the same behavior as in [RFC4861].

If the request was not answered at first -- due, for example, to a transitory condition -- an implementation following this algorithm would retry immediately and then back off for progressively longer periods. This would allow for a reasonably fast resolution time when the transitory condition clears.

Note that RetransTimer and ReachableTime are by default set from the protocol constants RETRANS_TIMER and REACHABLE_TIME but are overridden by values advertised in Router Advertisements as specified in [RFC4861]. That remains the case even with the protocol updates specified in this document. The key values that the operator would configure are BACKOFF_MULTIPLE, MAX_RETRANS_TIMER, MAX_UNICAST_SOLICIT, and MAX_MULTICAST_SOLICIT.

It is useful to have a maximum value for ($BACKOFF_MULTIPLE^$solicit_attempt_num)*$RetransTimer so that the retransmissions are not too far apart. The above value of 60 seconds for this MAX_RETRANS_TIMER is consistent with DHCPv6.

5. Acknowledgements

The comments from Thomas Narten, Philip Homburg, Joel Jaeggli, Hemant Singh, Tina Tsou, Suresh Krishnan, and Murray Kucherawy have helped improve this document.

6. Security Considerations

Relaxing the retransmission behavior for NUD is believed to have no impact on security. In particular, it doesn't impact the application of Secure Neighbor Discovery [RFC3971].

7. References

7.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC3971] Arkko, J., Kempf, J., Zill, B., and P. Nikander, "SEcure Neighbor Discovery (SEND)", RFC 3971, March 2005.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

7.2. Informative References

[RFC0826] Plummer, D., "Ethernet Address Resolution Protocol: Or converting network protocol addresses to 48.bit Ethernet address for transmission on Ethernet hardware", STD 37, RFC 826, November 1982.

[RFC6583] Gashinsky, I., Jaeggli, J., and W. Kumari, "Operational Neighbor Discovery Problems", RFC 6583, March 2012.

Authors' Addresses

Erik Nordmark
Arista Networks
Santa Clara, CA
USA

EMail: nordmark@acm.org

Igor Gashinsky
Yahoo!
45 W 18th St
New York, NY
USA

EMail: igor@yahoo-inc.com