

Network Working Group
Request for Comments: 5508
BCP: 148
Category: Best Current Practice

P. Srisuresh
Kazeon Systems
B. Ford
MPI-SWS
S. Sivakumar
Cisco Systems
S. Guha
Cornell U.
April 2009

NAT Behavioral Requirements for ICMP

Status of This Memo

This document specifies an Internet Best Current Practices for the Internet Community, and requests discussion and suggestions for improvements. Distribution of this memo is unlimited.

Copyright Notice

Copyright (c) 2009 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents in effect on the date of publication of this document (<http://trustee.ietf.org/license-info>). Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Abstract

This document specifies the behavioral properties required of the Network Address Translator (NAT) devices in conjunction with the Internet Control Message Protocol (ICMP). The objective of this memo is to make NAT devices more predictable and compatible with diverse application protocols that traverse the devices. Companion documents provide behavioral recommendations specific to TCP, UDP, and other protocols.

Table of Contents

1. Introduction and Scope	3
2. Terminology	4
3. ICMP Query Handling	6
3.1. ICMP Query Mapping	6
3.2. ICMP Query Session Timeouts	7
4. ICMP Error Forwarding	8
4.1. ICMP Error Payload Validation	8
4.2. ICMP Error Packet Translation	10
4.2.1. ICMP Error Packet Received from the External Realm ..	11
4.2.2. ICMP Error Packet Received from the Private Realm ..	13
4.3. NAT Sessions Pertaining to ICMP Error Payload	15
5. Hairpinning Support for ICMP Packets	16
6. Rejection of Outbound Flows Disallowed by NAT	17
7. Conformance to RFC 1812	17
7.1. IP Packet Fragmentation	19
7.1.1. Generating "Packet Too Big" ICMP Error Message	19
7.1.2. Forwarding "Packet Too Big" ICMP Error Message	20
7.2. Time Exceeded Message	20
7.3. Source Route Options	20
7.4. Address Mask Request/Reply Messages	20
7.5. Parameter Problem Message	21
7.6. Router Advertisement and Solicitations	21
7.7. DS Field Usage	21
8. Non-QueryError ICMP Messages	22
9. Summary of Requirements	22
10. Security Considerations	25
11. Acknowledgements	26
12. References	27
12.1. Normative References	27
12.2. Informative References	27

1. Introduction and Scope

As pointed out in RFC 3424 [UNSAF], NAT implementations vary widely in terms of how they handle different traffic. The purpose of this document is to define a specific set of requirements for NAT behavior with regard to ICMP messages. The objective is to reduce the unpredictability and brittleness the NAT devices (NATs) introduce. This document is an adjunct to [BEH-UDP], [BEH-TCP], and other protocol-specific BEHAVE document(s) in the future that define requirements for NATs when handling protocol-specific traffic.

The requirements of this specification apply to traditional NATs as described in [NAT-TRAD]. A traditional NAT has two variations, namely Basic NAT and Network Address Port Translator (NAPT). Of these, NAPT is by far the most commonly deployed NAT device. NAPT allows multiple private hosts to share a single public IP address simultaneously.

This document only covers the ICMP aspects of NAT traversal, specifically the traversal of ICMP Query messages and ICMP Error messages. Traditional NAT inherently mandates firewall-like filtering behavior [BEH-UDP]. However, firewall functionality in general or any other middlebox functionality is out of the scope of this document.

In some cases, ICMP message traversal behavior on a NAT device may be overridden by local administrative policies. Some administrators may choose to entirely prohibit forwarding of ICMP Error messages across a NAT device. Some others may choose to prohibit ICMP-Query-based applications across a NAT device. These are local policies and not within the scope of this document. For this reason, some of the ICMP requirements listed in the document are preceded with a constraint of local policy permitting.

This document focuses strictly on the behavior of the NAT device, and not on the behavior of applications that traverse NATs. Application designers may refer to [BEH-APP] and [ICE] for recommendations and guidelines on how to make applications work robustly over NATs that follow the requirements specified here and the adjunct protocol-specific BEHAVE documents.

Per [RFC1812], ICMP is a control protocol that is considered to be an integral part of IP, although it is architecturally layered upon IP -- it uses IP to carry its data end-to-end. As such, many of the ICMP behavioral requirements discussed in this document apply to all IP protocols.

In case a requirement in this document conflicts with protocol-specific BEHAVE requirement(s), protocol-specific BEHAVE documents will take precedence. The authors are not aware of any conflicts between this and any other IETF document at the time of this writing.

Section 2 describes the terminology used throughout the document. Section 3 is focused on requirements concerning ICMP-Query-based applications traversing a NAT device. Sections 4 and 5 describe requirements concerning ICMP Error messages traversing a NAT device. Section 6 describes requirements concerning ICMP Error messages generated by a NAT device. Section 7 reviews RFC 1812 conformance requirements and applicability to NATs when handling ICMP messages. Section 8 reviews a requirement for ICMP messages that are neither ICMP Query nor ICMP Error kind. Section 9 summarizes all the requirements in one place. Section 10 has a discussion on security considerations.

2. Terminology

Definitions for the majority of the NAT terms used throughout the document may be found in [NAT-TERM] and [BEH-UDP].

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 [RFC2119].

The term "Realm" is adapted from [NAT-TERM] and is defined as follows. "Realm" is often interchanged for "network domain" or simply "network" throughout the document.

Address realm or Realm - An address realm is a network domain in which the network addresses are uniquely assigned to entities such that datagrams can be routed to them. Routing protocols used within the network domain are responsible for finding routes to entities given their network addresses. Note that this document is limited to describing NAT in the IPv4 environment and does not address the use of NAT in other types of environments (e.g., the IPV6 environment).

The term "NAT Session" is adapted from [NAT-MIB] and is defined as follows:

NAT Session - A NAT session is an association between a session as seen in the private realm and a session as seen in the public realm, by virtue of NAT translation. If a session in the private realm were to be represented as (PrivateSrcAddr, PrivateDstAddr, TransportProtocol, PrivateSrcPort, PrivateDstPort) and the same session in the public realm were to be represented as (PublicSrcAddr, PublicDstAddr, TransportProtocol, PublicSrcPort, PublicDstPort), the

NAT session would provide the translation glue between the two session representations. NAT sessions in the document are restricted to sessions based on TCP, UDP, and ICMP. In the future, NAT sessions may be extended to be based on other transport protocols such as Stream Control Transmission Protocol (SCTP), UDP-lite, and Datagram Congestion Control Protocol (DCCP).

ICMP Message Classification - Section 3.2.2 of [RFC1122] and Section 4.3.1 of [RFC1812] broadly group ICMP messages into two main categories, namely "ICMP Query" messages and "ICMP Error" messages. All ICMP Error messages listed in RFC 1122 and RFC 1812 contain part of the Internet datagram that elicited the ICMP error. All the ICMP Query messages listed in RFC 1122 and RFC 1812 contain an "Identifier" field, which is referred to in this document as the "Query Identifier". There are however ICMP messages that do not fall into either of these two categories. We refer to them as "Non-QueryError ICMP Messages". All three ICMP message classes are described as follows:

- o ICMP Query Messages - ICMP Query messages are characterized by an Identifier field in the ICMP header. The Identifier field used by the ICMP Query messages is also referred to as "Query Identifier" or "Query Id", for short throughout the document. A Query Id is used by Query senders and responders as the equivalent of a TCP/UDP port to identify an ICMP Query session. ICMP Query messages include ICMP messages defined after RFC 1122 or RFC 1812 (for example, Domain Name Request/Reply ICMP messages defined in RFC 1788), as they include request/response pairs and contain an "Identifier" field.
- o ICMP Error Messages - ICMP Error messages provide signaling for IP. All ICMP Error messages are characterized by the fact that they embed the original datagram that triggered the ICMP Error message. The original datagram embedded within the ICMP Error payload is also referred to as the "Embedded packet" throughout the document. Unlike ICMP Query messages, ICMP Error messages do not have a Query Id in the ICMP header.
- o Non-QueryError ICMP Messages - ICMP messages that do not fall under either of the above two classes are referred to as "Non-QueryError ICMP Messages" throughout the document. For example, Router Discovery ICMP messages [RFC1256] are "request/response" type ICMP messages. However, they are not characterized as ICMP Query messages in this document as they do not have an "Identifier" field within the messages. Likewise, there are other ICMP messages defined in [RFC4065] that do not fall in either of the ICMP Query or ICMP Error message categories, but will be referred to as Non-QueryError ICMP messages.

The reason for categorizing ICMP messages for NAT behavioral properties is that each category has different characteristics used for mapping (i.e., the Query Id and the Embedded datagram), which leaves the Non-QueryError ICMP messages in a separate, distinctive group.

3. ICMP Query Handling

This section lists the behavioral requirements for a NAT device when processing ICMP Query packets. The following subsections discuss requirements specific to ICMP Query handling in detail.

3.1. ICMP Query Mapping

Unless explicitly overridden by local policy, a NAT device MUST permit ICMP Queries and their associated responses, when the Query is initiated from a private host to the external hosts. ICMP Query mapping by NAT devices is necessary for current ICMP-Query-based applications to work. This entails a NAT device to transparently forward ICMP Query packets initiated from the nodes behind NAT, and the responses to these Query packets in the opposite direction. As specified in [NAT-TRAD], this requires translating the IP header. A NAT device further translates the ICMP Query Id and the associated checksum in the ICMP header prior to forwarding.

NAT mapping of ICMP Query Identifiers SHOULD be external-host independent. Say, an internal host A sent an ICMP Query out to an external host B using Query Id X. And, say, the NAT assigned this an external mapping of Query Id X' on the NAT's public address. If host A reused the Query Id X to send ICMP Queries to the same or different external host, the NAT device SHOULD reuse the same Query Id mapping (i.e., map the private host's Query Id X to Query Id X' on NAT's public IP address) instead of assigning a different mapping. This is similar to the "endpoint independent mapping" requirement specified in the TCP and UDP requirement documents [BEH-UDP], [BEH-TCP].

Below is justification for making the endpoint-independent mapping for ICMP Query Id a SHOULD [RFC2119] requirement. ICMP Ping [RFC1470] and ICMP traceroute [MS-TRCRT] are two most commonly known legacy applications built on top of ICMP Query messages. Neither of these applications require the ICMP Query Id to be retained across different sessions with external hosts. But, that may not be the case with future applications. In the future, when an end host application reuses the same Query Identifier in sessions with different target hosts, the end host application might require that the endpoint identity (i.e., the tuple of IP address and Query Identifier) appears the same across all its target hosts. In an IP network without NAT requirements, such a requirement will be valid.

In a world with NAT devices, the above assumption will be valid when NAT devices enforce endpoint mapping that is external-host independent. Given the dichotomy between legacy applications not requiring endpoint-independent mapping and future applications that might require it, the requirement level is kept at SHOULD [RFC2119].

REQ-1: Unless explicitly overridden by local policy, a NAT device MUST permit ICMP Queries and their associated responses, when the Query is initiated from a private host to the external hosts.

- a) NAT mapping of ICMP Query Identifiers SHOULD be external-host independent.

3.2. ICMP Query Session Timeouts

NATs maintain a mapping timeout for the ICMP Queries that traverse them. The mapping timeout is the time a mapping will stay active without packets traversing the NAT. There is great variation in the values used by different NATs. The ICMP Query session timeout requirement is necessary for current ICMP Query applications to work. Query response times can vary. ICMP-Query-based applications are primarily request/response driven.

Ideally, the timeout should be set to Maximum Round Trip Time (Maximum RTT). For the purposes of constraining the maximum RTT, the Maximum Segment Lifetime (MSL), defined in [RFC793], could be considered a guideline to set packet lifetime. Per [RFC793], MSL is the maximum amount of time a TCP segment can exist in a network before being delivered to the intended recipient. This is the maximum duration an IP packet can be assumed to take to reach the intended destination node before declaring that the packet will no longer be delivered. For an application initiating an ICMP Query message and waiting for a response for the Query, the Maximum RTT could in practice be constrained to be the sum total of MSL for the Query message and MSL for the response message. In other words, Maximum RTT could be constrained to no more than 2x MSL. The recommended value for MSL in [RFC793] is 120 seconds, even though several implementations set this to 60 seconds or 30 seconds. When MSL is 120 seconds, the Maximum RTT (2x MSL) would be 240 seconds.

In practice, ICMP Ping [RFC1470] and ICMP traceroute [MS-TRCRT], the two most commonly known legacy applications built on top of ICMP Query messages, take less than 10 seconds to complete a round trip when the target node is operational on the network.

Setting the ICMP NAT session timeout to a very large duration (say, 240 seconds) could potentially tie up precious NAT resources such as Query mappings and NAT Sessions for the whole duration. On the other hand, setting the timeout very low can result in premature freeing of NAT resources and applications failing to complete gracefully. The ICMP Query session timeout needs to be a balance between the two extremes. A 60-second timeout is a balance between the two extremes. An ICMP Query session timer MUST NOT expire in less than 60 seconds. It is RECOMMENDED that the ICMP Query session timer be made configurable.

REQ-2: An ICMP Query session timer MUST NOT expire in less than 60 seconds.

- a) It is RECOMMENDED that the ICMP Query session timer be made configurable.

4. ICMP Error Forwarding

Many applications make use of ICMP Error messages from end hosts and intermediate devices to shorten application timeouts. Some applications will not operate correctly without the receipt of ICMP Error messages. The following sub-sections discuss the requirements a NAT device must conform to in order to ensure reliable forwarding.

4.1. ICMP Error Payload Validation

An ICMP Error message checksum covers the entire ICMP message, including the payload. When an ICMP Error packet is received, if the ICMP checksum fails to validate, the NAT SHOULD silently drop the ICMP Error packet. This is because NAT uses the embedded IP and transport headers for forwarding and translating the ICMP Error message (described in Section 4.2). When the ICMP checksum is invalid, the embedded IP and transport headers, which are covered by the ICMP checksum, are also suspect.

[RFC1812] and [RFC1122] require a router or an end host that receives an IP packet with an invalid IP header checksum to silently drop the IP packet. As such, end hosts and routers do not generate an ICMP Error message in response to IP packets with invalid IP header checksums. For this reason, if the IP checksum of the embedded packet within an ICMP Error message fails to validate, the NAT SHOULD silently drop the Error packet.

When the IP packet embedded within the ICMP Error message includes IP options, the NAT device must not assume that the transport header of the embedded packet is at a fixed offset (as would be the case when there are no IP options associated with the packet) from the start of

the embedded packet. Specifically, if the embedded packet includes IP options, the NAT device MUST traverse past the IP options to locate the start of transport header for the embedded packet.

It is possible to compute the transport checksum of the embedded packet within an ICMP Error message when the ICMP Error message contains the entire transport segment. However, ICMP Error messages do not contain the entire transport segment in many cases. This is because [ICMP] stipulates that an ICMP Error message should embed an IP header and only a minimum of 64 bits of the IP payload. Even though Section 4.3.2.3 of [RFC1812] recommends an ICMP Error originator include as much of the original packet as possible in the payload, the length of the resulting ICMP datagram cannot exceed 576 bytes. ICMP Error originators truncate IP packets that do not fit within the stipulations.

A NAT device SHOULD NOT validate the transport checksum of the embedded packet within an ICMP Error message, even when it is possible to do so. This is because a NAT dropping an ICMP Error message due to an invalid transport checksum will make it harder for end hosts to receive error reporting for certain types of corruption. End-to-end validation of ICMP Error messages is best left to end hosts. Many newer revision end host TCP/IP stacks implement the improvements in [TCP-SOFT] and do not accept ICMP Error messages with a mismatched IP or TCP checksum in the embedded packet, if the embedded datagram contains a full IP packet and the TCP checksum can be calculated.

In the case that the ICMP Error payload includes ICMP extensions [ICMP-EXT], the NAT device MUST exclude the optional zero-padding and the ICMP extensions when evaluating transport checksum for the embedded packet. Readers are urged to refer to [ICMP-EXT] for information on identifying the presence of ICMP extensions in an ICMP message.

REQ-3: When an ICMP Error packet is received, if the ICMP checksum fails to validate, the NAT SHOULD silently drop the ICMP Error packet. If the ICMP checksum is valid, do the following:

- a) If the IP checksum of the embedded packet fails to validate, the NAT SHOULD silently drop the Error packet; and
- b) If the embedded packet includes IP options, the NAT device MUST traverse past the IP options to locate the start of the transport header for the embedded packet; and

- c) The NAT device SHOULD NOT validate the transport checksum of the embedded packet within an ICMP Error message, even when it is possible to do so; and
- d) If the ICMP Error payload contains ICMP extensions [ICMP-EXT], the NAT device MUST exclude the optional zero-padding and the ICMP extensions when evaluating transport checksum for the embedded packet.

4.2. ICMP Error Packet Translation

Section 4.3 of [NAT-TRAD] describes the fields of an ICMP Error message that a NAT device translates. In this section, we describe the requirements a NAT device must conform to while performing the translations. Requirements identified in this section are necessary for the current applications to work correctly.

Consider the following scenario in Figure 1. Say, NAT-xy is a NAT device connecting hosts in private and external networks. Router-x and Host-x are in the external network. Router-y and Host-y are in the private network. The subnets in the external network are routable from the private as well as the external domains. By contrast, the subnets in the private network are only routable within the private domain. When Host-y initiated a session to Host-x, let us say that the NAT device mapped the endpoint on Host-y into Host-y' in the external network. The following subsections describe the processing of ICMP Error messages on the NAT device(NAT-xy) when the NAT device receives an ICMP Error message in response to a packet pertaining to this session.

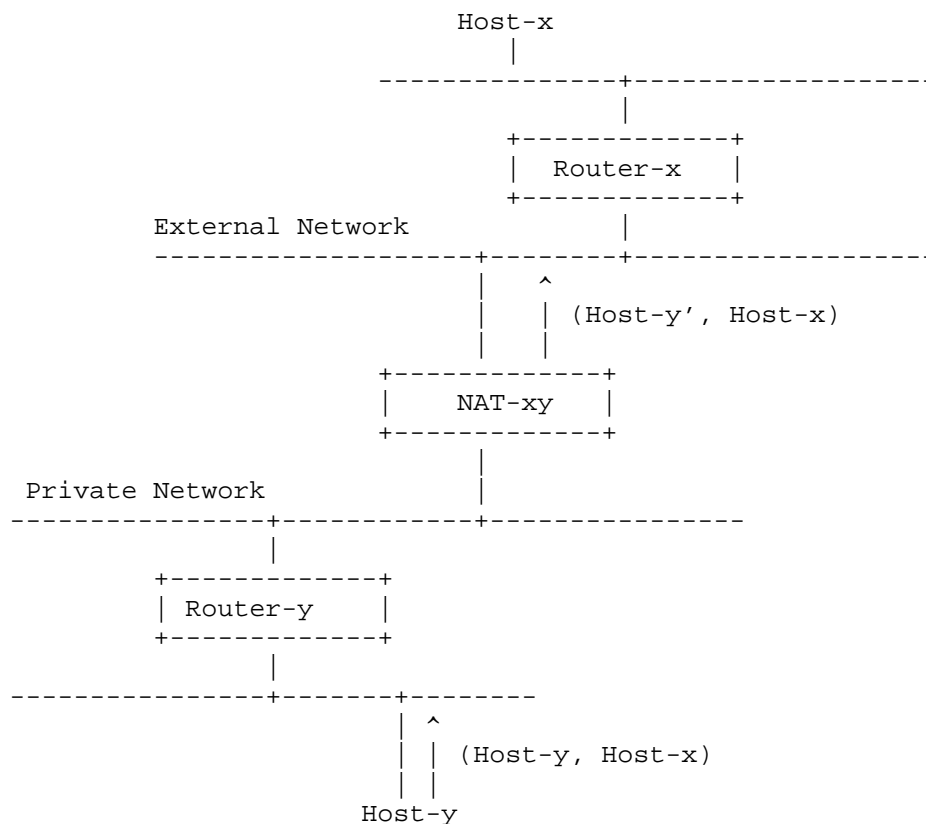


Figure 1. A Session from a Private Host Traversing a NAT Device

4.2.1. ICMP Error Packet Received from the External Realm

Say, a packet from Host-y to Host-x triggered an ICMP Error message from one of Router-x or Host-x (both of which are in the external domain). Such an ICMP Error packet will have one of Router-x or Host-x as the source IP address and Host-y' as the destination IP address as described in Figure 2 below.

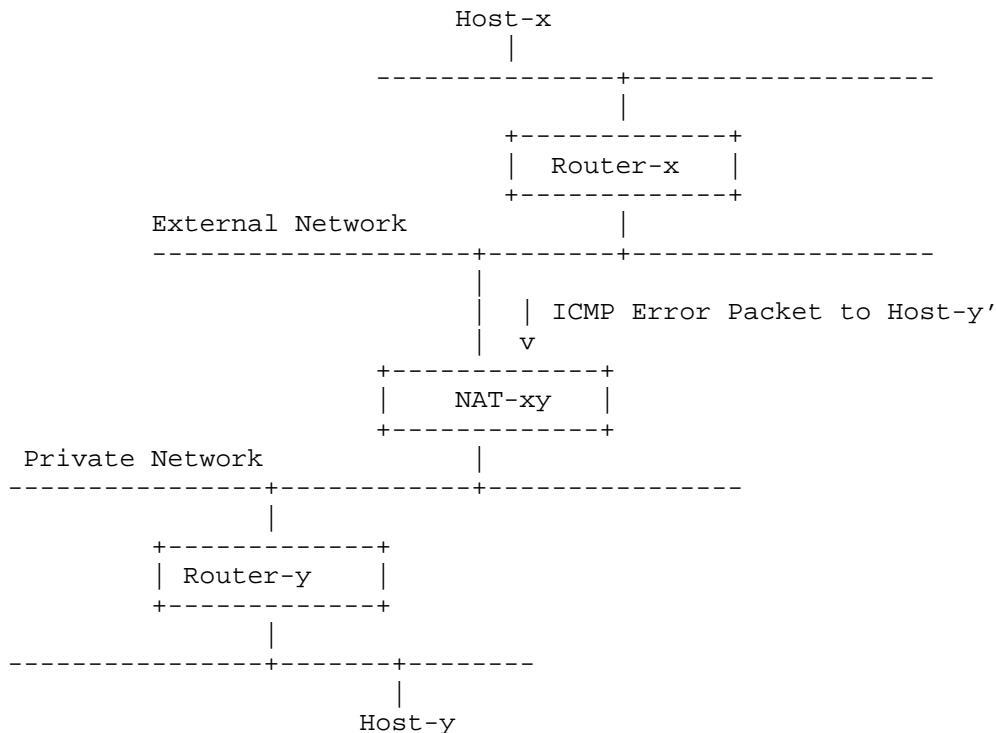


Figure 2. ICMP Error Packet Received from External Network

When the NAT device receives the ICMP Error packet, the NAT device uses the packet embedded within the ICMP Error message (i.e., the IP packet from Host-y' to Host-x) to look up the NAT Session to which the embedded packet belongs. If the NAT device does not have an active mapping for the embedded packet, the NAT SHOULD silently drop the ICMP Error packet. Otherwise, the NAT device MUST use the matching NAT Session to translate the embedded packet; that is, translate the source IP address of the embedded packet (e.g., Host-y' -> Host-y) and transport headers.

The ICMP Error payload may contain ICMP extension objects [ICMP-EXT]. NATs are encouraged to support ICMP extension objects. At the time of this writing, the authors are not aware of any standard ICMP extension objects containing realm-specific information.

The NAT device MUST also use the matching NAT Session to translate the destination IP address in the outer IP header. In the outer header, the source IP address will remain unchanged because the originator of the ICMP Error message (Host-x or Router-x) is in an external domain and is routable from the private domain.

REQ-4: If a NAT device receives an ICMP Error packet from an external realm, and the NAT device does not have an active mapping for the embedded payload, the NAT SHOULD silently drop the ICMP Error packet. If the NAT has active mapping for the embedded payload, then the NAT MUST do the following prior to forwarding the packet, unless explicitly overridden by local policy:

- a) Revert the IP and transport headers of the embedded IP packet to their original form, using the matching mapping; and
- b) Leave the ICMP Error type and code unchanged; and
- c) Modify the destination IP address of the outer IP header to be the same as the source IP address of the embedded packet after translation.

4.2.2. ICMP Error Packet Received from the Private Realm

Now, say, a packet from Host-x to Host-y triggered an ICMP Error message from one of Router-y or Host-y (both of which are in the private domain). Such an ICMP Error packet will have one of Router-y or Host-y as the source IP address and Host-x as the destination IP address as specified in Figure 3 below.

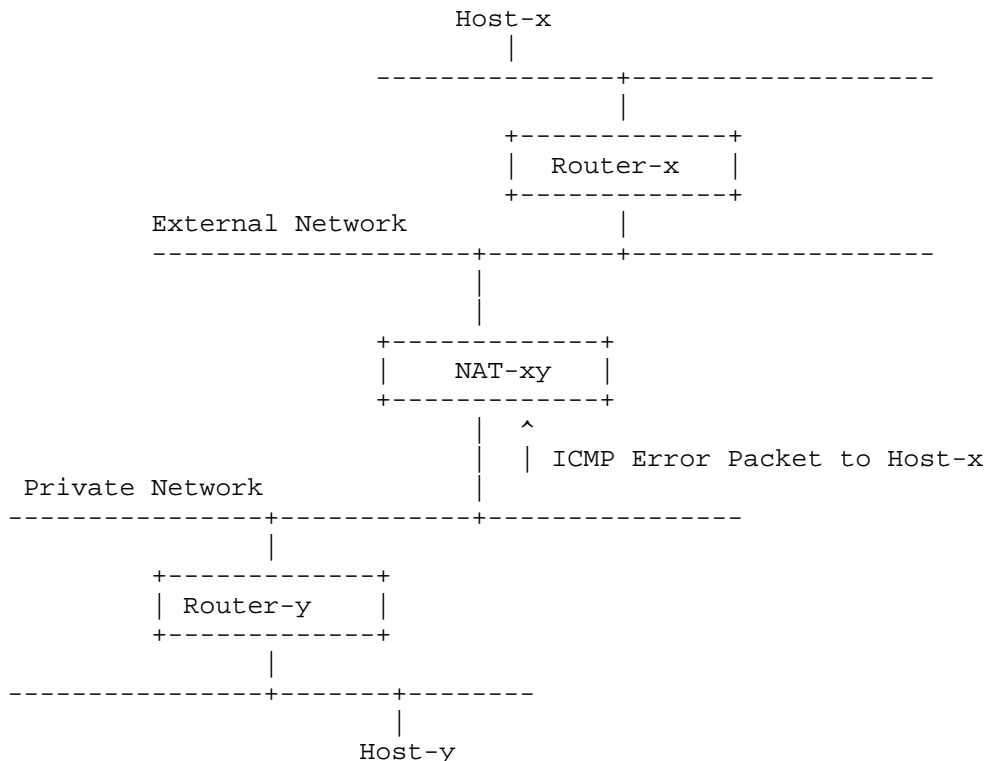


Figure 3. ICMP Error Packet Received from Private Network

When the NAT device receives the ICMP Error packet, the NAT device MUST use the packet embedded within the ICMP Error message (i.e., the IP packet from Host-x to Host-y) to look up the NAT Session to which the embedded packet belongs. If the NAT device does not have an active mapping for the embedded packet, the NAT SHOULD silently drop the ICMP Error packet. Otherwise, the NAT device MUST use the matching NAT Session to translate the embedded packet.

The ICMP Error payload may contain ICMP extension objects [ICMP-EXT]. NATs are encouraged to support ICMP extension objects. At the time of this writing, the authors are not aware of any standard ICMP extension objects containing realm-specific information.

In the outer header, the destination IP address will remain unchanged, as the IP address for Host-x is already in the external domain. If the ICMP Error message is generated by Host-y, the NAT device must simply use the NAT Session to translate the source IP address Host-y to Host-y'. If the ICMP Error message is originated by the intermediate node Router-y, translation of the source IP

address varies depending on whether the Basic NAT or NAPT function [NAT-TRAD] is enforced by the NAT device. A NAT device enforcing the Basic NAT function has a pool of public IP addresses and enforces address mapping (which is different from the endpoint mapping enforced by NAPT) when a private node initiates an outgoing session via the NAT device. So, if the NAT device has active mapping for the IP address of the intermediate node Router-y, the NAT device MUST translate the source IP address of the ICMP Error packet with the public IP address in the mapping. In all other cases, the NAT device MUST simply use its own IP address in the external domain to translate the source IP address.

REQ-5: If a NAT device receives an ICMP Error packet from the private realm, and the NAT does not have an active mapping for the embedded payload, the NAT SHOULD silently drop the ICMP Error packet. If the NAT has active mapping for the embedded payload, then the NAT MUST do the following prior to forwarding the packet, unless explicitly overridden by local policy:

- a) Revert the IP and transport headers of the embedded IP packet to their original form, using the matching mapping; and
- b) Leave the ICMP Error type and code unchanged; and
- c) If the NAT enforces Basic NAT function ([NAT-TRAD]), and the NAT has active mapping for the IP address that sent the ICMP Error, translate the source IP address of the ICMP Error packet with the public IP address in the mapping. In all other cases, translate the source IP address of the ICMP Error packet with its own public IP address.

4.3. NAT Sessions Pertaining to ICMP Error Payload

While processing an ICMP Error packet pertaining to an ICMP Query or Query response message, a NAT device MUST NOT refresh or delete the NAT Session that pertains to the embedded payload within the ICMP Error packet. This is in spite of the fact that the NAT device uses the NAT Session to translate the embedded payload. This ensures that the NAT Session will not be modified if someone is able to spoof ICMP Error messages for the session. [ICMP-ATK] lists a number of potential ICMP attacks that may be attempted by malicious users on the network. This requirement is necessary for current applications to work correctly.

REQ-6: While processing an ICMP Error packet pertaining to an ICMP Query or Query response message, a NAT device MUST NOT refresh or delete the NAT Session that pertains to the embedded payload within the ICMP Error packet.

5. Hairpinning Support for ICMP Packets

[BEH-UDP] and [BEH-TCP] mandate support for hairpinning for UDP and TCP sessions, respectively, on NAT devices. A NAT device needs to support hairpinning for ICMP Query sessions as well. Specifically, NAT devices enforcing Basic NAT [NAT-TRAD] MUST support the traversal of hairpinned ICMP Query sessions. Say, for example, individual private hosts register their NAT assigned external IP address with a rendezvous server. Other hosts that wish to initiate ICMP Query sessions to the registered hosts might do so using the public address registered with the rendezvous server. For this reason, Basic NAT devices are required to support the traversal of hairpinned ICMP Query sessions. This requirement is necessary for current applications to work correctly.

Packets belonging to any of the hairpinned sessions could, in turn, trigger ICMP Error messages directed to the source of hairpinned IP packets. Such hairpinned ICMP Error messages will traverse the NAT devices en route. All NAT devices (i.e., Basic NAT as well as NAPT devices) MUST support the traversal of hairpinned ICMP Error messages. Specifically, the NAT device must translate not only the embedded hairpinned packet, but also the outer IP header that is hairpinned. This requirement is necessary for current applications to work correctly.

A hairpinned ICMP Error message is received from a node in a private network. As such, the ICMP Error processing requirement specified in Req-5 is applicable in its entirety in processing the ICMP Error message. In addition, the NAT device MUST translate the destination IP address of the outer IP header to be same as the source IP address of the embedded IP packet after the translation.

REQ-7: NAT devices enforcing Basic NAT [NAT-TRAD] MUST support the traversal of hairpinned ICMP Query sessions. All NAT devices (i.e., Basic NAT as well as NAPT devices) MUST support the traversal of hairpinned ICMP Error messages:

- a) When forwarding a hairpinned ICMP Error message, the NAT device MUST translate the destination IP address of the outer IP header to be same as the source IP address of the embedded IP packet after the translation.

6. Rejection of Outbound Flows Disallowed by NAT

A NAT device typically permits all outbound sessions. However, a NAT device may disallow some outbound sessions due to resource constraints or administration considerations. For example, a NAT device may not permit the first packet of a new outbound session if the NAT device is out of resources (out of addresses or TCP/UDP ports, or NAT Session resources) to set up a state for the session, or, if the specific session is administratively restricted by the NAT device.

When a NAT device is unable to establish a NAT Session for a new transport-layer (TCP, UDP, ICMP, etc.) flow due to resource constraints or administrative restrictions, the NAT device SHOULD send an ICMP destination unreachable message, with a code of 13 (Communication administratively prohibited) to the sender, and drop the original packet. This requirement is meant primarily for future use. Current applications do not require this for them to work correctly. The justification for using ICMP code 13 in the ICMP Error message is as follows: Section 5.2.7.1 of [RFC1812] recommends routers use ICMP code 13 (Communication administratively prohibited) when they administratively filter packets. ICMP code 13 is a soft error and is on par with other soft error codes generated in response to transient events such as "network unreachable" (ICMP type=3, code=0).

Some NAT designers opt to never reject an outbound flow. When a NAT runs short of resources, they prefer to steal a resource from an existing NAT Session rather than reject the outbound flow. Such a design choice may appear conformant to REQ-8 below. However, the design choice is in violation of the spirit of both REQ-8 and REQ-2. Such a design choice is strongly discouraged.

REQ-8: When a NAT device is unable to establish a NAT Session for a new transport-layer (TCP, UDP, ICMP, etc.) flow due to resource constraints or administrative restrictions, the NAT device SHOULD send an ICMP destination unreachable message, with a code of 13 (Communication administratively prohibited) to the sender, and drop the original packet.

7. Conformance to RFC 1812

This document specifies NATs to have a behavior that is consistent with the way routers handle ICMP messages, as specified in Section 4.3 of [RFC1812]. However, since the publication of [RFC1812], some of its requirements are no longer best current practices. Thus, the following requirements are derived from [RFC1812] and apply to NATs compliant with this specification:

REQ-9: A NAT device MAY implement a policy control that prevents ICMP messages being generated toward certain interface(s). Implementation of such a policy control overrides the MUSTs and SHOULDs in REQ-10.

REQ-10: Unless overridden by REQ-9's policy, a NAT device needs to support ICMP messages as below, some conforming to Section 4.3 of [RFC1812] and some superseding the requirements of Section 4.3 of [RFC1812]:

a. MUST support:

1. Destination Unreachable Message, as described in Section 7.1 of this document.
2. Time Exceeded Message, as described in Section 7.2 of this document.
3. Echo Request/Reply Messages, as described in REQ-1.

b. MAY support:

1. Redirect Message, as described in Section 4.3.3.2 of [RFC1812].
2. Timestamp and Timestamp Reply Messages, as described in Section 4.3.3.8 of [RFC1812].
3. Source Route Options, as described in Section 7.3 of this document.
4. Address Mask Request/Reply Message, as described in Section 7.4 of this document.
5. Parameter Problem Message, as described in Section 7.5 of this document.
6. Router Advertisement and Solicitations, as described in Section 7.6 of this document.

c. SHOULD NOT support:

1. Source Quench Message, as described in Section 4.3.3.3 of [RFC1812].
2. Information Request/reply, as described in Section 4.3.3.7 of [RFC1812].

In addition, a NAT device is RECOMMENDED to conform to the following implementation considerations:

- d. DS Field Usage, as described in Section 7.7 of this document.
- e. When Not to Send ICMP Errors, as described in Section 4.3.2.7 of [RFC1812].
- f. Rate Limiting, as described in Section 4.3.2.8 of [RFC1812].

7.1. IP Packet Fragmentation

Many networking applications (which include TCP- as well as UDP-based applications) depend on ICMP Error messages from the network to perform end-to-end path MTU discovery [PMTU]. Once the path MTU is discovered, an application that chooses to avoid fragmentation may do so by originating IP packets that fit within the path MTU en route and setting the DF (Don't Fragment) bit in the IP header, so the intermediate nodes en route do not fragment the IP packets. The following sub-sections discuss the need for NAT devices to honor the DF bit in the IP header and be able to generate "Packet Too Big" ICMP Error message when they cannot forward the IP packet without fragmentation. Also discussed is the need to seamlessly forward ICMP Error messages generated by other intermediate devices.

7.1.1. Generating "Packet Too Big" ICMP Error Message

When a router is unable to forward a datagram because it exceeds the MTU of the next-hop network and its Don't Fragment (DF) bit is set, the router is required by [RFC1812] to return an ICMP Destination Unreachable message to the source of the datagram, with the code indicating "fragmentation needed and DF set". Further, [PMTU] states that the router MUST include the MTU of that next-hop network in the low-order 16 bits of the ICMP header field that is labeled "unused" in the ICMP specification [ICMP].

A NAT device MUST honor the DF bit in the IP header of the packets that transit the device. The NAT device may not be able to forward an IP packet without fragmentation if the MTU on the forwarding interface of the NAT device is not adequate for the IP packet. If the DF bit is set on a transit IP packet and the NAT device cannot forward the packet without fragmentation, the NAT device MUST send a "Packet Too Big" ICMP message (ICMP type 3, code 4) with the next-hop MTU back to the sender and drop the original IP packet. The sender will usually resend after taking the appropriate corrective action.

If the DF bit is not set and the MTU on the forwarding interface of the NAT device mandates fragmentation, the NAT device MUST fragment the packet and forward the fragments [RFC1812].

7.1.2. Forwarding "Packet Too Big" ICMP Error Message

This is the flip side of the argument for the above section. By virtue of the address translation NAT performs, NAT may end up being the recipient of "Packet Too Big" messages.

When the NAT device is the recipient of a "Packet Too Big" ICMP message from the network, the NAT device MUST forward the ICMP message back to the intended recipient, pursuant to the previously stated requirements (REQ-3, REQ-4, and REQ-5).

7.2. Time Exceeded Message

A NAT device MUST generate a "Time Exceeded" ICMP Error message when it discards a packet due to an expired Time to Live (TTL) field. A NAT device MAY have a per-interface option to disable origination of these messages on that interface, but that option MUST default to allowing the messages to be originated.

When a NAT device conforms to the above requirement, it ensures that legacy applications such as Traceroute [RFC1470], [MS-TRCRT], which depend upon the "Time Exceeded" ICMP Error message, will continue to operate even as NAT devices are en route.

7.3. Source Route Options

A NAT device MAY support modifying IP addresses in the source route option so the IP addresses in the source route option are realm relevant. If a NAT device does not support forwarding packets with the source route option, the NAT device SHOULD NOT forward outbound ICMP messages that contain the source route option in the outer or inner IP header. This is because such messages could reveal private IP addresses to the external realm.

7.4. Address Mask Request/Reply Messages

Section 4.3.3.9 of [RFC1812] says an IP router MUST implement support for receiving ICMP Address Mask Request messages and responding with ICMP Address Mask Reply messages. However, several years (more than 13 years at the time of this document) have elapsed since the text in RFC 1812 was written. In the intervening time, DHCP [DHCP] has replaced the use of address mask request/reply. At the current time,

there is rarely any host that does not meet host requirements [RFC1122] and needs a NAT device to support address mask request/reply.

For this reason, a NAT device is not required to support this ICMP message.

A NAT device MAY support address mask request/reply messages.

7.5. Parameter Problem Message

Section 4.3.3.5 of [RFC1812] says an IP router MUST generate a Parameter Problem message for any error not specifically covered by another ICMP message. However, this message is rarely used in practice in networks where IPv4 NATs are deployed.

For this reason, a NAT device is not required to support this ICMP message.

A NAT device MAY support parameter problem messages.

7.6. Router Advertisement and Solicitations

Section 4.3.3.10 of [RFC1812] says an IP router MUST support the router part of the ICMP Router Discovery Protocol on all connected networks on which the router supports either IP multicast or IP broadcast addressing. However, this message is rarely used in practice in networks where IPv4 NATs are deployed.

For this reason, a NAT device is not required to support this ICMP message.

A NAT device MAY support Router Advertisement and Solicitations.

7.7. DS Field Usage

[RFC1812] refers to the Type of Service (TOS) octet in the IP header, which contains the TOS and IP precedence fields. However, the TOS and IP precedence fields are no longer in use today. [RFC2474] renamed the TOS octet as the DS field and defined diffserv classes within the DS field.

When generating an ICMP message, a NAT device SHOULD copy the diffserv class of the message that causes the sending of the ICMP error message. A NAT device MAY allow configuration of the diffserv class to be used for the different types of ICMP messages.

8. Non-QueryError ICMP Messages

In the preceding sections, ICMP requirements were identified for NAT devices, with a primary focus on ICMP Query and ICMP Error messages, as defined in the Terminology Section (see Section 2). This document provides no guidance on the handling of Non-QueryError ICMP messages by the NAT devices. A NAT MAY drop or appropriately handle Non-QueryError ICMP messages.

REQ-11: A NAT MAY drop or appropriately handle Non-QueryError ICMP messages. The semantics of Non-QueryError ICMP messages is defined in Section 2.

9. Summary of Requirements

Below is a summary of all the requirements.

REQ-1: Unless explicitly overridden by local policy, a NAT device MUST permit ICMP Queries and their associated responses, when the Query is initiated from a private host to the external hosts.

a) NAT mapping of ICMP Query Identifiers SHOULD be external host independent.

REQ-2: An ICMP Query session timer MUST NOT expire in less than 60 seconds.

a) It is RECOMMENDED that the ICMP Query session timer be made configurable.

REQ-3: When an ICMP Error packet is received, if the ICMP checksum fails to validate, the NAT SHOULD silently drop the ICMP Error packet. If the ICMP checksum is valid, do the following:

a) If the IP checksum of the embedded packet fails to validate, the NAT SHOULD silently drop the Error packet; and

b) If the embedded packet includes IP options, the NAT device MUST traverse past the IP options to locate the start of the transport header for the embedded packet; and

c) The NAT device SHOULD NOT validate the transport checksum of the embedded packet within an ICMP Error message, even when it is possible to do so; and

- d) If the ICMP Error payload contains ICMP extensions [ICMP-EXT], the NAT device MUST exclude the optional zero-padding and the ICMP extensions when evaluating transport checksum for the embedded packet.

REQ-4: If a NAT device receives an ICMP Error packet from an external realm, and the NAT device does not have an active mapping for the embedded payload, the NAT SHOULD silently drop the ICMP Error packet. If the NAT has active mapping for the embedded payload, then the NAT MUST do the following prior to forwarding the packet, unless explicitly overridden by local policy:

- a) Revert the IP and transport headers of the embedded IP packet to their original form, using the matching mapping; and
- b) Leave the ICMP Error type and code unchanged; and
- c) Modify the destination IP address of the outer IP header to be same as the source IP address of the embedded packet after translation.

REQ-5: If a NAT device receives an ICMP Error packet from the private realm, and the NAT does not have an active mapping for the embedded payload, the NAT SHOULD silently drop the ICMP Error packet. If the NAT has active mapping for the embedded payload, then the NAT MUST do the following prior to forwarding the packet, unless explicitly overridden by local policy.

- a) Revert the IP and transport headers of the embedded IP packet to their original form, using the matching mapping; and
- b) Leave the ICMP Error type and code unchanged; and
- c) If the NAT enforces Basic NAT function [NAT-TRAD], and the NAT has active mapping for the IP address that sent the ICMP Error, translate the source IP address of the ICMP Error packet with the public IP address in the mapping. In all other cases, translate the source IP address of the ICMP Error packet with its own public IP address.

REQ-6: While processing an ICMP Error packet pertaining to an ICMP Query or Query response message, a NAT device MUST NOT refresh or delete the NAT Session that pertains to the embedded payload within the ICMP Error packet.

- REQ-7: NAT devices enforcing Basic NAT ([NAT-TRAD]) MUST support the traversal of hairpinned ICMP Query sessions. All NAT devices (i.e., Basic NAT as well as NAT devices) MUST support the traversal of hairpinned ICMP Error messages.
- a) When forwarding a hairpinned ICMP Error message, the NAT device MUST translate the destination IP address of the outer IP header to be same as the source IP address of the embedded IP packet after the translation.
- REQ-8: When a NAT device is unable to establish a NAT Session for a new transport-layer (TCP, UDP, ICMP, etc.) flow due to resource constraints or administrative restrictions, the NAT device SHOULD send an ICMP destination unreachable message, with a code of 13 (Communication administratively prohibited) to the sender, and drop the original packet.
- REQ-9: A NAT device MAY implement a policy control that prevents ICMP messages being generated toward certain interface(s). Implementation of such a policy control overrides the MUSTs and SHOULDs in REQ-10.
- REQ-10: Unless overridden by REQ-9's policy, a NAT device needs to support ICMP messages as below, some conforming to Section 4.3 of [RFC1812] and some superseding the requirements of Section 4.3 of [RFC1812]:
- a. MUST support:
1. Destination Unreachable Message, as described in Section 7.1 of this document.
 2. Time Exceeded Message, as described in Section 7.2 of this document.
 3. Echo Request/Reply Messages, as described in REQ-1.
- b. MAY support:
1. Redirect Message, as described in Section 4.3.3.2 of [RFC1812].
 2. Timestamp and Timestamp Reply Messages, as described in Section 4.3.3.8 of [RFC1812].
 3. Source Route Options, as described in Section 7.3 of this document.

4. Address Mask Request/Reply Message, as described in Section 7.4 of this document.
 5. Parameter Problem Message, as described in Section 7.5 of this document.
 6. Router Advertisement and Solicitations, as described in Section 7.6 of this document.
- c. SHOULD NOT support:
1. Source Quench Message, as described in Section 4.3.3.3 of [RFC1812].
 2. Information Request/reply, as described in Section 4.3.3.7 of [RFC1812].

In addition, a NAT device is RECOMMENDED to conform to the following implementation considerations:

- d. DS Field Usage, as described in Section 7.7 of this document.
- e. When Not to Send ICMP Errors, as described in Section 4.3.2.7 of [RFC1812].
- f. Rate Limiting, as described in Section 4.3.2.8 of [RFC1812].

REQ-11: A NAT MAY drop or appropriately handle Non-QueryError ICMP messages. The semantics of Non-QueryError ICMP messages is defined in Section 2.

10. Security Considerations

This document does not introduce any new security concerns related to ICMP message handling in the NAT devices. However, the requirements in the document do mitigate some security concerns known to exist with ICMP messages.

[ICMP-ATK] lists a number of ICMP attacks that can be directed against end host TCP stacks. For example, a rogue entity could bombard the NAT device with a large number of ICMP Errors. If the NAT device did not validate the legitimacy of the ICMP Error packets, the ICMP Errors would be forwarded directly to the end nodes. End hosts not capable of defending themselves against such bogus ICMP Error attacks could be adversely impacted by such attacks. Req-3 recommends validating the ICMP checksum and the IP checksum of the

embedded payload prior to forwarding. These checksum validations by themselves do not protect end hosts from attacks. However, checksum validation mitigates end hosts from malformed ICMP Error attacks. Req-4 and Req-5 further mandate that when a NAT device does not find a mapping selection for the embedded payload, the NAT should drop the ICMP Error packets, without forwarding.

A rogue source could also try to send bogus ICMP Error messages for the active NAT sessions, with intent to destroy the sessions. Req-6 averts such an attack by ensuring that an ICMP Error message does not affect the state of a session on the NAT device.

Req-8 recommends a NAT device sending an ICMP Error message when the NAT device is unable to create a NAT session due to lack of resources. Some administrators may choose not to have the NAT device send an ICMP Error message, as doing so could confirm to a malicious attacker that the attack has succeeded. For this reason, sending of the specific ICMP Error message stated in REQ-8 is left to the discretion of the NAT device administrator.

Unfortunately, ICMP messages are sometimes blocked at network boundaries due to local security policy. Thus, some of the requirements in this document allow local policy to override the recommendations of this document. Blocking such ICMP messages is known to break some protocol features (most notably path MTU Discovery) and some applications (e.g., ping, traceroute), and such blocking is NOT RECOMMENDED.

11. Acknowledgements

The authors wish to thank Fernando Gont, Dan Wing, Carlos Pignataro, Philip Matthews, and members of the BEHAVE working group for doing a thorough review of early versions of the document and providing valuable input and offering generous amounts of their time in shaping the ICMP requirements. Their valuable feedback made this document a better read. Dan Wing and Fernando Gont were a steady source of encouragement. Fernando Gont spent many hours preparing slides and presenting the document in an IETF meeting on behalf of the authors. The authors wish to thank Carlos Pignataro and Dan Tappan, authors of the [ICMP-EXT] document, for their feedback concerning ICMP extensions. The authors wish to thank Philip Matthews for agreeing to be a technical reviewer for the document. Lastly, the authors highly appreciate the rigorous feedback from the IESG members.

12. References

12.1. Normative References

- [BEH-UDP] Audet, F., Ed., and C. Jennings, "Network Address Translation (NAT) Behavioral Requirements for Unicast UDP", BCP 127, RFC 4787, January 2007.
- [ICMP] Postel, J., "Internet Control Message Protocol", STD 5, RFC 792, September 1981.
- [ICMP-EXT] Bonica, R., Gan, D., Tappan, D., and C. Pignataro, "Extended ICMP to Support Multi-Part Messages", RFC 4884, April 2007.
- [NAT-TRAD] Srisuresh, P. and K. Egevang, "Traditional IP Network Address Translator (Traditional NAT)", RFC 3022, January 2001.
- [RFC793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.
- [RFC1812] Baker, F., Ed., "Requirements for IP Version 4 Routers", RFC 1812, June 1995.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

12.2. Informative References

- [BEH-APP] Ford, B., Srisuresh, P., and D. Kegel, "Application Design Guidelines for Traversal through Network Address Translators", Work in Progress, March 2007.
- [BEH-TCP] Guha, S., Ed., Biswas, K., Ford, B., Sivakumar, S., and P. Srisuresh, "NAT Behavioral Requirements for TCP", BCP 142, RFC 5382, October 2008.
- [DHCP] Droms, R., "Dynamic Host Configuration Protocol", RFC 2131, March 1997.
- [ICE] Rosenberg, J., "Interactive Connectivity Establishment (ICE): A Protocol for Network Address Translator (NAT) Traversal for Offer/Answer Protocols", Work in Progress, October 2007.
- [ICMP-ATK] Gont, F., "ICMP Attacks against TCP", Work in Progress, October 2008.

- [MS-TRCRT] Microsoft Support, "How to use the Tracert command-line utility to troubleshoot TCP/IP problems in Windows", <http://support.microsoft.com/kb/162326>, October, 2006.
- [NAT-MIB] Rohit, R., Srisuresh, P., Raghunarayan, R., Pai, N., and C. Wang, "Definitions of Managed Objects for Network Address Translators (NAT)", RFC 4008, March 2005.
- [NAT-TERM] Srisuresh, P. and M. Holdrege, "IP Network Address Translator (NAT) Terminology and Considerations", RFC 2663, August 1999.
- [PMTU] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191, November 1990.
- [RFC1122] Braden, R., Ed., "Requirements for Internet Hosts - Communication Layers", STD 3, RFC 1122, October 1989.
- [RFC1256] Deering, S., Ed., "ICMP Router Discovery Messages", RFC 1256, September 1991.
- [RFC1470] Enger, R. and J. Reynolds, "FYI on a Network Management Tool Catalog: Tools for Monitoring and Debugging TCP/IP Internets and Interconnected Devices", FYI 2, RFC 1470, June 1993.
- [RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.
- [RFC4065] Kempf, J., "Instructions for Seamoby and Experimental Mobility Protocol IANA Allocations", RFC 4065, July 2005.
- [TCP-SOFT] Gont, F., "TCP's Reaction to Soft Errors", RFC 5461, February 2009.
- [UNSAF] Daigle, L., Ed., and IAB, "IAB Considerations for UNilateral Self-Address Fixing (UNSAF) Across Network Address Translation", RFC 3424, November 2002.

Authors' Addresses

Pyda Srisuresh
Kazeon Systems, Inc.
1161 San Antonio Rd.
Mountain View, CA 94043
U.S.A.

Phone: +1 408 836 4773
EMail: srisuresh@yahoo.com

Bryan Ford
Max Planck Institute for Software Systems
Campus Building E1 4
D-66123 Saarbruecken
Germany

Phone: +49-681-9325657
EMail: baford@mpi-sws.org

Senthil Sivakumar
Cisco Systems, Inc.
7100-8 Kit Creek Road
PO Box 14987
Research Triangle Park, NC 27709-4987
U.S.A.

Phone: +1 919 392 5158
EMail: ssenthil@cisco.com

Saikat Guha
Cornell University
331 Upson Hall
Ithaca, NY 14853
U.S.A.

Phone: +1 607 255 1008
EMail: saikat@cs.cornell.edu