

Debugging Multicast on TEN-155

Jan Novak, DANTE

16.10.2000

1. Scope of the document

This document attempts to provide guidelines for multicast debugging specific to the TEN-155 set-up or more generally to a

purely PIM-SM/MSDP/MBGP based backbone. It does not attempt to describe anything related to the end systems, local area networks and PIM-SM domains, which source the data (basically NRN). The aim is to provide methodology to identify the multicast problems related to the multicast data distribution on TEN-155. The methodology is illustrated using Cisco specific commands. Juniper has always similar equivalents, but with different syntax - the author often also does not know them, so it is up-to the readers to investigate this bit.

For general documentation on multicast debugging please refer to <ftp://ftpeng.cisco.com/ipmulticast.html> or IETF draft " Multicast Debugging Handbook" available also at <http://www.dante.net/mbone/refs/draft-ietf-mboned-mdh-04.txt>

This is not multicast tutorial, person doing debugging must have preliminary knowledge of the protocol principles related to multicast.

2. Knowing the set-up

Both overall and detailed information about the TEN-155 set-up is available at <http://www.dante.net/mbone>. Before being ready for debugging one must know (the rules below are independent on the technology used):

Physical level

- a) all TEN-155 backbone lines and ATM PVCs are multicast enabled (not true as of 16.10.2000 - DE1 in the old PoP and corresponding ATM PVCs are not multicast enabled, but this is going to disappear soon)
- b) most of the NRN accesses are multicast enabled with the following exceptions: ACONET, HUNGARNET, CESNET (dedicated PVC), Poland (dedicated PVC), Renater (dedicated PVC), GARR (tunnel to CH), Israel (tunnel to the UK), Ireland
- c) peering to Abilene is multicast enabled
- d) there is tunnel to UUNET in the US and tunnel to Infonet (crosses the Amsterdam unicast peering) in Europe

Protocol level

To achieve successful multicast forwarding "multicast enabled" means:

- a) PIM enabled on the line/PVC
- b) every internal BGP session must have enabled NLRI=multicast
- c) every internal BGP session (internal here means both peering inside of one confederation AS and peering between two confederation ASs as well) must have parallel MSDP session (parallel means IDENTICAL IP addresses of both peerings) - **this is an absolute requirement especially in the environment of BGP confederations.**
- d) multicast peering to external peers (NRNs, UUNET, Abilene, Infonet) does not have to have parallel BGP and MSDP peering, but one rule must be kept - the MSDP and MBGP session must have end points in the same ASN - e.g. 8933 on the TEN-155 side and "something" on the other side (it is not possible to have MBGP from 8933 to 1275 and MSDP from 8933 to "some-AS-behind 1275" - whatever the AS path in the second case is), even if the "physical" RPF check points to the same interface in both cases.

3. Checking the basic configuration and network status

Before starting some "serious" debugging the following checks must be done (see above in fact):

- a) `sh ip pim interface` and `sh ip pim neighbours` - checks, if PIM is running on the particular interface and if the neighbour is properly seen - **if this condition is not fulfilled no multicast forwarding can occur !!!**
- b) `sh ip mbgp summary` - check if NLRI=multicast peerings are up
- c) `sh ip msdp summary` - check if MSDP sessions are up

The corresponding commands for Junipers are - `sh pim nei`, `sh pim int`, `sh msdp summary`, `sh bgp summary`.

Only after these checks it is reasonable to proceed further in debugging.

4. Common case - data expected but not received (steady data stream sources)

If the basic configuration checked in the point 3 is all right, there is still number of possible reasons for the data forwarding failure. The intra-domain reasons are discussed in the documents provided in paragraph 1. Here are described only few possible inter-domain reasons.

Preliminary knowledge necessary - IP addresses of the source (**S**) of multicast data and the group (**G**) it is sending to. On the TEN-155 network level the expected entry interface must be known.

4.1 Data do not arrive to TEN-155 and TEN-155 sends PIM joins

Check the multicast forwarding state on the TEN-155 neighbour to the external peer, where the data is expected from by issuing commands:

```
sh ip mr S G
```

```
sh ip mr S G count
```

```
sh ip mr S G active
```

If the outgoing interface list of the first command output is populated and all the counters corresponding to the last two commands for the source S are not increasing (needs to be checked at several time intervals) then:

TEN-155 users are sending PIM joins, they know about the source S, but there is no data forwarded to TEN-155 from the neighbour - report to the connected network NOC and ask them to investigate internally.

4.2 Data do arrive to TEN-155, but is not forwarded - first hop

When doing the tests in 4.1 on the TEN-155 router closest to the data originator, look at the output of:

```
sh ip mr S G count
```

carefully - there is RPF check failure counter. If this counter increases, it means data arrive to TEN-155, but is discarded at the entrance.

Issue:

```
sh ip rpf S
```

command. There are two possibilities:

a) the RFP check does not use a route originated by the neighbour (in other words, the neighbour does not announce this source in his MBGP routing table) but uses default route if available or something obscure in the routing tables (something more about this in the paragraph 7.1). Contact the neighbour's NOC to correct the situation - check TEN-155 BGP filters if TEN-155 does not filter the needed prefix out before doing that.

b) RPF check fails due to wrong choice done by TEN-155. Check this case by issuing command:

```
sh ip mbgp S
```

and the output shows correct route from the neighbour, containing the source S and coming from proper/expected interface. For example in the cases of countries with dedicated unicast and multicast interface, unicast routes are used instead of MBGP or backup routes are used etc. Contact the TEN-155 IP NOC with request to correct the situation.

4.3 Data do arrive to TEN-155, but is not forwarded - last hop

After checking the input side (and seeing the data arrive correctly), go to the TEN-155 router closest to the user, who expects the data from (S,G) or alternatively, follow one branch of the distribution tree as seen in the outgoing interface list in the outputs from:

```
sh ip mr S G
```

When there is a NRN interface in the outgoing interface list, follow the procedure of 4.1 again. Look at the (S,G) data counters of "sh ip mr S G count" output and if the RPF

check failure counter is increasing and if "sh ip mbgp S" shows the presence of the correct prefix, there is routing failure inside of TEN-155 - report to the TEN-155 IP NOC.

4.4. Data do not arrive to TEN-155, no PIM joins sent

If the output of:

```
sh ip mr S G
```

issued in the paragraph 4.1 on the TEN-155 router closest to the data source has empty outgoing interface list or more correctly, if this entry is not present in the multicast forwarding table (the output contains information about * G but not about S,G) then it means:

The source S sending to group G is not known to TEN-155 and its users. The reason is MSDP SA distribution failure. Issue the following command to check:

```
sh ip msdp sa | inc S
```

There should not be any SA message for the source S and group G (this requires msdp sa cache being enabled on the TEN-155 routers, currently as of 16.10.2000 on on all of them). These could be some reasons for this:

a) MSDP SA message arrived to TEN-155, but has been discarded on the entrance due to MSDP RPF check failure - follow paragraph 6.2 for MSDP RPF check.

b) MSDP SA message did not arrive to TEN-155 - if point a) has been checked and MSDP RPF check failure does not occur, then the neighbour did not send the SA message - contact the neighbour's IP NOC to investigate.

4.5 Data do arrive to TEN-155 but is distributed partially (to some users only)

Proceed as follows:

Go around all output TEN-155 routers, which users report lack of data and:

a) check the RPF for arriving data by issuing command "sh ip nr S G count" command looking for the RPF check failure counter:

if not OK, TEN-155 internal routing failure, contact TEN-155 IP NOC.

If data RPF OK then:

b) check the presence of MSDP SA message (sh ip ms sa | inc S):

if present contact the neighbour's IP NOC - failure inside of the neighbour's network

If SA not present then:

c) perform MSDP SA RPF check as described in paragraph 6.2:

if it failed contact the TEN-155 IP NOC - TEN-155 internal MSDP failure

If MSDP RPF OK then:

d) inform the site, which originates the SA message that the failure described at <http://www.dante.net/mbone/nop/tale> happens and ask them to upgrade to the appropriate IOS on the router closest to the source S (DR on the LAN). Check for the presence/non-presence of SA on the neighbour to the external peer beforehand also.

5 Debugging SDR - non-steady data stream source

In principle SDR is nothing special - just an ordinary multicast application sending multicast data. The difference is, SDR sends just one packet with the session announcement every 20 and sometimes up to 120 minutes. This is far too much to build and keep

some distribution tree (trees are maintained by periodical control messages triggered by the arriving multicast data traffic) because of too short PIM time-outs. Both PIM (see corresponding RFC) and MSDP have special arrangements to handle such sources. MSDP seeing just one data packet from a source during certain time period, encapsulates the data packet into SA message and forwards it to its neighbours. There are two (known -) problems related to this process:

5.1 SDR cache not populated

There is cisco which prevents the router de-capsulate the SDR packet to itself if "ip sdr listen" command is issued on the same interface, where MSDP session is configured to. Solution - configure it on any other interface, FE, loop back etc.

5.2 SDR packets distributed partially

This network "feature" concerns not only the router itself but all the affected users as well. The reason for this is exactly the same as in the case of paragraph 4.5 case c), MSDP RPF OK (provided all the other reasons have already been excluded - but do so with some other data source than SDR). The problem is, how to prove it with one packet per 20 minutes:

The router must have "sdr cache" enabled, if yes then issue:

```
sh ip sdr "name" detail
```

command. In the detailed session description look for the "Last heard: " field. The MSDP SA cache has 5 minutes timeout, if you have caught SDR announcement with "Last heard" less than 5 minutes and no corresponding SA message appears in the output of:

```
sh ip ms sa | inc S
```

(looking for group 224.2.127.254)

then the problem explained at www.dante.net/mbone/nop/tale has been proven - report to the site originating the SDR announcement and ask them to upgrade to an appropriate IOS.

6 Multicast RPF checks in details

6.1 RPF check for multicast data

PIM stands for Protocol Independent Multicast, which means, it can use whichever routing table available in the router for the RPF check. Some general rules, the explanation of which is out of the scope of this document are explained in DIP 40

In the inter-domain multicast routing there is (as opposed to the intra-domain case) only one relevant RPF check - the multicast data is accepted only on the interface, which is used by the unicast routing protocol to reach the source S of the multicast data. The rule is simple in the case of an external BGP peer under the check and is checked simply by the command:

```
sh ip rpf S
```

the output must show MBGP (this is specific to the TEN-155 configuration - see paragraph 7.1) prefix, which contains (S is /32 IP address while MBGP announces supernets like /16 etc) address S and which is originated by the appropriate AS and has an appropriate AS path - must be known beforehand from the TEN-155 set-up.

The difficulty with this rule comes in the internal BGP environment and especially in the BGP confederations environment:

a) internal BGP - sh ip rpf command shows the MBGP prefix used and the physical interface used. When looking for some TEN-155 external source S, one must learn the MBGP next hop for the prefix returned by the sh ip rpf command (first of all - check in the protocol output field, if the prefix comes from MBGP - if not, it is internal failure or very special case of mixture of MBGP internal routes and BGP external routes - see paragraph 7.1) and then find out, if this next hop is reached by OSPF over the interface stated in the sh ip rpf output. The preliminary knowledge of the peer, from which we expect the data, is required - one must know, that the source S is in the Infonet network for example and that the RPF check should point to the shortest path interface over which Infonet multicast tunnel is reached.

b) BGP confederations - the general rule is the same as above, but the outputs can be very confusing. In the environment, where there are several backup BGP sessions configured, the BGP has many equivalent routes to choose from. First rule is OSPF cost - "confederated peer" looks into the costs, over which he reaches his

peers and chooses the one with least cost path. If there are many with the same cost (and there are usually quite a few in the cases, where IP NOC keeps BGP full mesh even for confederated BGP peers) then the one with lowest BGP ID is chosen. In the case of TEN-155 it is the UK.TEN-155.NET router (as of 16.10.2000). So - do not get confused -) - the routes are "officially" accepted from the UK outer, but the nexthop of the prefix in question can lead to another interface than to the UK !!!! (**This is not the case of MSDP thought -) !!!!!**)

6.2 MSDP RPF check

Similarly to the multicast data, SA messages must be RPF checked to prevent indefinite message looping. Please note, there is no meaning to do RPF check against the route to the source S as the SA message is not originated by the source S. The route to the Rendezvous Point of the corresponding PIM-SM domain, which originated the SA must be used for RPF checks. When checking the MSDP RPF more preliminary information is needed - the IP address of the RP corresponding to the domain in question (in worst case, when setting up new connection, the administrators of the domain must be asked for the IP address of their RP to be capable to do MSDP RPF check when you are not seeing any SA messages yet).

MSDP has implicit (e.g. not explicitly stated in the IETF draft) and confusing RPF rules in different situations:

a) external BGP peer - if doing the MSDP RPF check on the router, to which the nexthop to the RP in question is external BGP peer then the RPF check is simple:

```
sh ip rpf RP
```

(where RP means IP address of the RP)

with the same rules as mentioned in 6.1 for external peers.

b) internal BGP peer - if doing the MSDP RPF check on a router, which has an internal BGP peer as a physical next hop to the external RP, then the earlier mentioned "configuration" RPF rules must be fulfilled:

The IP addresses used to close MSDP and BGP peering must be identical for RPF check to pass:

```
ip msdp peer 1.1.1.1
```

```
ip bgp neighbour 1.1.1.1
```

This rule must be applied to all multicast enabled peers.

c) "confederation-al" peer - the configuration-al rule is the same as in point b). But which peer is chosen in the situation described in paragraph 6.1/b ??? Each MSDP peer must do the following when receiving a SA message:

first - MSDP RPF check

second - if passed, forward SA to all other MSDP peers, but the one, from which the SA was accepted.

These two rules lead to the following - due to the BGP selection rules, described in 6.1 de3.de.ten-155.net for example chooses the BGP routes received from the uk.ten-155.net router to reach fr.ten-155.net. This does not change anything for the data RPF checked - still the shortest path to FR is chosen, but changes a lot for MSDP SA messages. The rule for this case stands:

Accept SA messages from the MSDP peer, which corresponds (due to the necessary identity of IP addresses mentioned above) to the BGP peer, from which we accept the BGP routes!!!

Consequence: Multicast data in DE is accepted from the direct connection to FR but MSDP SA messages from RENATER are accepted from the UK router !!!!

So, do not panic when seeing "MSDP RPF check failed"

in the debug messages on the DE router, when debugging the FR peer!!!

P.S. The DE-FR-UK example has been valid for the "pre-STM-4" migration situation and will certainly change once uk.ten-155.net disappears as new lowest BGP-ID router will have to be elected (and hopefully most of the excessive confed-BGP peerings will disappear).

The best way how to prove MSDP RPF check failure is to run debug "debug ip msdp peer "address"", which lists all SAs forwarded by the peer "address" and their RPF check results on the local router.

7.The TEN-155 configuration specialities

7.1 Multiple routing tables

TEN-155 routers have several routing tables available for RPF checks listed in the order of highest preference (unfortunately there is no output on cisco displaying these preferences for MBGP routes):

- a) few static routes to special destinations
- b) external MBGP routes (SAFI=2)
- c) external unicast BGP routes (SAFI=1)
- d) internal and local MBGP routes (SAFI=2)
- e) OSPF routes
- f) internal and local BGP routes (SAFI=1)

This set-up has following consequences for the RPF checks:

- a) Static routes should never appear in the output of RPF checks
- b) in some very specific cases collision between rules b-c-d mentioned above can happen:

DANTE offices have been advertised as 193.63.211.0/26 from the DE test router. In the UK this route appears as MBGP internal and is superseded by the external unicast BGP route 193.62.0.0/14 announced by JANET and RPF for

DANTE fails in all UK connected countries. In all other countries both routes become internal and rule d) takes over again. It would be probably worthwhile to swap rules c) and d) to make RPF "protocol" consistent.

c) TEN-155 originates 212.1.192.0/21 in both (SAFI=1 and SAFI=2) BGP tables. Due to the rule b) this prevents TEN-155 backbone to originate any multicast data (RPF check always fails as it points to the interfaces used to originate the BGP prefix) even for test or measurement purposes. Two solution are possible:

first - stop to originate the prefix and stop to distribute MBGP default, which is now present in the SAFI=2 table. This would have as consequence RPF check failure for MSDP SA messages of TEN-155 sources outside of TEN-155

second - very careful "tweaking" of OSPF/BGP/MBGP distances to create this preference table:

a) static

b) OSPF

c) all MBGP routes (internal/external)

d) all BGP routes (internal/external)

The rule b) and d) changes the default preferences between OSPF and BGP routes, but it shouldn't harm any unicast routing on TEN-155 as there is no redistribution of OSPF into BGP and vice-versa, e.g. there are never same routes in OSPF and BGP.

7.2 Rendez-vous point configuration

Every TEN-155 cisco router is configured statically as a RP. This is clearly not necessary (now clearly, at the beginning there was simply lack of any consultancy from the outside -)) for the following reason:

TEN-155 does not have any internal data sources, which need the RP functionality of source registering and originating the MSDP SA message (the only exception is multicast trace route packets - the

distribution of these packets could be broken by stopped TEN-155 prefix distribution mentioned above). The network just relies data packets and SA messages from the others.

On the other hand, the core needs to know about at least one RP, otherwise the ciscos (don't know about the Juniper) start to handle all multicast traffic in PIM-DM mode and does not forward any PIM-SM joins - e.g. the whole things just stops(or more precisely, for some unknown reason, the router still sends some joins, but just once per 30 minutes or so)

The currently running configuration is probably not very elegant but does not harm anything and provides 100% redundancy against any RP failure.