Network Working Group Internet Engineering Task Force
Request for Comments: 1122 R. Braden, Editor
October 1989

Requirements for Internet Hosts -- Communication Layers

Status of This Memo

This RFC is an official specification for the Internet community. It incorporates by reference, amends, corrects, and supplements the primary protocol standards documents relating to hosts. Distribution of this document is unlimited.

Summary

This is one RFC of a pair that defines and discusses the requirements for Internet host software. This RFC covers the communications protocol layers: link layer, IP layer, and transport layer; its companion RFC-1123 covers the application and support protocols.

Table of Contents

1. INTRODUCTION 5
1.1 The Internet Architecture 6
1.1.1 Internet Hosts 6
1.1.2 Architectural Assumptions 7
1.1.3 Internet Protocol Suite 8
1.1.4 Embedded Gateway Code 10
1.2 General Considerations 12
1.2.1 Continuing Internet Evolution 12
1.2.2 Robustness Principle 12
1.2.3 Error Logging 13
1.2.4 Configuration 14
1.3 Reading this Document 15
1.3.1 Organization 15
1.3.2 Requirements 16
1.3.3 Terminology 17
1.4 Acknowledgments 20

2. LINK LAYER 21
2.1 INTRODUCTION 21

Dell et al., Ex. 1012
Dell et al., v. Corrigent Corporation
IPR2023-00464
Page 1 of 8

algorithm which is totally satisfactory, though it has identified several forbidden paths and promising techniques.

* A particular gateway SHOULD NOT be used indefinitely in the absence of positive indications that it is functioning.

* Active probes such as "pinging" (i.e., using an ICMP Echo Request/Reply exchange) are expensive and scale poorly. In particular, hosts MUST NOT actively check the status of a first-hop gateway by simply pinging the gateway continuously.

* Even when it is the only effective way to verify a gateway's status, pinging MUST be used only when traffic is being sent to the gateway and when there is no other positive indication to suggest that the gateway is functioning.

* To avoid pinging, the layers above and/or below the Internet layer SHOULD be able to give "advice" on the status of route cache entries when either positive (gateway OK) or negative (gateway dead) information is available.

DISCUSSION:

If an implementation does not include an adequate mechanism for detecting a dead gateway and re-routing, a gateway failure may cause datagrams to apparently vanish into a "black hole". This failure can be extremely confusing for users and difficult for network personnel to debug.

The dead-gateway detection mechanism must not cause unacceptable load on the host, on connected networks, or on first-hop gateway(s). The exact constraints on the timeliness of dead gateway detection and on acceptable load may vary somewhat depending on the nature of the host's mission, but a host generally needs to detect a failed first-hop gateway quickly enough that transport-layer connections will not break before an alternate gateway can be selected.

Passing advice from other layers of the protocol stack complicates the interfaces between the layers, but it is the preferred approach to dead gateway detection. Advice can come from almost any part of the IP/TCP

architecture, but it is expected to come primarily from the transport and link layers. Here are some possible sources for gateway advice:

- o TCP or any connection-oriented transport protocol should be able to give negative advice, e.g., triggered by excessive retransmissions.

- o TCP may give positive advice when (new) data is acknowledged. Even though the route may be asymmetric, an ACK for new data proves that the acknowleged data must have been transmitted successfully.

- o An ICMP Redirect message from a particular gateway should be used as positive advice about that gateway.

- o Link-layer information that reliably detects and reports host failures (e.g., ARPANET Destination Dead messages) should be used as negative advice.

- o Failure to ARP or to re-validate ARP mappings may be used as negative advice for the corresponding IP address.

- o Packets arriving from a particular link-layer address are evidence that the system at this address is alive. However, turning this information into advice about gateways requires mapping the link-layer address into an IP address, and then checking that IP address against the gateways pointed to by the route cache. This is probably prohibitively inefficient.

Note that positive advice that is given for every datagram received may cause unacceptable overhead in the implementation.

While advice might be passed using required arguments in all interfaces to the IP layer, some transport and application layer protocols cannot deduce the correct advice. These interfaces must therefore allow a neutral value for advice, since either always-positive or always-negative advice leads to incorrect behavior.

There is another technique for dead gateway detection that has been commonly used but is not recommended.

This technique depends upon the host passively receiving ("wiretapping") the Interior Gateway Protocol (IGP) datagrams that the gateways are broadcasting to each other. This approach has the drawback that a host needs to recognize all the interior gateway protocols that gateways may use (see [INTRO:2]). In addition, it only works on a broadcast network.

At present, pinging (i.e., using ICMP Echo messages) is the mechanism for gateway probing when absolutely required. A successful ping guarantees that the addressed interface and its associated machine are up, but it does not guarantee that the machine is a gateway as opposed to a host. The normal inference is that if a Redirect or other evidence indicates that a machine was a gateway, successful pings will indicate that the machine is still up and hence still a gateway. However, since a host silently discards packets that a gateway would forward or redirect, this assumption could sometimes fail. To avoid this problem, a new ICMP message under development will ask "are you a gateway?"

IMPLEMENTATION:

The following specific algorithm has been suggested:

- o Associate a "reroute timer" with each gateway pointed to by the route cache. Initialize the timer to a value Tr, which must be small enough to allow detection of a dead gateway before transport connections time out.

- o Positive advice would reset the reroute timer to Tr. Negative advice would reduce or zero the reroute timer.

- o Whenever the IP layer used a particular gateway to route a datagram, it would check the corresponding reroute timer. If the timer had expired (reached zero), the IP layer would send a ping to the gateway, followed immediately by the datagram.

- o The ping (ICMP Echo) would be sent again if necessary, up to N times. If no ping reply was received in N tries, the gateway would be assumed to have failed, and a new first-hop gateway would be chosen for all cache entries pointing to the failed gateway.

Note that the size of Tr is inversely related to the amount of advice available. Tr should be large enough to insure that:

* Any pinging will be at a low level (e.g., <10%) of all packets sent to a gateway from the host, AND

* pinging is infrequent (e.g., every 3 minutes)

Since the recommended algorithm is concerned with the gateways pointed to by route cache entries, rather than the cache entries themselves, a two level data structure (perhaps coordinated with ARP or similar caches) may be desirable for implementing a route cache.

3.3.1.5 New Gateway Selection

If the failed gateway is not the current default, the IP layer can immediately switch to a default gateway. If it is the current default that failed, the IP layer MUST select a different default gateway (assuming more than one default is known) for the failed route and for establishing new routes.

DISCUSSION:

When a gateway does fail, the other gateways on the connected network will learn of the failure through some inter-gateway routing protocol. However, this will not happen instantaneously, since gateway routing protocols typically have a settling time of 30-60 seconds. If the host switches to an alternative gateway before the gateways have agreed on the failure, the new target gateway will probably forward the datagram to the failed gateway and send a Redirect back to the host pointing to the failed gateway (!). The result is likely to be a rapid oscillation in the contents of the host's route cache during the gateway settling period. It has been proposed that the dead-gateway logic should include some hysteresis mechanism to prevent such oscillations. However, experience has not shown any harm from such oscillations, since service cannot be restored to the host until the gateways' routing information does settle down.

IMPLEMENTATION:

One implementation technique for choosing a new default gateway is to simply round-robin among the default gateways in the host's list. Another is to rank the

gateways in priority order, and when the current default gateway is not the highest priority one, to "ping" the higher-priority gateways slowly to detect when they return to service. This pinging can be at a very low rate, e.g., 0.005 per second.

3.3.1.6 Initialization

The following information MUST be configurable:

(1) IP address(es).

(2) Address mask(s).

(3) A list of default gateways, with a preference level.

A manual method of entering this configuration data MUST be provided. In addition, a variety of methods can be used to determine this information dynamically; see the section on "Host Initialization" in [INTRO:1].

DISCUSSION:
Some host implementations use "wiretapping" of gateway protocols on a broadcast network to learn what gateways exist. A standard method for default gateway discovery is under development.

3.3.2 Reassembly

The IP layer MUST implement reassembly of IP datagrams.

We designate the largest datagram size that can be reassembled by EMTU_R ("Effective MTU to receive"); this is sometimes called the "reassembly buffer size". EMTU_R MUST be greater than or equal to 576, SHOULD be either configurable or indefinite, and SHOULD be greater than or equal to the MTU of the connected network(s).

DISCUSSION:
A fixed EMTU_R limit should not be built into the code because some application layer protocols require EMTU_R values larger than 576.

IMPLEMENTATION:
An implementation may use a contiguous reassembly buffer for each datagram, or it may use a more complex data structure that places no definite limit on the reassembled datagram size; in the latter case, EMTU_R is said to be

"indefinite".

Logically, reassembly is performed by simply copying each fragment into the packet buffer at the proper offset. Note that fragments may overlap if successive retransmissions use different packetizing but the same reassembly Id.

The tricky part of reassembly is the bookkeeping to determine when all bytes of the datagram have been reassembled. We recommend Clark's algorithm [IP:10] that requires no additional data space for the bookkeeping. However, note that, contrary to [IP:10], the first fragment header needs to be saved for inclusion in a possible ICMP Time Exceeded (Reassembly Timeout) message.

There MUST be a mechanism by which the transport layer can learn MMS_R, the maximum message size that can be received and reassembled in an IP datagram (see GET_MAXSIZES calls in Section 3.4). If EMTU_R is not indefinite, then the value of MMS_R is given by:

```
MMS_R = EMTU_R - 20
```

since 20 is the minimum size of an IP header.

There MUST be a reassembly timeout. The reassembly timeout value SHOULD be a fixed value, not set from the remaining TTL. It is recommended that the value lie between 60 seconds and 120 seconds. If this timeout expires, the partially-reassembled datagram MUST be discarded and an ICMP Time Exceeded message sent to the source host (if fragment zero has been received).

DISCUSSION:

The IP specification says that the reassembly timeout should be the remaining TTL from the IP header, but this does not work well because gateways generally treat TTL as a simple hop count rather than an elapsed time. If the reassembly timeout is too small, datagrams will be discarded unnecessarily, and communication may fail. The timeout needs to be at least as large as the typical maximum delay across the Internet. A realistic minimum reassembly timeout would be 60 seconds.

It has been suggested that a cache might be kept of round-trip times measured by transport protocols for various destinations, and that these values might be used to dynamically determine a reasonable reassembly timeout

value. Further investigation of this approach is required.

If the reassembly timeout is set too high, buffer resources in the receiving host will be tied up too long, and the MSL (Maximum Segment Lifetime) [TCP:1] will be larger than necessary. The MSL controls the maximum rate at which fragmented datagrams can be sent using distinct values of the 16-bit Ident field; a larger MSL lowers the maximum rate. The TCP specification [TCP:1] arbitrarily assumes a value of 2 minutes for MSL. This sets an upper limit on a reasonable reassembly timeout value.

3.3.3 Fragmentation

Optionally, the IP layer MAY implement a mechanism to fragment outgoing datagrams intentionally.

We designate by EMTU_S ("Effective MTU for sending") the maximum IP datagram size that may be sent, for a particular combination of IP source and destination addresses and perhaps TOS.

A host MUST implement a mechanism to allow the transport layer to learn MMS_S, the maximum transport-layer message size that may be sent for a given {source, destination, TOS} triplet (see GET_MAXSIZES call in Section 3.4). If no local fragmentation is performed, the value of MMS_S will be:

```
MMS_S = EMTU_S - <IP header size>
```

and EMTU_S must be less than or equal to the MTU of the network interface corresponding to the source address of the datagram. Note that <IP header size> in this equation will be 20, unless the IP reserves space to insert IP options for its own purposes in addition to any options inserted by the transport layer.

A host that does not implement local fragmentation MUST ensure that the transport layer (for TCP) or the application layer (for UDP) obtains MMS_S from the IP layer and does not send a datagram exceeding MMS_S in size.

It is generally desirable to avoid local fragmentation and to choose EMTU_S low enough to avoid fragmentation in any gateway along the path. In the absence of actual knowledge of the minimum MTU along the path, the IP layer SHOULD use EMTU_S <= 576 whenever the destination address is not on a connected network, and otherwise use the connected network's