

中华人民共和国通信行业标准

YD/T 3308—2017

IP 组播 Ping 与路径探测协议

The ping and traceroute protocols for IP multicast

2017-11-07 发布

2018-01-01 实施

中华人民共和国工业和信息化部 发布

目 次

前言	II
1 范围	1
2 规范性引用文件	1
3 缩略语	2
4 组播 Ping	3
4.1 协议概述	3
4.2 协议操作规程	4
4.3 客户端行为	8
4.4 服务器行为	9
4.5 实现考虑	9
4.6 IANA 考虑	10
4.7 安全考虑	10
5 组播路由探测	10
5.1 概述	10
5.2 报文格式	11
5.3 路由器行为	17
5.4 客户端行为	19
5.5 组播协议相关处理	20
5.6 问题分析与故障定位	21
5.7 IANA 考虑	22
5.8 安全考虑	22
参考文献	23

前 言

本标准按照 GB/T 1.1-2009 给出的规则起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本标准由中国标准化协会提出并归口。

本标准起草单位：华为技术有限公司、中国信息通信研究院、中国电信集团公司、中国移动通信集团公司、中兴通讯股份有限公司、上海贝尔股份有限公司、成都迈普产业集团有限公司、杭州华三通信技术有限公司。

本标准主要起草人：刘晖、杜宗鹏、陈端。

IP 组播 Ping 与路径探测协议

1 范围

本标准规定了在 IP 组播网络中，如何支持组播 Ping 协议实现对组播转发路径可达性的检查，以及如何支持组播路由探测协议，实现对组播转发路径的追踪。组播 Ping 与组播路径探测的布署有利于提高组播网络的运维和管理能力。

本标准适用于支持 IP 组播业务的网络和设备。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

- | | |
|---------------|---|
| IETF RFC 768 | 用户数据报协议 (User Datagram Protocol) |
| IETF RFC 792 | 互联网控制报文协议 (Internet Control Message Protocol) |
| IETF RFC 1075 | 距离向量组播路由协议 (Distance Vector Multicast Routing Protocol) |
| IETF RFC 1305 | 网络时间协议版本 3: 规范, 实现和分析 (Network Time Protocol (Version 3) Specification, Implementation and Analysis) |
| IETF RFC 2113 | IP 路由器告警选项 (IP Router Alert Option) |
| IETF RFC 2711 | IPv6 路由器告警选项 (IPv6 Router Alert Option) |
| IETF RFC 2858 | 边界网络协议版本 4 的多协议扩展 (Multiprotocol Extensions for BGP-4) |
| IETF RFC 2863 | 接口组管理信息库规范 (The Interfaces Group MIB) |
| IETF RFC 3973 | 协议无关组播密集模式协议 (Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised)) |
| IETF RFC 4601 | 协议无关组播稀疏模式协议 (Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)) |
| IETF RFC 4605 | 基于互连网组管理协议/组播监听者发现协议的组播转发—IGMP/MLD 代理协议 (Internet Group Management Protocol (IGMP)/Multicast Listener Discovery (MLD)-Based Multicast Forwarding ("IGMP/MLD Proxying")) |
| IETF RFC 4607 | IP 源特定组播 (Source-Specific Multicast for IP) |
| IETF RFC 5132 | IP 组播管理信息库 (IP Multicast MIB) |
| IETF RFC 5405 | 应用设计者使用单播 UDP 的实现原则 (Unicast UDP Usage Guidelines for Application Designers) |
| IETF RFC 6450 | 组播 Ping 协议 (Multicast Ping Protocol) |

3 缩略语

以下缩略语适用于本文件。

AMT	自动组播隧道	Automatic Multicast Tunnel
ASM	任意源组播	Any Source Multicast
DF	指定转发者	Designated Forwarder
DOS	拒绝服务	Denial of Service
DVMRP	距离向量组播路由协议	Distance Vector Multicast Routing Protocol
G	组播组	Group
IANA	互联网网络号分配机构	Internet Assigned Numbers Authority
ICMP	互联网控制报文协议	Internet Control Message Protocol
ID	标识	Identification
IGMP	互连网组管理协议	Internet Group Management Protocol
IPv4	网际协议 (第四版)	Internet Protocol Version 4
IPv6	网际协议 (第六版)	Internet Protocol Version 6
MBGP	多协议边界网关协议	Multiprotocol Border Gateway Protocol
MBZ	必为 0 值	Must Be Zero
MIB	管理信息库	Management Information Base
MLD	组播监听发现协议	Multicast Listener Discovery
MTU	最大传输单元	Maximum Transmission Unit
NTP	网络时间协议	Network Time Protocol
PIM-SM	协议无关组播稀疏模式	Protocol Independent Multicast for Sparse Mode
RP	汇集点	Rendezvous Point
RPA	汇集点地址	Rendezvous Point Address
RPL	汇集点链路	Rendezvous Point Link
RTT	往返时间	Round Trip Time
S	组播源	Source
SSM	源特定组播	Source Specific Multicast
TLV	类型长度值	Time-Length-Value
TTL	生存期	Time To Live
UDP	用户数据报协议	User Datagram Protocol
UTF-8	单一码转换格式-8	Unicode Transformation Form at 8

4 组播 Ping

4.1 协议概述

组播 Ping 可用于检查组播可达性（见 IETF RFC6450）。除了检查源特定组播组（SSM）（见 IETF RFC4607）或任意源组播组（ASM）（见 IETF RFC4601）的接收，还可搜集其他相关的组播信息，如组播树构建时间，报文传递的跳数，以及报文延迟和丢失等。组播 Ping 与单播 Ping 所采用的 ICMP（见 IETF RFC792）响应请求/应答（Echo Request/Reply）机制是类似的，不同之处是需要使用 UDP（见 IETF RFC768）协议承载并要求客户端和服务端实现该协议。过渡路由器不要求支持该协议，只需采用通常的方式转发协议消息。

本协议是基于当前已经实现的 `ssmping` 和 `asmping` 工具描述的，这两个工具已经被互连网社团广泛使用以进行组播连通性测试。典型的协议使用方式为：服务器连续运行以响应客户端的请求，客户端通过某种方式了解到服务器的单播地址并测试从服务器的组播接收。客户端应用发送单播消息到服务器请求使用的组，可选地，用户可以选择一个特定的组或一个组的范围。服务器响应使用的组，或者如果无可用的组则响应一个错误信息。

对于 ASM 客户端加入一个 ASM 组 G (Group)，而对于 SSM 加入一个 SSM 组通道 (S,G) ((Source, Group))，其中 G 为服务器限定的组播组地址，而 S 为服务器的单播地址。在加入组 G 或通道 (S,G) 后，客户端向服务器单播发送组播 Ping 请求。请求发送时将 UDP 目的端口号设为组播 Ping 的专用端口号，请求周期性地发送（如每秒发送一次）。这些请求包含一个序列号，或典型地可包含一个时标，请求由服务器响应，服务器在响应时可增加一些新的选项。

对于每个请求，服务器发送两个应答：一个应答以请求客户端的源 IP 地址和源 UDP 端口号为目的地址和目的端口号单播发送，另一个应答以请求组地址和请求客户端的源端口号为目的地址和目的端口号组播发送。两个应答所采用的目的端口号与接收到的请求的源端口号相同。服务器单播或组播发送的中应答通过一个 TTL 选项限定 TTL 值（IPv4 TTL 或 IPv6 跳数，推荐设为 64），使得客户端可以计算跳数。客户在完成测量时应离开先前加入的组 G 或通道 (S,G) 。

通过使用本协议，客户端（或客户端的一个使用者）可以获得组播交付的若干特性。首先，通过接收单播应答，客户端可确证服务器接收到了单播请求，它正在运行着并能够响应客户端的请求。因而，假定客户端接收到单播响应，而如果此时不能够接收到组播响应，表明存在着组播转发问题或组播管理限制。如果能够接收组播响应，客户端不仅可以知道它可以接收组播流量，还可以估计建立组播树需要花费的时间，确定是否存在着报文丢失，测量环回时间（RTT）的大小和变化等。

对于单播而言，RTT 为从发送单播请求到接收单播应答所消耗的时间，组播 RTT 为发送单播请求到接收组播应答所花费的时间。通过限定应答的 TTL 值，客户端可以确定距离源的路由器跳数。主机可通过比较单播和组播的结果，检查单播与组播的跳数和 RTT 的差异。

组播跳数和跳数随时间的变化反映了组播树和组播树变化的细节。假定服务器同时发送单播和组播应答，客户端可以测量从服务器到客户端的单播和组播路径上的单向延时，以及单播和组播延时的差值。

服务器也为单播和组播分别限定一个时标，这是因为单播和组播不能够同时发送，发送的延时取决于主机的操作系统和当前的处理负荷。

4.2 协议操作规程

4.2.1 概述

组播 Ping 共使用四种报文类型：响应请求（Echo Request），响应应答（Echo Reply），初始消息（Init），和服务器响应（Server Resonse）消息。响应请求和响应应答消息用于实际的报文测量。初始消息用于初始化一个组播 Ping 会话和协商所要使用的组。服务器发送服务器响应消息用来响应初始消息。初始消息应以网络字节顺序表示，并且应使用 UDP 校验和。

所有的消息都采用相同的报文格式：一个字节表示消息类型，后面跟着几个 TLV（类型，长度，值）选项格式，以使协议更易于扩展。消息类型 0~191 的范围预留给 IANA 分配，范围 192~255 未在 IANA 注册，可预留为实验使用。

初始消息通常包含一些前缀选项，以便请求服务器从这些前缀中选择一个 Ping 使用的组播组。服务器发送一个服务器响应消息包含选定的组地址，或包含描述服务器可提供的组播组前缀选项。客户端在响应请求中通常会包含一组选项，服务器可做一个简单的应答（通过只改变报文类型）而不检查任何它不支持的选项，这对应于简单的组播 Ping 场景。但通常情况下，服务器应该增加一个 TTL 选项和其他可支持的选项，如响应客户端对特定请求的应答。响应应答（一个单播和一个组播）应首先包含响应请求中携带的相同选项（除会话 ID 选项之外），选项顺序与响应请求相同，然后在后面添加服务器希望添加的选项。服务器必不能处理未知的选项，但这些选项应包含在响应应答中。客户端应忽略未知的选项。

协议消息的大小通常小于路径 MTU，因而没必要考虑分片。然而对于路径 MTU 特别小的情况，或者客户端发送特别大的请求以验证它能够接收分片组播数据报的情况下，是可能产生分片的。文档中不限定路径 MTU 发现一定得到执行，这时可定义新的扩展选项使得客户端请求发现路径 MTU，并从服务器接收当前的路径 MTU。文档中定义了几种不同的选项，一些选项不要求服务器处理，服务器可以不加处理地返回。另一些是服务器关注的客户端选项和客户端请求的服务器选项，需要服务器进行特别的处理。除非特别限定，选项必不可以可在相同的消息中使用多次。

4.2.2 选项格式与定义

所有选项的格式如图 1 所示。

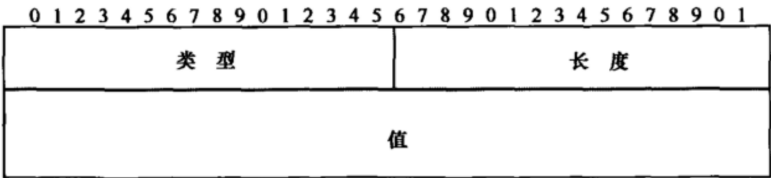


图1 选项格式

图 1 中。

- 类型（2 字节）：选项的类型；

- 长度（2 字节）：值的长度，取决于选项类型，取值范围可为 0~65535；
- 值（可变）：相关选项内容。如果长度为 0，则不应包括值的字段。

文档定义了下列选项：版本（0），客户端 ID（1），序列号（2），客户端时标（3），组播组（4），选项请求选项（5），服务器信息（6），TTL（9），组播前缀（10），会话 ID（11）和服务器时标（12），值 7 和值 8 是预留的，范围 0~49151 的选项类型预留给 IANA 机构分配。49152~65535 范围内的数值尚未注册可用于实验。各选项的详细定义如下。

a) 版本，类型 0：长度应为 1 字节。所有的消息中都应包含该选项。对于本协议该版本值设为 2。请注意该协议的早期实现只是部分地遵循该标准。它们可被认定为版本 1 并且不使用该选项。如果服务器接收到版本大于 2 或无版本的消息，服务器（除非支持特殊版本）发送服务器响应消息的时候，将版本置为 2。这时如果该消息中存在客户端 ID 和序列号选项则需原样返回，服务器不应包含其他任何选项。接收到版本号大于 2 的客户端应中止发送请求到服务器（除非它支持特殊的版本）。

b) 客户端 ID，类型 1：长度应为非 0。客户端应在所有消息的选项中包含该选项（无论初始消息还是响应请求）。客户端可使用任何值用以检测应答是否是针对初始/响应请求消息。服务器应把该选项看成透明数据，并且如果在请求中存在则需在应答中原样返回该选项。该选项值可以是一个处理 ID，也可以是一个与 IP 地址组合的处理 ID，以区分来自其他客户端的消息。客户端的实现者决定如何使用该选项。

c) 序列号，类型 2：长度为 4 字节。客户端应在响应请求消息中包含这个选项但不能够在初始消息中包含它。对响应请求消息进行应答的服务器应将它拷贝到响应应答中，或者在发生错误时将其拷贝到服务器响应消息中。该选项数值可以典型地从 1 开始，对每个顺序的响应请求逐 1 递增。

d) 客户端时标，类型 3：长度应为 8 字节。客户端在响应请求应包含此选项，但必不能在初始消息中包含该选项。对响应请求消息进行应答的服务器应将其拷贝到响应应答中。该选项可设为响应请求消息发送的时间，前 4 个字节限定了从初始历元（Epoch，1970 年 1 月 1 日，UTC0000）计算的以秒数表示的时间。下 4 个字节限定从前 4 个字节的开始的毫秒数。

e) 组播组，类型 4：长度应大于 2 字节；可以用于服务器响应消息中告诉客户端在后面的响应请求消息中使用哪个组。它应包含在响应请求中告诉服务器应响应哪个组地址（通常该组地址可以从前面服务器响应消息中获得）。它应在响应应答消息中使用（从响应请求消息中拷贝）。但不能在初始消息中使用。选项值的格式如图 2 所示。

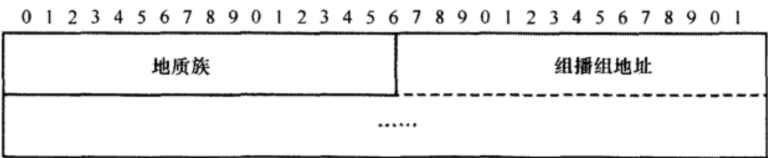


图2 组播组选项

地址族由 IANA 设定，值的范围为 0~65535，后面跟着组地址，选项值的长度对于 IPv4 为 6 字节，对于 IPv6 为 18 字节。

f) 选项请求选项，类型 5。

- 长度应大于 1 字节。该选项可以在客户端发送的消息（初始和响应请求消息）中使用。服务器必不能发送这个选项，除非该选项出现在响应请求消息中，服务器应在响应应答消息中将该选项原样返回。该选项包含客户端向服务器请求的一个选项类型的列表，对该选项的支持对客户端和服务器都是可选的。选项的长度为一个非 0 的偶数字节数，因为它包含一个或多个两字节的选项类型。选项值的格式如图 3 所示。

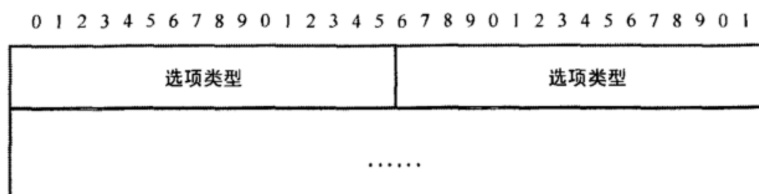


图3 选项请求选项

- 该选项可以被客户端用来请求服务器包含时标或服务器信息等选项。客户端可以在初始消息中请求服务器信息选项；它必不能在其他的消息中请求该选项。客户端可以在响应请求消息中请求时标选项，它必不能在其他消息中请求该选项。受上述限制的制约，支持选项请求选项的服务器，应该在对包含特定的选项请求选项的响应请求消息进行响应应答时包含该特定的选项。
 - 服务器可以根据实现或本地配置，不必包含所有客户端请求的选项，可以只包含一个选项。任何被请求的选项都被附加在其他原样返回的选项后面。
- g) 服务器信息，类型 6：长度应非 0，可用在服务器响应消息中，但必不能在其他消息中使用。对该选项的支持是可选的。支持该选项的服务器只有当客户端请求时才在服务器响应消息中添加该选项。选项只为 UTF-8 字符串，可以包含服务器的设备商或版本信息，也可以包含服务器支持的选项信息。交互式客户端可以支持该选项，也应当允许用户请求该字符串并显示它。
- h) 预留，类型 7：该选项曾为早期版本使用。客户端必不能使用该选项，服务器应把它看成是未知的选项（接收到不处理）。但如果在响应请求中接收到，则服务器应在响应应答消息中原样返回。
- i) 预留，类型 8：该选项曾为早期版本使用。客户端必不能使用该选项，服务器应把它看成是未知的选项（接收到不处理）。但如果在响应请求中接收到，则服务器应在响应应答消息中原样返回。
- j) TTL，类型 9。长度必须为 1 字节。该选项包含单一字节，限定一个响应应答消息的 TTL 值。每次服务器发送单播或组播响应应答消息，都应该包含此限定 TTL 的选项。该选项可被客户端使用用来确定消息传递的跳数。它必不能在其他消息中使用。如果服务器知道响应应答的 TTL 值，则服务器应该选定这个选项。通常情况下服务器可以将该 TTL 值设定到主机栈中。请注意 TTL 不一定对于单播和组播是相同的。也需注意即使客户端不请求这个选项，它也应被包含在响应应答中。
- k) 组播前缀，类型 10。
- 长度应大于 2 字节，如图 4 所示。可以在初始消息中请求前缀表示的组，也可在服务器响应消息中使用，以告诉客户端它可以使用哪些组。它必不能在响应请求和响应应答消息中使用。需注意该选项也可包含多次用于限定多个前缀。

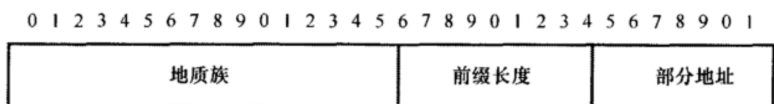


图4 组播前缀

— 地址族取值范围为 0~65535, 由 IANA 设定。前缀长度取值范围对于 IPv4 为 4 比特~32 比特, 对于 IPv6 为 8 比特~128 比特, 当取值为 0 时表示通配符。报文的最后是组地址。对任何组地址族, 前缀长度 0 意味着该地址族的任意组播地址都是可以接收的, 因而叫做通配符。组地址仅需包含覆盖前缀长度比特的足够的字节 (例如, 如果前缀长度为 12 比特~24 比特, 组地址应为 3 字节; 而在前缀长度为 0, 即通配符情况下, 则无需携带组地址)。任何超出前缀长度的比特必须被忽略。对于 IPv4, 选项值长度为 4 字节~7 字节, 而对于 IPv6, 选项值长度为 4 字节~19 字节, 而对于通配符, 选项值长度为 3 字节。

L) 会话 ID, 类型 11: 长度应为非 0 字节数。服务器应在服务器响应消息中包含该选项。如果客户端接收到该选项, 客户端应在接下来的响应请求消息中包含相同的会话 ID 选项。会话 ID 应被设为伪随机数, 使得其值难于预测。会话 ID 可用来预防在响应请求消息中进行源地址欺骗。

m) 服务器时标, 类型 12: 长度必 8 字节。如果客户端请求, 则支持该选项的服务器需要在响应应答消息中包含它。时标限定为发送响应应答消息的时间。前 4 个字节限定了从初始历元 (Epoch, 1970 年 1 月 1 日, UTC0000) 计算的以秒数表示的时间, 其后 4 个字节限定了秒数后面的毫秒数。

4.2.3 消息格式及定义

所有的消息都是由一个字节的类型加上可变长度的选项构成, 如图 5 所示。

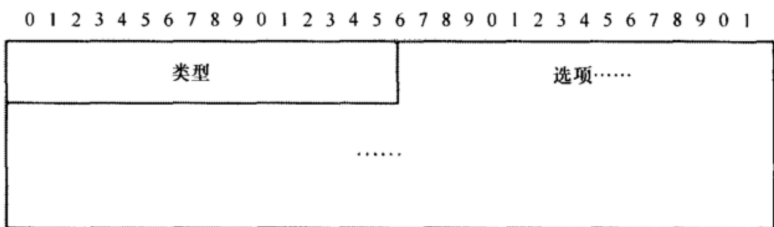


图5 消息格式

共定义了四种类型的消息。类型 81 (ASCII 码的字符 Q) 表示响应请求; 类型 65 (ASCII 码的字符 A) 表示响应应答; 类型 73 (ASCII 码的字符 I) 表示初始消息; 类型 83 (ASCII 码的字符 S) 表示服务器响应消息。选项立即跟随着类型字节并不以任何方式对齐, 即选项可以在任何字符边界开始。选项格式如 4.4.2 节所示。

每种消息中包含的选项种类如下所示。

a) 初始消息, 类型 73: 该信息由客户端发向服务器以请求特定的信息。它主要用于请求一个组播组地址, 也可用于检查服务器提供哪些组前缀。它应包含一个版本选项, 也应该包含一个客户端 ID 选项。它可包含选项请求选项和组播前缀选项。只有包含组前缀选项时才可用来请求一个组地址。如果包含多个前缀选项, 则前缀应以优先级顺序排列, 服务器将根据限定的优先级顺序考虑前缀, 如果找到了对应于一个前缀的组, 它将只返回这个组, 而不考虑其余的前缀。

b) 服务器响应消息, 类型 83: 该消息由服务器发送, 或者是初始消息的一个响应, 或者是响应请求的一个响应。当作为初始消息的响应时, 它可以向客户端提供一个组播组 (如果为客户端所请求), 或者提供其他的服务器信息。当作为响应请求的响应, 可以告诉客户端停止发送响应请求。该消息中总是包含版本选项。客户端 ID 选项和序列号选项如果在客户端消息中出现, 将在该响应中被原样返回。当向客户端提供一个组时, 消息中需包含组播组选项, 如果被请求也可包含服务器信息和前缀选项。

c) 响应请求, 类型 81: 该消息由客户端发送, 请求服务器发送单播和组播响应应答。它应包含版本、序列号和组播组选项。如果客户端在前面的服务器响应消息中接收到会话 ID, 则应包含会话 ID。此外它应该包含客户端 ID 和客户端时标选项, 可以包含选项请求选项。

d) 响应应答, 类型 65: 该消息由服务器发送, 作为响应请求的响应。此消息总是成对出现, 一个单播发送, 一个组播发送, 两个消息的内容大致相同。服务器总是原样返回响应请求中的所有选项 (会话 ID 除外)。响应请求中任何服务器不支持的选项, 也将被原样返回。两个响应应答消息总是包含 TTL 选项 (单播和组播设定的值可能不同)。两个响应应答消息当被请求时也应包含服务器时标选项 (单播和组播可设定不同的数值)。

4.2.4 速率限制

客户端应默认每秒发送至多一个响应请求, 服务器应默认执行限速, 以防止该协议被利用来实行 DoS 攻击。服务器应默认对于一个给定的客户端, 平均每秒至多对一个响应请求消息进行应答。服务器实现也应提供配置选项以允许特定的客户端以更快速的发送响应请求。如果特定的客户端 IP 地址允许更高的速率, 则应使用初始消息和会话 ID 选项以防止欺骗的发生。

该协议和应用的实现者应考虑 UDP 的使用规则 (见 IETF RFC5405), 特别是当客户端发送和服务器接收的速率超过每秒 1 个的情况下。

4.3 客户端行为

客户端仅要求用户限定服务器的单播地址, 然后即可发送携带前缀选项 (包含组地址选项和零前缀长度) 的初始消息。然后, 服务器决定应在该地址族上返回哪个组。客户端可允许用户限定组地址或前缀 (对于 IPv6, 用户可以仅要求限定一个范围或一个 RP 地址, 客户端可构建所需的前缀, 如嵌入式 RP)。客户端在初始消息中可限定一个或多个前缀选项以告诉服务器它希望使用的地址。如果用户限定了一个组地址, 组地址可以采用最大前缀长度编码 (例如 IPv4 为 32 比特)。前缀选项以优先级顺序排列, 将优先级最高的前缀放在最前面。

如果客户端接收到包含组地址的服务器响应消息时, 它可发送响应请求消息。如果没有组地址选项, 客户端将典型的错误退出。服务器可能在服务器响应消息中包含某些前缀选项, 客户端使用前缀选项向用户提供可用的前缀或可用的前缀范围。

假定客户端在服务器响应消息中获得一个组地址, 在让用户知道该组播组将被使用后, 可启动组播 Ping。通常, 客户端应每秒钟至多发送一个响应请求。

当发送响应请求时, 客户端应总是包含组选项。如果服务器响应消息中包含会话 ID, 则响应请求中应包含取值完全相同的同一选项。如果客户端发送响应请求后接收到服务器响应消息 (即, 服务器响

应消息中包含一个序列号)，这意味发生了错误，客户端立即停止发送响应请求消息。这种情形通常在服务器重启时出现。

客户端可以允许用户请求服务器信息。如果用户请求服务器信息，客户端可发送不携带前缀选项的初始消息，但携带选项请求选项，请求服务器返回一个服务器信息选项。服务器将返回支持的服务器信息，也可返回支持的前缀列表。但服务器不会返回一个组地址。客户端也可以通过发送无前缀并且不请求任何选项的初始消息获得一个前缀列表。

一种不被推荐的实现方式是，客户端可以挑选一个组播组，不进行初始消息—服务器响应消息协商而是直接发送响应请求。如果服务器支持并且可以接收这个组播组，服务器则发送一个通常的响应应答消息。否则，服务器将发送一个服务器响应消息中止客户端的请求。

4.4 服务器行为

如果初始消息包含前缀选项，服务器将顺序检查这些选项，以查看它是否可以从给定的前缀设定一个组播地址。服务器可以配置一组可提供的组地址。服务器可以从一个地址池中随机地挑选一个组播组，也可以基于客户端的 IP 地址或标识的哈希选定一个组，或简单地使用一个固定组。服务器可决定是否根据客户端 IP 地址包含站点范围的组地址。服务器决定是否允许多个客户端同时共享相同的组地址。

如果服务器找到了一个合适的组地址，它将在服务器响应消息中返回这个组地址选项。服务器应额外包含一个会话 ID。这将帮助服务器保存一些状态，例如确保客户端使用的是为它设定的组。一个好的会话 ID 将是一个难于预测的伪随机字节串。如果服务器不能发现合适的组地址，或者在初始消息中不存在前缀，可发送一个列出所有可用前缀选项的服务器响应消息。最后，如果初始消息请求服务器信息选项，服务器也应包含这个选项。

当服务器接收到响应请求消息，它可首选检查组地址和会话 ID 是否有效。如果服务器能够满足条件，它将发送一个单播的响应应答消息给客户端，同时也组播一个响应应答消息给到组地址上。响应应答消息以与响应请求完全相同的顺序包含选项（但不包含会话 ID），然后服务器添加一个 TTL 选项，若需要也可添加其他必要的选项，例如，它可能添加一个客户端请求的时标选项。如果服务器不接受响应请求（如错误的组地址或会话 ID，或请求报文过大），它可发送一个服务器响应消息，请求客户端中止请求。该服务器响应必须返回与响应请求相同的序列号。服务器响应消息可包含客户端希望请求的组地址的组前缀。单播和组播响应应答消息有相同的 UDP 载荷（TTL 和时标选项除外）。

需注意服务器可能没有接收到初始消息即接收到了响应请求消息，这通常发生在服务器重启或客户端直接发送响应请求而不发送初始消息的情况。服务器如果判定组地址可用，则可以选择响应这个请求，如果组地址不可用，服务器发送一个服务器响应消息。

4.5 实现考虑

服务器管理者应能够配置一个或多个组前缀。当在互联网和其他环境中部署服务器时，服务器管理者应能够限制服务器仅响应几个当前不为组播应用使用的组播组。服务器实现应灵活地提供给管理者使用不同的策略提供一个或多个组前缀以限定客户选择，例如站内客户端使用的站内地址。

服务器应默认地, 对于一个客户端, 在一秒之内至多对一个响应请求消息进行应答。当发送速率在数秒之内高于这个限制, 则建议使用漏桶算法, 但平均速率应默认地限制到每秒每客户端一个消息。服务器应能实现针对客户端 IP 地址的管理控制, 也允许特定的客户端发送多个快速的响应请求。

如果服务器使用不同的策略用于不同的 IP 地址, 它应请求客户端发送初始消息, 并返回一个不可预测的会话 ID 以防止欺骗攻击, 这是一个当速率超出限制时的绝对要求。

4.6 IANA 考虑

IANA 要求在 1024~49151 范围内设定 UDP 用户端口号, 为该协议使用; IANA 也需提供对消息类型和选项类型的注册服务。消息类型取值范围在 0~255。0~191 为该协议相关的标准使用, 类型 192~255 为实验使用。选项类型取值范围在 0~49151 数值上用于该协议的相关标准使用, 而 49152~65535 用于实验使用。

4.7 安全考虑

一个主机可能以其他主机的地址为源地址发送响应请求, 然后令互联网上任意一个组播 Ping 服务器发送报文到其他主机。这种行为应不会带来什么恶果, 最坏的情况是如果接收单播响应应答的主机恰巧也加入了所用的组播组, 主机将会接收到两份应答。

对于 ASM (任意源组播), 主机也可能使组播 Ping 服务器发送组播报文用于其他用途, 如扰乱其他人使用这个组。但是, 服务器实现应允许管理和限制服务器响应的组。这时主要的考虑是带宽, 为了限制使用的带宽, 服务器应默认地实现速率限制, 即一个响应请求每秒钟最多只被响应一次。

为了帮助防止欺骗, 服务器应要求客户端发送一个初始消息, 然后返回一个不可预测的会话 ID。该 ID 应与 IP 地址相关, 并有有限的寿命。然后, 服务器应只对与响应请求源地址相关的具有有效会话 ID 的响应请求消息进行应答。

5 组播路由探测

5.1 概述

单播路径探测允许追踪网络中一个节点到另一个节点的路径, 采用的是 ICMP TTL 超时消息。组播路径探测 (Mtrace) 允许追踪 IP 组播路由路径。Mtrace 需要组播路由器的支持, 而由诊断分析程序访问, 提供除了路径追踪之外的报文速率和报文丢失等其他信息。

Mtrace 的主要特点有以下 4 个。

- a) 可追踪组播报文从某个组播源到某个目的地 (即组播接收者) 所走的路径;
- b) 可以定位报文丢失问题 (如拥塞);
- c) 可定位配置问题 (如 TTL 门限);
- d) 尽可能减少报文发送 (如不会导致泛洪或报文数量急剧增加)。

其基本机制为：请求 Mtrace 的查询者（或客户端，不要求一定是组播源或组播接收者）发送 Mtrace 查询（Query）报文到与组播接收者相连的末跳组播路由器，末跳路由器将查询报文转换为请求（Request）报文，在请求报文中添加包含接口地址和报文统计特性等信息的响应数据块（Response Data Block），将请求报文向给定组播源、组播组的转发路径的前跳路由器（处于末跳的上游）单播发送，每一跳路由器都在请求报文的后面添加本跳的响应数据块，单播向前跳路由器（处于本跳的上游）发送。当请求报文到达首跳路由器（组播源所在的网络与之相连）时，首跳将请求报文转换为响应（Response）报文，并将完整的响应发送给响应目的地址。如果路径发生异常（如不存在组播路由等），响应也有可能到达首跳之前返回给响应目的地址。

Mtrace 使用路由器中任何可用的信息试图确定组播转发路径的前跳。不同种类的组播路由协议维护的协议状态数量不同，Mtrace 尽可能利用这些协议提供可用信息。例如如果 DVMRP（见 IETF RFC1075）路由器不存在到某个源有效的状态但是存在一条 DVMRP 路由，则该 DVMRP 路由器将选择 DVMRP 路由的父节点作为前一跳；如果 PIM-SM（见 IETF RFC4601）路由器处于（*,G）树上，则该 PIM-SM 路由器选择到汇集点 RP 的父节点作为前跳，这时虽然不存在源/组特定状态，路径仍可以得到追踪。

5.2 报文格式

Mtrace 消息以类型-长度-值（TLV）格式编码，如果一个实现接收到长度超过长度字段限定的 TLV 长度，则长度限定范围内的 TLV 可被接受，而超过长度的数据部分应被忽略。TLV 格式如图 6 所示。

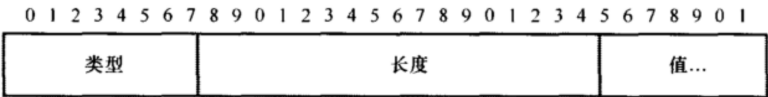


图6 TLV格式

图 6 中。

- 类型：报文类型（8 比特）；
- 长度：报文长度（16 比特）；
- 值：内容长度可变，取决于报文类型。

Mtrace 目前定义了四种 TLV，见表 1。

表 1 Mtrace TLV

代码	类型
1	Mtrace查询或请求
2	Mtrace响应
3	Mtrace标准响应块
4	Mtrace增量响应块

Mtrace 消息应包含一个 Mtrace 查询（请求）或响应，可以包含一个或多个标准或增量响应块。发送 Mtrace 请求的每个组播路由器必不能包含多个 Mtrace 标准块，但是可以包含多个增量响应块。

类型字段取“0x1”时为 Mtrace 查询或请求，当向查询者发送响应报文时类型字段设为“0x02”。

5.2.1 Mtrace 查询报文头

Mtrace 报文携带在 UDP 报文中。UDP 源端口号由本机操作系统唯一地选择。UDP 目的端口号为 IANA 预留的 Mtrace 端口号。Mtrace 消息中的 UDP 校验和必须是有效的。

Mtrace 消息包含公共的 Mtrace 查询报文头，查询报文头如图 7 所示。查询报文头由生成查询报文的查询者填充，过渡路由器不允许修改查询头的任何字段。

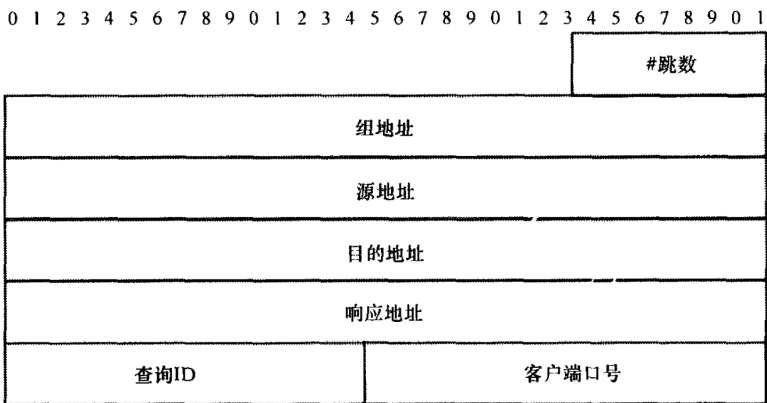


图7 查询报文头

图 7 中。

- #跳数：8 比特，限定查询者希望追踪的跳数。如果在 mtrace 请求到达首跳路由器前出现了错误条件，可以设定该字段进行扩展环搜索，直至到达问题节点之前的路径。
- 组地址：待追踪的组播地址，IPv4 长度为 32 比特，IPv6 长度为 128 比特。如果不限定特定组地址信息，IPv4 地址设定为全“1”，IPv6 地址设定为(::)。但需注意对于不限定组的情况，必须限定源地址。
- 源地址：待追踪路径的组播源地址，IPv4 为 32 比特，IPv6 为 128 比特。若不限定组播源，IPv4 地址设为全“1”，IPv6 地址设为(::)。需注意对于不限定组播源的 Mtrace，对某些路由协议是不适用的(如 PIM-SSM)[RFC4607]。
- 目的地址：待追踪路径的组播接收者，IPv4 为 32 比特，IPv6 为 128 比特，路径追踪从接收者开始，向组播源的方向进行。
- 响应地址：完整的 Mtrace 响应报文发送的地址，IPv4 为 32 比特，IPv6 为 128 比特。响应地址必须为全球范围的单播 IP 地址。
- 查询 ID：16 比特，作为 Mtrace 请求报文的唯一标识符，以检测重复或延迟的响应，和最小化报文的冲突。
- 客户端口号：发送 Mtrace 响应报文所使用的 UDP 端口号，路由器在接收到响应报文时必不能修改该 UDP 端口号。

5.2.2 IPv4 Mtrace 标准响应块

在追踪路径上的每个过渡 IPv4 路由器将该标准响应数据块添加到要转发的 Mtrace 请求报文中时，标准的响应数据块的格式，如图 8 所示。

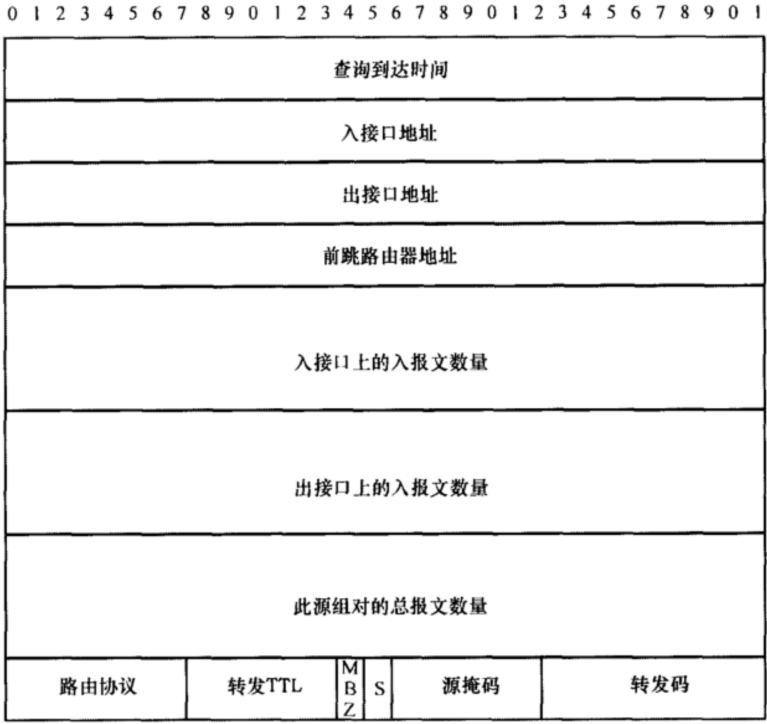


图8 IPv4响应数据块的格式

图 8 中。

- 查询到达时间（32 比特）：为 32 比特的 NTP 时标，限定 Mtrace 请求到达此路由器的时间。其中低 16 比特为整数部分，高 16 比特为分数部分。例如 Unix 时间到 NTP[RFC1305]时标的转换公式为：查询到达时间= (tv.tv_sec + 32384) << 16 + ((tv.tv_usec << 10) / 15625)；其中常数 32384 为从 1900 年 1 月 1 日到 1970 年 1 月 1 日的整秒数截取为 16 比特数得到的数值。((tv.tv_usec << 10) / 15625)为((tv.tv_usec / 100000000) << 16)的被截取的部分。
- 入接口地址（32 比特）：该源-组组播数据报文到达的接口地址。如果未知则设为 0。
- 出接口地址(32 比特)：该源-组的组播数据报文流向目的地址所经过的接口。如果未知则设为 0。
- 前跳路由器地址(32 比特)：本路由器希望报文从组播源到达所经过的路由器的地址，如果前跳未知则设为一个组播地址（如所有路由器地址—224.0.0.2）。但如果入接口未知则设为 0。
- 入接口上的入报文数量（64 比特）：入接口上接收到的所有源和组的组播报文数量。如果无数量可报则设为全“1”，该计数器的数值可以与该接口的接口 MIB 表[RFC2863]中的 ifHCIn MulticastPkts 值相等。
- 出接口上的出报文数量（64 比特）：出接口已发送的或排队等待发送的所有源和组的组播报文数量，如果无数量可报则设为全“1”，该计数器的数值可以与该接口的接口 MIB 表的 ifHCOut MulticastPkts 值相等。

- 该源-组对的报文总数（64 比特）：该路由器转发的从该组播源到达该组播组的报文数量，若无数量可报则设为全“1”。如果后面的 S 比特为 1，则该数量是针对源网络，源网络由“源掩码”字段限定；如果 S 比特为 1，同时“源掩码”字段设为 63，则表明没有源特定状态，则报文数量是针对与发向该组播组的所有源的。该计数器值应该与对应转发项的 IPMROUTE-STD-MIB[RFC5132]中的 ipMcastRoutePkts 值相同。
- 路由协议（8 比特）：描述本路由器和前跳路由器使用的路由协议，已定义的数值包括。
 - 0 未知
 - 1 PIM[RFC4601]
 - 2 使用特殊路由表的 PIM
 - 3 使用静态路由的 PIM
 - 4 使用 MBGP[RFC2858]路由的 PIM
 - 5 使用由断言（Assert）处理创建状态的 PIM
 - 6 双向 PIM[RFC5015]
 - 7 IGMP/MLD 代理[RFC4605]
 - 8 AMT 中继
 - 9 AMT 网关

为获得该项数值，组播路由器读取 IPMROUTE-STD-MIB 中限定的 ipMcastRouteProtocol, ipMcastRouteRtProtocol, 和 ipMcastRouteRtType 值，根据这些 MIB 值确定上述路由协议值。

- 转发 TTL（8 bits）：包含报文在出接口转发时要求设定的 TTL 值。
- MBZ(1 比特)：应在发送时设为 0，在接收时忽略。
- S(1 比特)：该标志用于表示指明源-组对对应的源网络的报文数量，源网络由“源地址”和“源掩码”字段共同确定。
- 源掩码（6 比特）：路由器为该组播源设定的网络掩码中“1”的数量。如果路由器只根据组状态进行转发，则该字段设为 63(0x3f)。
- 转发码（8 比特）：包含转发信息/错误码，已定义的数值，见表 2。

表 2 转发码

值	名称	描述
0x00	NO_ERROR	无错误
0x01	WRONG_IF	Mtrace请求报文从一个不会为该源、组和目的转发的路由器接口到达。
0x02	PRUNE_SENT	对于Mtrace请求中标识的源/组，该路由器已经发送一个剪枝到上游路由器。
0x03	PRUNE_RCVD	该路由器已经停止为此源/组进行转发，原因是收到了下一跳（下游）路由器的剪枝请求。
0x04	SCOPED	该组在本跳受到管理范围的限制。
0x05	NO_ROUTE	该路由器不存在到此源或组的路由，并且无法确定潜在的路由。
0x06	WRONG_LAST_HOP	该路由器不是合适的末跳路由器。

表 2 转发码（续）

值	名称	描述
0x07	NOT_FORWARDING	该路由器由于无法限定的原因不向出接口转发对应该源-组对的报文。
0x08	REACHED_RP	到达了汇集点RP或核
0x09	RPF_IF	Mtrace请求到达对应该源-组对预期的RPF接口。
0x0A	NO_MULTICAST	Mtrace请求到达不支持组播的接口。
0x0B	INFO_HIDDEN	一个或多个跳为此Mtrace隐藏了。
0x81	NO_SPACE	已没有足够的空间在报文中插入新的响应数据块
0x82	OLD_ROUTER	前跳路由器不理解Mtrace请求报文
0x83	ADMIN_PROHIB	Mtrace被管理禁止

请注意如果一个路由器发现已没有足够的空间插入其响应数据块，它会在前跳路由器的转发码字段设定 0x81 错误码，覆盖前面路由器已设定的任何错误代码。在路由器发送响应到报文头中的响应地址以后，Mtrace 客户端可以在报文中列出的末跳重新启动 Mtrace。转发码中的 0x80 比特可用于指明一个致命错误，致命错误是指路由器可能知道前跳但不能够将 Mtrace 请求报文转发给它。

5.2.3 IPv6 Mtrace 标准响应块

追踪路径上的每个过渡 IPv6 路由器将“响应数据块”添加到要转发的 Mtrace 请求报文后面，标准响应数据块的格式如图 9 所示。

0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1
查询到达时间																															
入接口ID																															
出接口ID																															
本地地接																															
远端地址																															
如接口上的入报文数量																															
出接口上的入报文数量																															
此源组对的总报文数量																															
路由协议								MBZ				S	源前缀长度								转发码										

图9 IPv6响应数据块的格式

图 9 中。

- 查询到达时间 (32 比特)：与 5.2.2 中定义的相同。
- 入接口 ID (32 比特)：该源-组组播数据报文预期到达的接口 ID。如果未知则设为 0。该 ID 值应与 IF-MIB 中为该接口定义的 InterfaceIndex 值相同，以网络字节顺序表示。
- 出接口 ID (32 比特)：该源-组组播数据报文流向目的地址所经过的接口 ID。如果未知则设为 0。该 ID 值应与 IF-MIB 中为该接口定义的 InterfaceIndex 值相同，以网络字节顺序表示。
- 本地地址 (128 比特)：唯一标识该路由器的全球 IPv6 地址。
- 远端地址 (128 比特)：前跳路由器地址。当前跳未知时设为所有路由器地址 (FF02::2)，但如果入接口未知则设为 0。
- 入接口上的入报文数量 (64 比特)：与 5.2.2 中定义的相同。
- 出接口上的出报文数量 (64 比特)：与 5.2.2 中定义的相同。
- 该源-组对的报文总数 (64 比特)：该路由器转发的从该组播源到达该组播组的报文数量，若无数量可报则设为全“1”。如果 S 比特为 1，则该数量是针对源网络，源网络由“源前缀长度”字段限定；如果 S 比特为 1，同时“源掩码”字段设为 255，表明没有源特定状态，则报文数量是针对与发向该组播组的所有源。该计数器值应该与对应转发项的 IPMROUTE-STD-MIB 中的 ipMcastRoutePkts 值相同。
- 路由协议 (8 比特)：与 5.2.2 中定义的相同。
- MBZ (1 比特)：必须在发送时设为 0，在接收时忽略。
- S (1 比特)：标志用于表明源-组对对应的源网络的报文数量，源网络由“源地址”和“源前缀”字段共同确定。
- 源前缀长度 (8 比特)：该路由器为该源包含的前缀长度。如果路由器只根据组状态进行转发，则该字段设为 255 (0xff)。
- 转发码 (8 比特)：与 5.2.2 中定义的相同。

5.2.4 Mtrace2 增量响应块

除了标准响应块，路径上的组播路由器当发送请求到上游路由器或发送响应到响应地址时可以添加“增量响应块”。使用增量响应块可以灵活地添加不同的信息，如图 10 所示。

增量响应块以类型为 0x04 的 Mtrace TLV 头开始，其 16 比特类型字段有各种不同的用途，如分析和协议验证等。

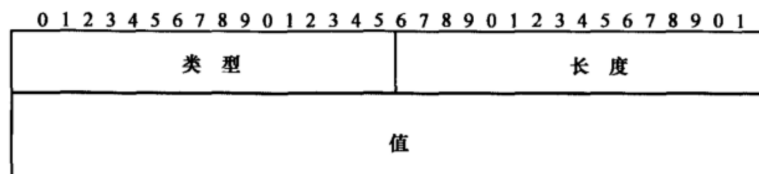


图10 增量响应块

5.3 路由器行为

5.3.1 概述

路由器在转发 Mtrace 报文时需检查 TTL 值或跳数值是否符合要求，并检查确定报文不是发给自己的。

5.3.2 Mtrace 查询报文的处理

Mtrace 查询消息没有填充任何响应数据块，它的 TLV 类型为 0x1。

当接收到 Mtrace 查询消息后，路由器必须检查该报文以确认自己是否是报文中到目的地址的最后一跳。如果它在目的地址相同的子网上有一个组播使能的接口，说明在网络中有活动的组成员，并且该路由器能够转发 (S,G) 或 (*, G) 相关的组播流量，据此判定自己为合适的末跳路由器，则在请求报文中封装本跳的响应数据块，将请求报文向位于转发路径上游的前一跳发送。

如果路由器确定它不是合适的末跳路由器，如该路由器不在目的网络中，或该网络中没有活动的组播成员，或者如果它不能够判定自己是否是末跳路由器，则该路由器将根据查询是通过单播还是组播方式接收的采取不同的操作：如果查询通过组播方式接收，则路由器必须默默地将报文丢弃；如果查询报文通过单播方式接收，则路由器生成一个转发码为 WRONG_LAST_HOP 的响应报文，向响应地址发送。

如果路由器在目的地址所在的网络中，并且它并不转发组播数据流，说明该路由器是到达接收者的单播末跳而不是组播末跳，如果该路由器判定该目的网络中有活动的组播成员，则将查询报文的目的地地址设为所有路由器地址 (IPv4 为 224.0.0.2, IPv6 为 FF02::2)，TTL 值设为 1，将其发送到接收者所在的网络中。组播末跳路由器接收到该查询报文后，生成请求报文向其前跳发送。

重复的查询消息可通过 (IP 源地址，查询 ID) 识别，并应被忽略。这可以通过简单的 1 个存储堆的 cache 实现，即保存前一个被处理的查询消息的 IP 源地址和查询 ID，而忽略新出现的同源同查询 ID 的查询报文。但重复的请求消息不能以这种方式被忽略。

5.3.3 Mtrace 请求报文的处理

当路由器接收到 Mtrace 查询报文，并可以确定自己就是合适的末跳路由器时，它将这个查询报文当作请求报文。Mtrace 请求是填充响应块的报文，其 TLV 类型为 1，与查询报文相同，路由器通过检查报文的长度确定报文的类型是查询还是请求。

如果路由器判定 Mtrace 请求不是发向这个路由器的，则路由器必须将请求报文默默地丢弃。对 Mtrace 请求的转发方法是：如果前跳路由器已知并且响应块的数量少于请求的数量（即 Mtrace 头中的“#跳数”字段），报文被发向前跳路由器；如果入接口已知、但前跳路由器未知，报文发向入接口的一个合适的组地址上。组地址的取值取决于使用的路由协议（如对于 IPv4 为 224/24，对于 IPv6 为 FF02::/16），可以为所有路由器地址（对于 IPv4 为 224.0.0.2，对于 IPv6 为 FF02::2），但必不能为所有主机地址（对于 IPv4 为 224.0.0.1，对于 IPv6 为 FF02::1）。如不满足上述条件，路由器需封装相应的错误转发码，将响应报文向查询者发送。

当路由器接收到 Mtrace 请求时，它执行如下操作（需注意有可能转发码的场景发生过多次，则只报告最先发生的场景）。

a) 如果路由器有足够的空间，则它将插入一个新的响应块到请求报文中，并在响应块中设定“查询到达时间”、出接口地址（IPv4）或出接口 ID（IPv6），出报文数量、转发 TTL（IPv4）。如果没有足够的空间，则在前跳响应块的转发码中填充 NO_SPACE，并将响应报文向查询者发送。

b) 确定对于特定源和组的转发信息，采用的处理方式与接收组播报文的处理相同。转发状态不一定要实例化，可以专为 Mtrace 生成虚状态。如果使用共享树协议并且不存在源特定转发状态，或如果源是以全“1”表示的，则需使用组状态。如果不存在组状态或组地址为 0，则应使用可能的源状态（即源特定加入将要使用的路径）；如果该路由器为核或 RP 并且存在着非源特定信息，若此时尚未设定转发码，将转发码设为 REACHED_RP。

c) 如果不能确定转发信息，若此时尚未设定转发码，将转发码设为 NO_ROUTE，将其他未设定字段设为 0，将响应报文向查询者转发。

d) 根据转发信息填充入接口地址，前跳路由器地址，入报文数量，总报文数量，路由协议，源掩码字段。

e) 如果 Mtrace 被管理禁止或前跳路由器不理解 Mtrace 请求，若此时尚未设定转发码，将转发码设为 ADMIN_PROHIB 或 OLD_ROUTER；如果 Mtrace 被管理禁止，则步骤 4 中的填充的所有信息都认为是私有信息，将被重新清 0，然后将响应报文向查询者发送。

f) 如果接收接口没有使能组播，若此时尚未设定转发码，将转发码设为 NO_MULTICAST；如果接收接口是路由器预计组播数据从组播源到达的接口，若此时尚未设定转发码，将转发码设为 RPF_IF，否则，如果接收接口不是路由器从组播源转发组播数据的接口，若此时尚未设定转发码，将转发码设为 WRONG_IF。

g) 如果管理范围的组播组或者在出接口或者在入接口上，若此时尚未设定转发码，将转发码设为 SCOPED。

h) 如果该路由器为组的汇集点或核，若此时尚未设定转发码，将转发码设为 REACHED_RP。

i) 如果该路由器已经向上游发送了一个对应于该源和组的剪枝消息，若此时尚未设定转发码，将转发码设为 PRUNE_SENT；如果该路由由于接收到下游的下一跳路由器发送的剪枝消息而停止了转发，若此时尚未设定转发码，将转发码设为 PRUNE_RCVD。如果路由器应正常地转发组播流量而实际没有转发，若此时尚未设定转发码，将转发码设为 NOT_FORWARDING。

j) 如果组播转发状态正常，该路由器生成新的请求报文发向前跳路由器。该路由器可使用相应的确认机制，向下游路由器或查询者发送确认报文，表明本跳已经正常地对请求报文进行了处理。确认报文与被确认的查询或请求报文有相同的查询头，以不同的类型码与查询或请求报文区别。如果本跳在预定时间间隔内未收到来自前跳的确认，则判定上游网络可能发生异常，需将接收到的请求文转换为响应报文并携带表示上游网络故障原因的转发码向查询者发送。

5.3.4 Mtrace 响应报文

路由器应正常地转发所有的 Mtrace 响应报文，而无需特殊处理。如果路由器自己发起了 Mtrace 查询或请求报文，它也需处理响应报文以使得该报文得以正常转发。

响应报文在 Mtrace 请求到达首跳路由器时，由首跳路由器向查询者发送。在某些情况下，由于设备或网络故障，请求报文不能够到达首跳路由器，如果过渡路由器能够感知故障，则可由过渡路由器向查询者响应。响应数据可由接收查询报文或请求报文的每个路由器向查询者响应，响应可携带转发路径上所有已被探测的路由器的响应数据块，也可以携带本跳路由器的响应数据块。查询者如果在预定时间间隔内没有收到响应报文，则可以判定组播路由异常。

路由器发送响应报文时，目的地址设为 Mtrace 头中的响应地址。如果响应地址为单播地址，则路由器在 IP 头中插入通常的单播 TTL 或跳数上限；如果响应地址是组播地址，则路由器将响应 TTL 或跳数上限从 Mtrace 头中拷贝到 IP 头中。

如果响应地址为单播地址，则路由器可使用其任意接口地址为源地址，因为某些组播路由协议基于源地址进行转发；如果响应地址为组播地址，则路由器必须使用组播路由拓补中已知的地址作为源地址。

当路由器确定响应报文的源接口时，响应报文必须在单个接口上发送，然后好似它在该接口上接收一样被转发。该路由器必不能在每个接口上各生成一个响应报文，以避免重复报文的产生。

5.4 客户端行为

5.4.1 Mtrace 查询的发送

当 Mtrace 的目的地址为运行 Mtrace 客户端时，Mtrace 查询报文将被发送到所有组播路由器地址（IPv4 为 224.0.0.2，IPv6 为 FF02::2），以保证报文被子网上的末跳路由器接收到。否则，如果 Mtrace 目的地已知合适的末跳，或者 Mtrace 客户端希望从响应 NO_SPACE 的过渡路由器重启 Mtrace 查询，查询报文将被单播发送到末跳或过渡路由器。

否则，查询报文将被组播到将要查询的组；如果 Mtrace 的目的地址是该组的组成员，查询报文会被发送到合适的末跳路由器上。最后，报文需要包含路由器告警（Router Alert）[RFC2113][RFC2711]选项，已确保不是组成员的路由器注意到该组播报文。

5.4.2 路径的探测和统计信息搜集

客户端可以发送少量包含大“#跳数”的初始查询消息，以尽可能追踪完整的路径。如果尝试失败，一个策略是执行线性追踪（即传统的单播 Traceroute 程序），将“#跳数”设为 1，然后设为 2，等等。如果在某跳接收不到响应，跳数可设为更大，以期更大的跳数将会响应。这种试图将继续直至用户定义的定时器超时为止。

如果客户端已经确认已经追踪了整条路径，它可以等待短短的时间在发起第二次追踪搜集统计特性。如果两个追踪的路径相同，则可以如 5.6 节所述，显示统计特性。

5.4.3 末跳路由器处理

Mtrace 查询者有可能不知道末跳路由器，或者路由器可能在禁止单播报文而不禁止组播报文的防火墙后面。在这些情况下，Mtrace 查询将被组播到要探测的组上。除了末跳之外的其他所有路由器将会忽略组播发送的 Mtrace 查询报文。组播发送的 Mtrace 请求需包含路由器告警选项。

另一种方法是将查询报文本播到追踪的目的地址上。单播到 Mtrace 目的地址的请求将包含路由器告警选项,以使未跳路由器能够注意到该报文。如果查询者与被请求的目的地与同一个路由器相连,如果未跳未知,则 Mtrace 查询应组播发送到 224.0.0.2 (ALL-ROUTERS.MCAST.NET)上。

5.4.4 首跳路由器处理

如果 Mtrace 查询者从首跳路由器单播不可达,在此情形下,查询者需要将 Mtrace 响应地址设为组播地址,将响应 TTL (或跳数上限)设为可从首跳路由器到达查询者的足够大的数值。可以将从一个小的 TTL 开始,随后递增 TTL 数值,直至到达一个合适的上限。

IANA 将 224.0.1.32(MTRACE.MCAST.NET)设为默认的 IPv4 Mtrace 响应使用的组播组,但是并未为 IP Mtrace 响应预留任何 IPv4/IPv6 组地址,因为 Mtrace 通常不以组播方式发送响应报文。

5.4.5 中断的过渡路由器

中断的过渡路由器可能不理解 mtrace 报文或将其丢弃,查询者则不会接收到 Mtrace 请求的响应。这时即可采取逐跳的探测,方法是递增设定响应数目直至得到一个响应(可采用线性或二进制搜索,但二进制搜索可能更慢,因为需要等待一个超时时间才可能定位错误)。

5.4.6 Mtrace 中止

当执行逐跳的扩展追踪时,需要确定何时中止扩展追踪,具体如下。

- 到达一个源:可根据 MTrace 报文中的最后一个路由器的入接口不为 0,而前跳路由器的入接口为 0 判定到达了组播源。
- 致命错误:如果 Mtrace 报文中的转发错误代码的 0x80 比特被设定,说明发生了致命错误。
- 无前跳:如果 Mtrace 报文中的最后一个前跳字段为 0,则不能继续追踪。
- 追踪路径短于预期:如果返回的路径短于请求的路径(即响应数据块的数量少于“#跳数”字段的值),则判定出现问题,追踪不能继续进行。

5.5 组播协议相关处理

5.5.1 PIM-SM

当组播 Mtrace 到达 PIM-SM RP 并且 RP 不再继续将 Mtrace 报文转发,这意味着 RP 未发起源特定加入,因而没有需要追踪的路径。但是,如果希望探测从源到 RP 的已被使用的转发路径,可将 MTrace 目的地址设为 RP,要追踪的组设为 0,将 Mtrace 查询报文本播到 RP。

5.5.2 双向 PIM

双向 PIM 是 PIM-SM 的一个变种,用来建立连接组播源到接收者的双向共享树。在双向共享树上,组播数据从源发送到 RPA (汇集点地址),然后从 RPA 发向接收者,无需源特定状态。与 PIM-SM 不同,RP 总有要追踪的状态。

给定 RPA 的指定转发者(DF)负责将下游流量向其转发链路发送,将上游流量从其链路发向 RPA 所属的 RPL (汇集点链路),因而 Mtrace 报告的是路径上的 DF 地址或 RPA。

5.5.3 PIM-DM

运行 PIM 密集模式 (PIM-DM) (见 IETF RFC3973) 的路由器并不知道报文走的路径, 除非报文正在转发。若没有额外的协议机制, 这意味这在共享介质上有分支点存在多个可能路径的情况下, Mtrace 只能追踪现有的路径, 而不能追踪潜在的路径。当有多个可能的路径, 但分支点不在共享介质上, 前跳路由器是可以知道的, 但最后一跳路由器并不知道自己是最后一跳。

当组播流量正在转发时, PIM 密集模式路由器知道自己是否时最后一跳转发者 (因为它们赢得或失去了断言 Assert 竞争), 并知道谁是前跳, 因而, Mtrace 在组播数据流量转发是总能够确定合适的路径。

5.5.4 IGMP/MLD 代理

当 Mtrace 查询报文当到达 IGMP/MLD 代理(见 IETF RFC4605)的入接口时, 它将 Mtrace 标准响应数据块的转发响应码设为 WRONG_IF(0x01)值, 并将响应数据块发送回响应地址。当 Mtrace 查询报文到达 IGMP/MLD 代理的出接口时, 它通过入接口向上游路由器转发。

5.5.5 AMT

AMT(见 IETF RFC7450)为不支持组播的网络提供组播连接性, 组播报文可封装在单播包文中从一个站点发送或接收。当 Mtrace 查询报文到达 AMT 网关的 AMT 伪接口时, AMT 网关将其封装为一个特定的 AMT 中继, 该 AMT 中继通过纯单播网络可达。

5.6 问题分析与故障定位

5.6.1 转发不一致性

转发错误代码可以用来确定一个组播组未预期地被剪枝或被管理限制。

5.6.2 TTL 或跳限问题

通过取所有跳数的最大值, 可以发现从源到达目的 TTL 或跳限。

5.6.3 报文丢失

通过两个 Mtrace, 比较报文中前跳的入报文数量和出报文数量的差异获得报文丢失信息。在点到点链路上, 这些数值的差异意味着可能发生的报文丢失。虽然 Mtrace 查询在传递过程报文数量可能会变化 (出入为 1 个、2 个或更多报文), 但当测量周期扩展时不会累计。但在共享链路上, 前跳的入报文数量可能比出报文数量要多, 原因是链路上的其他路由器或主机正在注入报文, 这表现为“负丢失”而可能会屏蔽真正的报文丢失。

除了要包含接口上所有的入和出的报文数量, 响应数据还需包含每个节点为特定源-组对转发的报文数量。比较两个 Mtrace 之间这些数值的差异, 然后比较两跳之间的这些值的差异可以测量从特定源以特定组发向特定接收者的报文丢失, 这项测量不会受到共享链路的影响。

在组播隧道的点到点链路上, 报文丢失通常是因沿着隧道路径上的单播路由的拥塞导致的。在不适用隧道的通常的组播链路上, 丢失很可能在某一跳的出队列中发生 (如由于优先级导致的丢弃) 也有可

能发生在下一跳的入队列中。响应数据的计数器不能够区分这些情况。单个节点上的入、出接口之间的报文数量的差异通常用来测量节点的队列溢出特性。

5.6.4 链路利用率

同样,通过两个 Mtrace,可以将同一跳的入报文数量和出报文数量之差除以对应的时标之差获得链路的报文速率。如果平均报文大小是已知的,则链路利用率也可用来估计报文丢失是因速率上限还是由于物理容量超出导致的。

5.6.5 时延

如果路由器有同步的时钟,可以根据相邻节点之间的时标之间的差异估计传播延迟和排队延迟。但是,该延迟包括控制处理开销,因而并不一定是组播数据传播延时的真实反应。

5.7 IANA 考虑

5.7.1 转发码

新的转发码应由修改该文档的 RFC 创建,完全描述新的转发码使用的条件。IANA 可以作为唯一的查询转发码定义的中心机构。

5.7.2 UDP 端口号和 IPv6 地址

IANA 应在 RFC 发布时为 Mtrace 分配 Mtrace UDP 目的端口号。

5.8 安全考虑

5.8.1 拓补发现和隐藏

Mtrace 可以用来发现在用的组播拓补。如果网络拓补是秘密的, Mtrace 可在区域边界受限,使用 ADMIN_PROHIB 转发码。如果一个域的拓补信息和连接性信息可能对 Mtrace 查询者是不可见的,此时可使用 INFO_HIDDEN 转发码。

5.8.2 流量速率

Mtrace 可用来发现哪些源向哪些组以何种速率发送组播流量。如果这个信息是秘密的, Mtrace 可限制在区域的边界,使用 ADMIN_PROHIB 转发码。

参 考 文 献

- [1] IETF draft-ietf-mboned-mtrace-v2-12 MTrace 版本 2: IP 组播路由探测协议 (Mtrace Version 2: Traceroute Facility for IP Multicast)
-

中华人民共和国通信行业标准
IP 组播 Ping 与路径探测协议
YD/T 3308—2017

*

人民邮电出版社出版发行
北京市丰台区成寿寺路 11 号邮电出版大厦
邮政编码：100064
北京康利胶印厂印刷
版权所有 不得翻印

*

开本：880 × 1230 1/16 2018 年 6 月第 1 版
印张：2 2018 年 6 月北京第 1 次印刷
字数：48 千字

15115 • 1403

定价：20 元

本书如有印装质量问题，请与本社联系 电话：(010)81055492