Network and Protocol Team
Internet-Draft
Expires: December 11, 2006

M. Hoerdt
Q2S
B. Hilt
GRTC
J-J. Pansiot
LSIIT
June 9, 2006

Simple join failure notification for PIM-SM multicast routing
draft-hoerdt-pim-group-unreachable-00.txt

Status of this Memo

Copyright Notice

Abstract

Currently, PIM-SM is the most widely deployed multicast routing protocol in the Internet. However in IPv6 this multicast protocol lacks a simple connectivity debugging tool, making it difficult to identify multicast routing problems. When a host wants to receive a multicast group or channel, if the corresponding PIM-Join message

cannot be propagated in the network, nothing exists to inform the receiver or the network about this failure. We propose a simple error notification message based on ICMP/ICMP6 which is able to indicate where and why the error happened in the network without using a multicast traceroute mechanism.

Table of Contents

1. Introduction

With the current specification of PIM-SM [1], when a router cannot propagate a join message, nothing can be done to inform either receivers or downstream routers. Therefore a useless multicast tree is maintained from receivers to the point of failure, while receivers needlessly wait for multicast data. Such a failure may be due for example to a temporary routing problem (a source or RP is not currently reachable), a configuration problem (RP adressing, ...) or a user error (the source address of an SSM channel is not an existing address).

The basic proposed idea is to send a downstream error message when a Join message propagation fails at some point in the network. This may occur towards a source (S) both for a PIM-SM[5] SSM channel or for a SPT tree for a PIM-SM[1] ASM group; This may also occur between the DR and the rendez-vous point (RP) for an ASM group. This error message, later on called Group Unreachable or GU, is carried over the ICMP/ICMP6[2][3] protocol. This draft specifies:

- the rules used to send and propagate this message down the multicast tree.
- the ICMP/GU message format.
- the proposed error codes.

The message is forwarded hop-by-hop by routers, it is not sent to the group address contained in the failing join. The rules to process this message by multicast hosts is not in the scope of this draft.

This mechanism is simpler than a multicast traceroute but still very useful. It provides multicast listeners with a report including both problem location and problem description. It helps network administrators reporting the problem to the person responsible of the good operation of multicast routing for the indicated location. The proposed message could be used by diagnostic applications such as ssmping or asmping[6] to locate the point of failure in case of unreachability. Some user applications could also make use of it, for example to warn the user and/or to leave the corresponding group or channel. As this mechanism requires to be deployed on all multicast routers to operate properly, it requires standardisation.

2. Protocol description

When a router R fails to send a Join message towards a rendez-vous point RP or a source S, it triggers the sending of an ICMP/ICMP6 Group Unreachability (GU) message. GU messages MAY aggregate Join failure reasons for several unreachable RP(s)/source(s). In order to locate the point of failure, the GU message contains the address of R.

To inform all the downstream tree routers and hosts about a failure, a GU message is propagated hop-by-hop downstream the multicast tree concerned by the failure (the tree corresponding to the failing Join message). Along these trees, this information could be taken into account in order to reduce the cost of useless/problematic multicast branches. This is beyond the scope of this document.

2.1. Sending GU Messages

GU messages are sent both when a failure occurs on an existing tree or when a new branch is under construction. In both cases, each failed Join message triggers the sending of a GU message. If the failure is due to a transient error such as a temporary routing problem, the router may choose to suppress the GU message.

On interdomain links and routers, the number of maintained groups/channels can be high. A single failure on those links may trigger a high number of GU messages sharing the same point of failure. To reduce the cost of the protocol in this case, GU messages sent on a link MAY aggregate unreachability information for several groups/channels having receivers through this link.

The GU message size is limited to the link MTU. If the number of groups/channels to aggregate exceeds this size, unreachability information is fragmented into multiple GU messages. To avoid the sending of a burst of GU messages, the messages should be rate limited.

When a multicast router R is unable to send a group/channel Join message, it builds for each failed group/channel a Group Record containing its failure information like group/channel identification, error code, etc... (detailled format in Section 3). For each outgoing interface of at least one failing tree, the corresponding Group Records are gathered in the same GU Record.

Then a GU message, possibly aggregating several GU records for this interface is built. The destination address of the GU messages using this interface MUST be All_Systems_on_this_Subnet group address (224.0.0.1) in IPv4, or ALL_NODES_ADDRESS group address (FF02::1) in IPv6. This ensures that all receivers for the failing group will received the GU message, together with all PIM Routers on the same interface. A further study could determine if the use of a specific multicast address would be more efficient. In this case the destination address would be All_Multicast_Reception_Hosts group address (224.0.0.X) in IPv4 and ALL_MULTI_REC_NODES group address (FF02::X) in IPv6 (X to be defined by IANA).

2.2. Receiving/Forwarding GU Messages

GU messages are forwarded hop-by-hop according to the TIB (Tree Information Base) maintained by the multicast routing protocol. When a router R receives a GU message on an interface:

- If the source address of the received message is not the address of a PIM neighbor for the interface the GU message MUST be silently ignored.
- If there is no group/channel in the GU message for which the receiving interface is the incoming interface, the message must be silently ignored.

The GU message is forwarded on all interested interfaces after a possible trimming. An interface i is interested by the GU if and only if the GU contains a group or channel such that:

(1) the GU was received by the incoming interface for this group/channel

(2) interface i is an outgoing interface for this group/channel. The GU sent through interface i is trimmed such that it contains only groups/channels satisfying conditions (1) and (2).

(3) Channel unreachability information (concerned by specific source S and group G) is only forwarded hop-by-hop down to a shared tree (*,G) using TIB states if the bit R is set in the GU Record. RP unreachability information is propagated on the shared (*,G) tree without setting this bit.

Normally, if a join (S,G) fails the GU is sent only down the (S,G) tree to the DRs (and eventually to the receivers). In some ASM cases a join (S,G) may fail between the RP and the FHR. In this case the GU will travel down to the RP, and if the RP has no outgoing interface for (S,G), the GU will stop there. The RP may choose to propagate the GU(S,G) message down the shared tree. If it do so, the RP MUST take care that this message can be distinguished from regular GU(S,G) messages. The R bit tells that. Note that such a message may help debug a problem between the RP and the source, but this problem may have no direct impact on receivers, if they reach the source by another route.

This bit is set by the router propagating a GU(S,G) through the shared tree, and then this GU should be transparently forwarded hop-by-hop down to the shared tree.

Several GU messages for the same interface, possibly originating from different routers MAY be aggregated in the same ICMP message.

3. Message format

Join messages sent by a router summarize for a given interface the group(*,Gn)/channel(Sn,Gn) join state information. These messages are sent to maintain a connection towards the root of the tree (Rendez-vous Point or Source). To inform downstream tree members of a Join propagation failure, each group/channel which cannot be connected to its tree root will be included into a GU message.

GU messages are a part of ICMP (in IPv4) or ICMP6 (in IPv6) messages.

Over IPv4, GU messages are encapsulated in IPv4 datagrams with a protocol number of 1. Every GU message is sent with an IP Time-To-Live of 1. IPv4 GU message MUST only contain unreachability information about IPv4 addresses.

Over IPv6, GU messages are encapsulated in IPv6 packets with a Next Header value of 58. An IPv6 GU messages MUST be sent with a link-local source address, and hop limit set to 1. IPv6 GU message MUST only contain unreachability information about IPv6 addresses.

A GU message uses the common ICMP/ICMP6 structure. Each ICMP/GU messages can carry several GU records concerning each an error location (i.e. router interface). In the following text we use ICMP/GU to refer to ICMP/GU or ICMP6/GU messages. So, information specific to a protocol family (IPv4/IPv6) will be specified explicitely.

A single Group Records informs about a single group address associated with a single problem notification. GU Records agregate Group Records with the same failing router address. GU messages agregate GU Records which MAY come from different router address. Respective figures 1,2 and 3 explicit the format for GU messages, GU Records and Group Records.

New Type and Code values must be used to identify new ICMP/ICMP6 Group Unreachability (GU) messages. Type and Code values must be defined by IANA. This GU header can be common to several GU messages (until reaching MTU value). Each of them regarding failure on a specific router.

GU Records on figure 2 have the format presented for 32 bits IPv4 addresses. When using IPv6, all addresses will have a size of 128 bits.

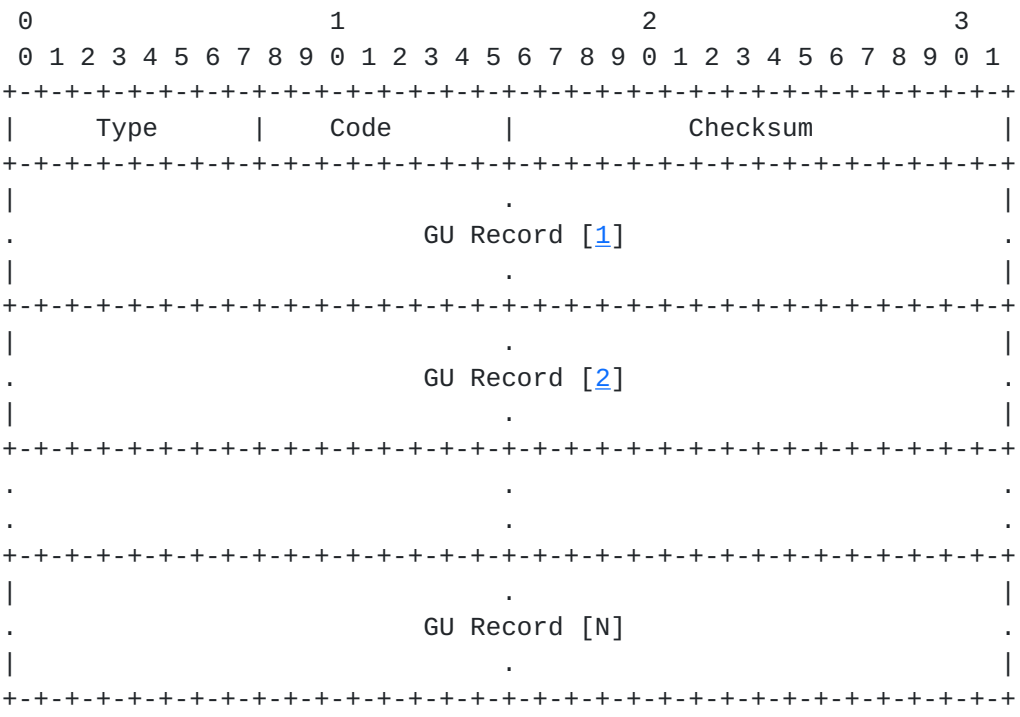

Figure 1 : GU Messages format

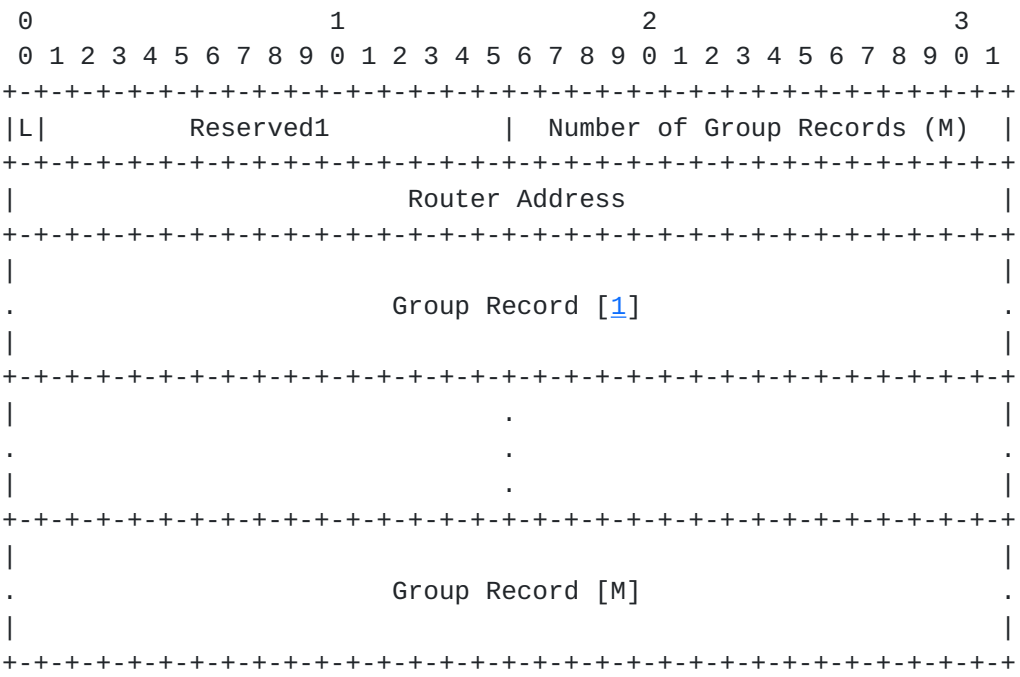

Figure 2 : GU Record format

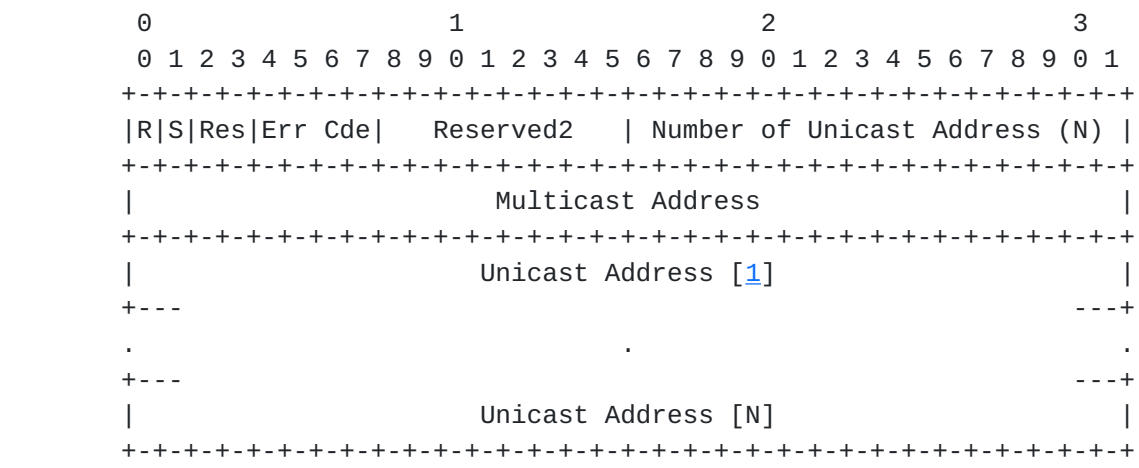

Figure 3 : Group Record format

3.1. Type and Code

Identifies the GU messages in the ICMP/ICMP6 messages family.

3.2. Checksum

The checksum is the 16-bit one's complement of the one's complement sum of the whole ICMP/ICMP6 GU message (the entire IP payload). For computing the checksum, the Checksum field is set to zero. When receiving packets, the checksum must be verified before processing the message.

3.3. L (The 'Last GU Record' bit)

When an ICMP/GU message carries several GU Records, flag L of the last GU Record MUST be set to 1. All intermediate GU Records MUST have flag L set to 0. If an ICMP/GU message carries a single GU Record, flag L of this record MUST be set to 1.

3.4. Reserved1

The Reserved1 field is reserved for future usage. It is set to zero on emission and ignored on reception.

3.5. Number of Group Records

The Number of Group Records (M) specifies how many Group Records (information regarding unreachable group/channel) are present in the GU Record.

3.6. Router Address

The Router Address is used to identify the router on which the Join message propagation failure occurred. This address should be chosen among the routers addresses with the largest scope (ie public addresses for IPv4, global addresses for IPv6 if available).

3.7. Group Record

Each Group Record of a GU Record contains information corresponding to a group/channel that could not be propagated in a Join message by the router designated by the Router Address (section Section 3.6).

3.8. Error Code

The four fields R,S,Res,Err Cde (Error Code) pertains to the same global error code which gives an indication about the type of failure occuring during the Join message propagation.

3.8.1. R (The 'shaRed' tree bit)

The R bit is set when a GU information triggered by a Join (S,G) must be propagated in the shared tree. It is typically used to propagated a failure related to the PIM-SM RP-SPT Switch.

3.8.2. S (The 'Source' or RP bit)

The S bit indicates if the Join error is related to a forwarding error towards a Rendez-vous Point (S is set to 0) or towards a Source (S is set to 1).

3.8.3. Res

These two bits are reserved for a future usage. They are set to zero on emission and ignored on reception.

3.8.4. Error Cde

The Err Cde (Error Code) field is a code that defines a more precise error type if known. A first list of possible error Code values and their signification is given in Section 4.

3.9. Reserved2

The Reserved field is reserved for future usage. It is set to zero on emission and ignored on reception.

3.10. Number of Unicast Address

The Number of Unicast Address (N) field specifies how many unicast IPv4/IPv6 addresses (RP/Sources) are present in this Group Record. This allows to determine the size of the current Group Record

3.11. Multicast Address

The Multicast Address field contains the IPv4/IPv6 multicast group address to which this Group Record pertains.

3.12. Unicast Address [i]

The Unicast Address list is a vector of n IPv4/IPv6 unicast addresses, where n is the value in this record's Number of Unicast Addresses (N) field. The unicast addresses are either Rendez-vous Point or Source addresses depending on the S flag of this record.

If the S bit is set to 0, this error occured in the ASM shared tree towards the RP. In this case Unicast Address [0] is the RP address contained in the failing Join message, while Unicast Address [1] is the RP address known for this group in this router or the NULL address if there is no local RP information for this group. If the S bit is set to 1, this error occured in an SSM or an ASM SPT tree. The list of Unicast Adresses is the list of failing sources for the same group, usually only one in the SSM case.

4. GU Error codes

The error code indicates the reason of the failure incured by a Join message forwarded hop-by-hop towards a source (S), if bit S is set or a Rendez-vous Point (RP), if bit S is cleared. Code values and signification are proposed in the table below. They are only suggestions to future definition by IANA.

S	Value	Name	Description
x	0x1	NO_MCAST_IF	The interface pointing to the nexthop towards the source or the RP is not enabled for multicast.
x	0x2	NO_MCAST_NEIGH	The next hop towards the source or the rendez-vous point is not a multicast capable neighbor.
x	0x3	NO_ROUTE	The router has no route for the source or for the RP of the group.
0	0x4	ERR_RP	The router cannot match a proper Rendez-vous point for the desired joined groupe. For instance, the RP mentioned in the (*,G) part of a Join message may be different from the local RP information for that group.
x	0x5	SCOPED	The group or channel is subject to administrative scoping at this hop.
x	0x6	FILTERED	The group or channel is subject to administrative filtering at this hop.
0	0x7	NO_ASM_ADDR	The group address specified in an ASM Join pertains to a SSM group address for that router.
x	0xF	NOT_FWD	The router is not forwarding this group/source for an unspecified reason.

Table 1 : GU error codes

The x value indicates that the error code may be used for both values of S.

5. Security Consideration

A GU message is an informational message. In this proposal it is only forwarded by routers and does not interact much more with them.

With this proposal, an important number of non-existent trees could

be joined by a set of malicious hosts in the network, possibly triggering a storm of GU messages and exausting the ressources available by trying to signal a malicious error case. This control plane attack is already possible without the GU message [4]. GU messages rate limitation MUST be available in order to minimize the additionnal cost on the network ressources in this kind of attack.

The detailed usage of information carried by the GU message is out of the scope of this draft but Forged GU messages could be sent in non failing trees making the receivers believe that the multicast connection failed at some point in time. To prevent that, only GU messages from known neighbors SHOULD be accepted by routers propagating the unreachability information.

6. Acknowledgments

The authors would like to thanks Stig Venaas and Jerome Durand for the early reviewing and comments on this work.

7. References

7.1. Normative References

[1] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", mar 2006.

[2] Postel, J., "Internet Control Message Protocol", STD 5, RFC 792, September 1981.

[3] Conta, A. and S. Deering, "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", RFC 2463, December 1998.

7.2. Informative References

[4] Savola, P., Lethonen, R., and D. Meyer, "PIM-SM Multicast Routing Security Issues and Enhancements (draft-ietf-mboned-mroutesec-04.txt)", oct 2004.

[5] Bhattacharyya, S., "An Overview of Source-Specific Multicast (SSM)", jul 2003.

[6] Venaas, S., "Source Specific and Any Source multicast ping (http://www.venaas.no/multicast/ssmping/)", apr 2005.

Authors' Addresses

Mickael Hoerdt
Q2S
E-216 Elektro-E
Q2S Center of Excellence
Trondheim NO-7491
NORWAY

Phone: +47 73 55 02 67
Email: hoerdt@q2s.ntnu.no
URI: http://www.q2s.ntnu.no/~hoerdt/

Benoit Hilt
GRTC
IUT de Colmar
BP 50568
Colmar Cedex 68008
FRANCE

Phone: +33 3 89 20 23 64
Email: benoit.hilt@uha.fr

Jean-Jacques Pansiot
LSIIT Universite Louis Pasteur Strasbourg
Pole API, bureau C447
Boulevard Sebastien Brant
Illkirch 67400
FRANCE

Phone: +33 3 90 24 45 63
Email: pansiot@clarinet.u-strasbg.fr
URI: http://clarinet.u-strasbg.fr/~pansiot/

Intellectual Property Statement

Disclaimer of Validity

Copyright Statement

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.