6MAN WG
Internet-Draft
Updates: 4861 (if approved)
Expires: January 2, 2013
E. Nordmark
Cisco Systems, Inc.
I. Gashinsky
Yahoo!
Jul 2012

# Neighbor Unreachability Detection is too impatient
## draft-ietf-6man-impatient-nud-02.txt

Abstract

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

Table of Contents

## 1. Introduction

IPv6 Neighbor Discovery [RFC4861] includes Neighbor Unreachability Detection (NUD), which detects when a neighbor is no longer reachable.The timeouts specified are very short (by default three transmissions spaced one second apart). That can be appropriate when there are alternative paths over 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. The effect of NUD reporting a failure in those cases is that the host will try the alternative; the next router in the Default Router List, or discard the NCE which will also send using a different router.

For that reason the timeouts in [RFC4861] were chosen to be short; this ensures that if a default router fails the host can use the next router in less than 45 seconds (taking into account a default ReachableTime of 30 seconds and the time spent in the DELAY state).

However, when there is no alternative there are several benefits in making NUD try probing for a longer time. One of those benefits is to be 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 every second instead of the unicast Neighbor Solicitations that NUD sends.

As a result IPv6 Neighbor Discovery is operationally more brittle than IPv4 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 we need to update the standard if we want 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

Giving up after three packets spaced one second apart is only REQUIRED when there is an alternative, such as an additional default route or a redirect.

If implementations transmit more than MAX_*CAST_SOLICIT packets it SHOULD use (binary) exponential backoff of the retransmit timer. This is to avoid any significant load due to a steady background level of retransmissions from implementations that try for a long time.

Even if there is no alternative, the protocol needs to be able to handle the case when the link-layer address of the destination 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]. A 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, using a timeout that follows a (binary) exponential backoff.

In the places where RFC4861 says to to discard/delete the NCE after N probes (Section 7.3, 7.3.3 and Appendix C) we will instead transition to the UNREACHABLE state.

If the Neighbor Cache Entry was created by a redirect, 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 unreachable state. Packets should be sent following the next-hop selection algorithm in section 5.2 in [RFC4861] which disregards NCEs that are not reachable.

The default router selection in section 6.3.6 says to prefer default routers that are "known to be reachable". For the purposes of that section, if the NCE for the router is in UNREACHABLE state, it is not known to be reachable. Thus the particular text in section 6.3.6 which 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 (binary) 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 as any time as specified in RFC 4861. This does not change with the introduction of the UNREACHABLE state in the conceptual model.

The UNREACHABLE state is conceptual and not a required part of this specification. A node merely needs to satisfy the externally observable behavior of this specification.

There is a non-obvious extension to the state machine description in Appendix C in RFC 4861 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 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. RFC 4861 has this:

```
PROBE           Retransmit timeout,     Discard entry         -
                N or more
                retransmissions.
```

That needs to be replaced by:

```
PROBE           Retransmit timeout,     Double timeout    UNREACHABLE
                N or more               Send multicast NS
                retransmissions.

UNREACHABLE     Retransmit timeout      Double timeout    UNREACHABLE
                                        Send multicast NS
```

The binary exponential backoff SHOULD be clamped at some reasonable maximum retransmit timeout, such as 60 seconds. If there is no IPv6 packets sent using the UNREACHABLE NCE, then it makes sense 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 binary exponential backoff can be applied on the condition of the continued use of the NCE to send IPv6 packets to the recorded link-layer address.

A node MAY unicast the first few Neighbor Solicitation messages while in UNREACHABLE state, but it MUST switch to multicast Neighbor Solicitations. Otherwise it would not detect a link-layer address change for the target.

## 4. Example Algorithm

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

next_retrans = ($BACKOFF_MULTIPLE^$solicit_attempt_num)*$RetransTimer + jittered value.

After MARK_UNREACHABLE retransmissions the implementation would mark the NCE UNREACHABLE and switch to multicast NUD probes.

The recommended behavior is to have 5 attempts, with timing spacing of 0 (initial request), 1 second later, 3 seconds later, then 9, and then 27, and switch to UNREACHABLE after 3 transmissions, which represents:

MAX_UNICAST_SOLICIT=5

RETRANS_TIMER=1 (default)

BACKOFF_MULTIPLE=3

MARK_UNREACHABLE=3

After 3 retransmissions the implementation would mark the NCE UNREACHABLE and switch to multicast NUD probes. Thus we enter UNREACHABLE, and try any available alternative, after 4 seconds compared to the current 2 seconds. That additional delay is small compared to the default 30 seconds ReachableTime.

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

An Implementation following this algorithm would, if the request was not answered at first due for example to a transitory condition, 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_UNICAST_SOLICIT and MAX_MULTICAST_SOLICIT.

It would be useful to have a maximum value for ($BACKOFF_MULTIPLE^$solicit_attempt_num)*$RetransTimer so that the retransmissions are not too far apart. A value 60 seconds is consistent with DHCP.

## 5. Acknowledgements

The comments from Thomas Narten, Philip Homburg, and Joel Jaeggli have helped improve this draft.

## 6. Security Considerations

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

## 7. IANA Considerations

This are no IANA considerations for this document.

## 8. References

### 8.1. 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.

## 8.2. 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
Cisco Systems, Inc.
510 McCarthy Blvd.
Milpitas, CA, 95035
USA

Phone: +1 408 527 6625
Email: nordmark@cisco.com

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

Email: igor@yahoo-inc.com