#40

TEN-155 Multicast

Jan Novak (DANTE)

Peter Heiligers (DFN)

This paper was first made available by Data Communications as a case study on http://www.data.com in December 1999.

DANTE IN PRINT is a track record of papers and articles published by, or on behalf of DANTE. HTML and Postscript versions are available from: http://www.dante.net/pubs/dip

For more information about DANTE or *DANTE IN PRINT* please contact:

DANTE
Francis House
112 Hills Road
Cambridge CB2 1PQ
United Kingdom

Tel: +44 1223 302992
Fax: +44 1223 303005
E-mail: dante@dante.org.uk

TEN-155 Multicast

Jan Novak (DANTE)

Peter Heiligers (DFN)

Abstract

This paper addresses the deployment of native multicast on a pan-European scale on the TEN-155 network. It also outlines various migration issues from the tunnelled MBONE infrastructure which was previously used. IP Multicast routing/forwarding in the backbone of TEN-155, the pan-European research network, and towards the European National Research Networks (NRNs) connected to it, make use of the PIM-SM (Protocol Independent Multicast - Sparse Mode), MBGP (Multiprotocol Border Gateway Protocol) and MSDP (Multicast Source Discovery Protocol) protocols. The implementation of these protocols is a great step forward compared to the MBONE based on DVMRP, but leads to other problems.

KEYWORDS: TEN-155, Multicast, MSDP, MBGP, PIM-SM, DVMRP, ATM, Full Mesh.

1. DANTE, QUANTUM and TEN-155

The Quantum[1] project foresees the exploration and implementation of providing Quality of Service across a pan-European network of very high speed dedicated to the European National Research Networks to support co-operative research. The Quantum project also calls for experimentation of new IP and ATM technology using a Wide Area and international test network (Quantum Test Programme - QTP[2]). The goal of the QTP is to validate emerging technologies with the aim to deploy operational services. The activity on native multicast described in this paper stems from this QTP activity.

TEN-155[3] is the operational network built as a result of the Quantumproject. A group of 16 national Research Networks and one regional network, co-ordinated by DANTE [4], are responsible for the Quantum Project which is co-funded under a joint initiative by DG XIII (Telematics Applications, Esprit and ACTS) of the European Commission.

DANTE is a non-profit company set up in 1993 by European National Research Network organisations. DANTE plans, builds and manages advanced networking services for the European research community.

2. Introduction

The developments and improvements madee by the MBONE service introduced on the TEN-34 network [5] are described in "DANTE in Print No.36"[6]. The TEN-155 MBONE infrastructure was configured during the migration from the TEN-34 to the TEN-155 network with DVMRP tunnels that matched the physical infrastructure.

The migration plan to PIM-SM / MBGP [7] was based on Cisco's implementation of PIM-SM[8], MBGP[9], and MSDP[10]. The advantages of this protocol stack compared to DVMRP tunnelling can be summarised as follows:

Jan Novak is a Network Engineer at DANTE.Peter Heiligers is also a Network Engineer at DFN, the German national research network.
Their email addresses are Jan.Novak@dante.org.uk and heiligers@noc.dfn.de

1) routing scalability - DVMRP is not suitable for inter domain routing;

2) same policy routing functionality as the current unicast routing in the Internet (AS-path / other filtering, broad range of attributes to influence path decisions);

3) no interference between unicast and multicast routing tables - there is the possibility to create a separate routing table which is used only for multicast RPF (Reverse Path Forwarding) checks; and

4) interconnection of Sparse Mode multicast domains using MSDP which in turn removes the typical non-pruning difficulties of the MBONE based on DVMRP.

The first phase of the migration plan was the testing of the relatively new MSDP / MBGP software. The purpose of MBGP is to provide policy routing for multicast, in the same way that BGP provides this functionality for unicast. The purpose of MSDP is to distribute information about multicast sources between different SM domains.

3. First test

DANTE installed a test router (Cisco 7507) in the TEN-155 Frankfurt PoP. This test router (DE2.ten-155.net) was needed for several reasons:

1) To test new software, a unicast peering was configured with the local production router to verify that MBGP NLRI=unicast (Network Layer Reachability Information) and NLRI=multicast routing table do not interfere. There are several important rules for path selection on a router running both unicast and multicast:
   a. an unicast path lookup never takes into account the NLRI multicast table (including the DVMRP routing table)
   b. a PIM RPF check can use any routing table created by any routing protocol running in the router. The way to point to the appropriate table is to use a Cisco internal parameter called distance. This is arouting protocol preference parameter which builds a protocol hierarchy for the route lookup (RPF check). It can be configured separately for every routing protocol running on the router.
   c. the longest match rule (the most important rule for unicast path selection) is no longer valid for RPF lookups - e.g. when 212.1.192 / 19 is available in the routing table of one routing protocol with a distance 15 and 212.1.192 / 21 is available in the routing table of another routing protocol with a distance 20, the RPF check selects the 212.1.192 / 19 prefix.

2) A first functionality test of MSDP was needed as well as interoperability tests between a DVMRP based cloud and a PIM-SM based cloud. The main purpose of MSDP is to interconnect PIM-SM domains operated by NRNs with the TEN-155 PIM-SM domain to enable any-to-any multicast data exchange.

3) To provide connectivity between the DVMRP cloud and the NRNs which use PIM-SM / MBGP / MSDP - for technical reasons DANTE did not want to support anything else than PIM-SM / MSDP on production TEN-155 routers.

The DVMRP - PIM-SM interoperability tests have been performed in co-operation with the German National Research Net-

work organisation DFN. DANTE and DFN performed various data transfer tests (using MGEN and, later on, normal audio / video session with MBONE tools) with the set-up depicted in Figure 1

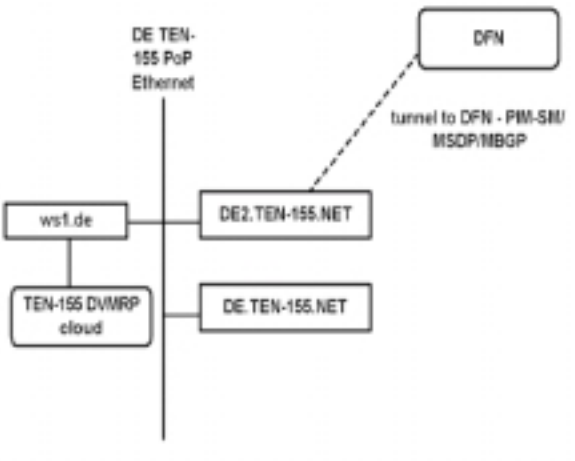

Figure 1: Initial Test Setup

Both DANTE and DFN generated data from both domains and tested the creation of forwarding states, their stability, expiration and finally data delivery. During this period unicast routing and CPU utilisation were monitored. This test set-up provided an excellent opportunity to test all the features mentioned above due to the "broad" range of routing protocols run by the DE2 router - MBGP to DFN, DVMRP to TEN-155 MBONE, BGP and OSPF to the DE TEN-155 unicast production router, static unicast and multicast routes.

4. The SE-NL-DE triangle

After two weeks of testing DANTE upgraded IOS and enabled IP multicast routing on the TEN-155 production routers in NL and SE. The respective NRNs, SURFnet and NORDUnet were connected to these routers. Both SE and NL TEN-155 routers were connected to the DE2 test router via separate dedicated ATM PVCs for multicast. The DE2 test router provided interconnection to the TEN-155 DVMRP cloud, DFN and US MBONE. The Oslo IETF in July 1999 was transmitted over this infrastructure and proved to be an excellent opportunity to stress test the set-up. All European countries in both clouds, DVMRP and MBGP, could accept high-speed multimedia data streams in very high quality. The scheme of the network topology at this stage is depicted in Figure 2.

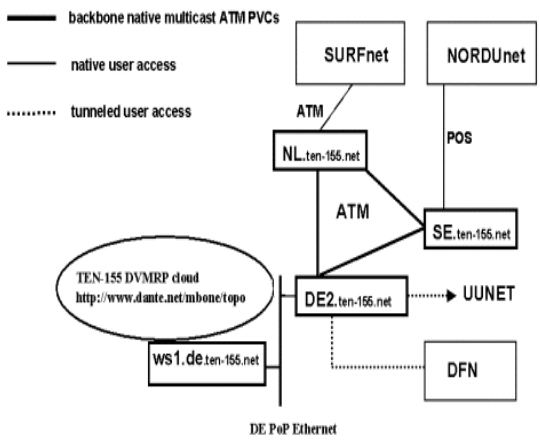

Figure 2: Network topology including NL and SE TEN-155 production routers

5. Current status

Currently multicast is enabled on the FR, NL, SE and UK TEN-155 production routers. All these routers are in the same BGP Autonomous System. The logical scheme is shown in Figure 3. Every router is statically configured as a Rendezvous Point (RP), creating a PIM-SM domain. The DE2 test router is still in use, but a new production router is planned to be used for both unicast and multicast routing. The reason for this set-up is the distribution of MSDP connections (TEN-155 has 19 NRNs connected) over several physical devices. A detailed network topology is available at http://www.dante.net/MBONE/topo/mcast155.html . The connection to the US was migrated from DVMRP to MBGP/MSDP. This connection, however, runs over a tunnel. DANTE plans to migrate it to a native connection over a dedicated ATM PVC. The rest of the DVMRP cloud consists of five countries (UK, IT, SI, LU, CH-CERN), which are planned to be migrated before the end of 1999. The set-up shows the advantages of having separate routing tables for unicast

and multicast: DANTE can offer a flexible multicast service to the NRNs - they can also use their production unicast router for multicast or have a separate multicast router connected either over an ATM PVC dedicated to multicast or over a tunnel. Due to the problems described in the next paragraphs, DANTE does not plan to support DVMRP from January 2000 at all.

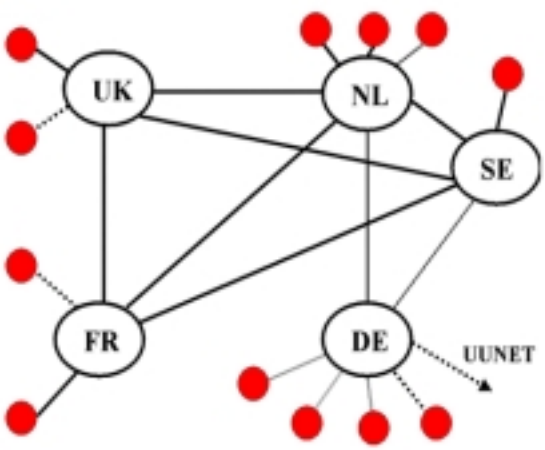

Figure 3: Scheme of the current network topology

The single lines are ATM PVCs/lines used for both unicast and multicast. Double lines are ATM PVCs dedicated to multicast and dashed lines are tunnels.

In the UK node, RCCN (PT) and MACHBA/ILAN (IL) are connected. On the Netherlands router there are connections to SURFnet (NL), GRNET (GR), BELNET (BE) and recently Unisource was connected also here. In Sweden all Nordic countries are connected via NORDUnet (SE, NO, FI, DK, IS). In France, RedIRIS (ES) and RENATER (FR) are connected. In Germany, DFN (DE), CESNET (CZ), SWITCH (CH), POL34 (PL) and the US multicast connections are configured for the moment.

The whole set-up was again stress tested during the Telecom99 conference in Geneva, September 1999, when several streams of up-to 1.5 Mbit/s of multicast data were transported to all participants.

6. The MBGP paradox

The purpose of MBGP is to enable inter domain multicast routing on the same physical infrastructure used for IP unicast traffic but also to allow one creation of non-congruent unicast and multicast topologies (multicast can be enabled on some peerings, but not necessarily on all of them). This is perfectly valid in an External BGP (EBGP) environment but has serious limitations for Internal BGP (IBGP) peerings. The BGP TCP connection always handles both NLRI together and a full mesh of IBGP must be kept inside an AS in order to maintain full any-to-any connectivity. There seem to be only two possibilities with respect to the multicast configuration of such an AS:

1) The whole network is fully enabled for both multicast and unicast losing the possibility of non-congruent multicast and unicast topologies inside an AS

2) Multicast is enabled only on some lines (either physical lines or ATM/Frame Relay PVCs), for example in the cases where mapping of multicast is required onto the connections, matching the physical topology in an ATM/Frame Relay network (TEN-155 case, see later). In this case, when a multicast enabled line fails, the RPF check can point to non-multicast enabled interfaces with multicast data being discarded as consequence.

DANTE, in co-operation with the TEN-155 IP NOC, carried out an experiment, to verify if it was possible to avoid this problem. In the topology in Figure 4, a DVMRP routing table was created for BGP next hops to eBGP peers.

The main objective of the experiment was to avoid multicast traffic on the ATM PVC between UK and SE, because this does not match the physical topology. The idea for

achieving this objective was the following: if the next-hop lookup of a BGP process was done independently for NLRI=unicast and NLRI=multicast, the unicast part could choose a normal OSPF routing table to reach the next hop. The unicast lookup does not see multicast routing table and uses the best OSPF path. On the other hand, the multicast part of the BGP process could use whatever routing table was available and choose one using Cisco distances (similarly to PIM RPF lookup) - e.g. in the particular case of this experiment using the DVMRP routing table created with default distance 0. Unfortunately, the RPF lookup points also to non-multicast enabled, e.g. in this case it points simply to the direct SE-UK PVC. This is a general design issue as the RPF lookup does not have any other possibility to choose another interface. Recently Cisco discussed with representatives of BELNET the possibility to configure two BGP sessions between two routers and set different next hops for unicast and multicast routes. This would again allow to create another routing table for multicast next hops and map the multicast data distribution back to the physical topology.

Print 37 [11]). This set-up is highly optimal for unicast routing, enabling shortest IP level path between all participants. The ideal set-up for multicast would be the mapping of the multicast data distribution tree only to the ATM PVCs which match the physical infrastructure (e.g. direct PVCs which span only one physical line, for example UK-NL and NL-SE in figure 4). Unfortunately, this is not possible as explained in the previous paragraph. For example, in the case of the experiment described above, the ATM PVC between SE and UK must be multicast enabled to allow multicast data exchange between these two countries. DANTE decided not to enable the whole full mesh for multicast at the moment, but to enable only the nodes (and corresponding iBGP sessions to keep full connectivity between all multicast enabled core routers) where two and more participants are connected. In other places direct eBGP sessions are configured via dedicated multicast ATM PVCs to bypass the nearest TEN-155 node. The disadvantage of this set-up is reduced redundancy, as RPF checks can still point to non-multicast enabled PVCs. The analysis made recently by DANTE shows that in the

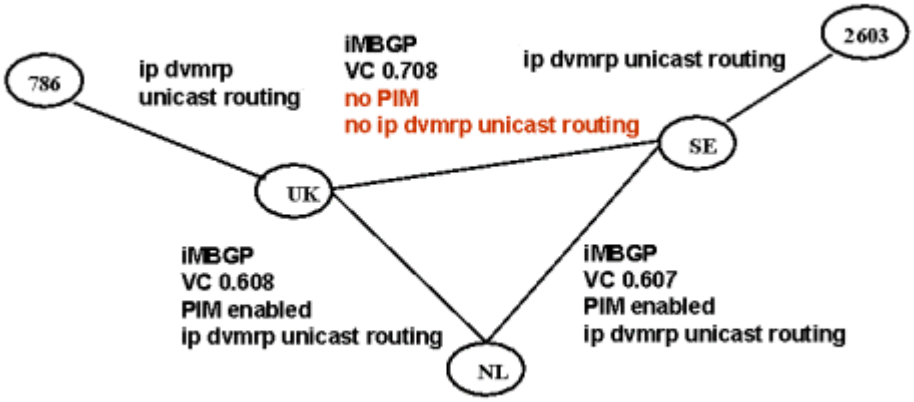

Figure 4: Scheme of the BGP next hop experiment

7. The TEN-155 Full Mesh

TEN-155 is an ATM based network, fully meshed on the ATM level between all core routers (described in detail in DANTE in

particular physical topology of TEN-155 (available at www.dante.net/operations/ten-155/topology.gif) the unnecessary multiplication of multicast data on a full

mesh of ATM PVCs is not as bad as it seemed to be initially. The basic idea of this comparison is shown in figures 5 and 6.

Let us assume there is one sender in one of the nodes (SE in Figure 5; the numbers show the count of the number of the TEN-155 circuits used to deliver multicast data to all nodes) and receivers are in all the other nodes. In the ideal case, when multicast is mapped to the physical topology, every multicast packet appears on 7 physical lines to reach all participants - in other words when transmitting a packet with multicast, the network multiplies the packet seven times when delivering multicast data to all participants. This number is the same for any of the nodes, when sending a multicast packet to all other nodes, in the TEN-155 topology:

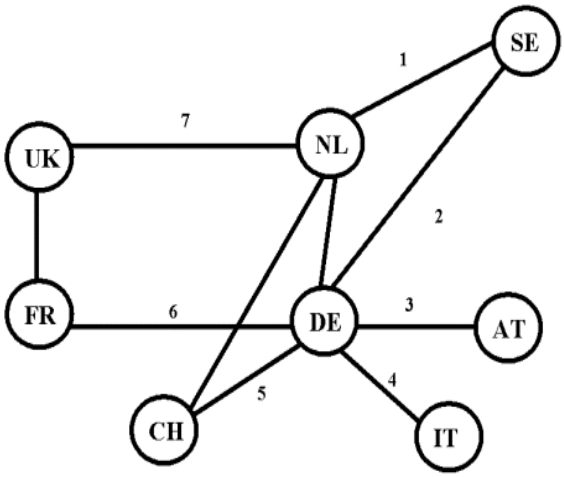

Figure 5: Scheme of TEN-155 physical topology

Now, let us imagine a full mesh of ATM PVCs between all 8 nodes, the number of appearances of a packet on the physical lines changes. The topology as seen on IP level (while the physical topology remains the same as in the previous case) by the routers is:

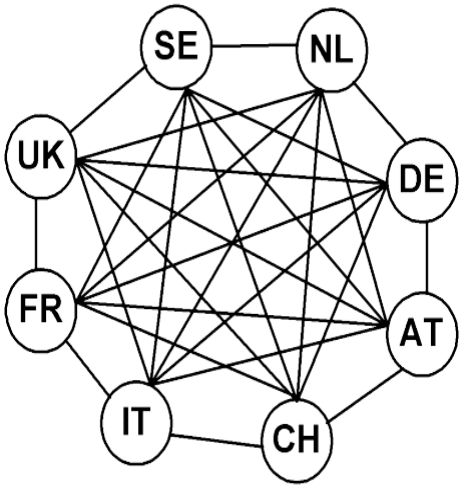

Figure 6: TEN-155 topology as seen on IP level

Assuming again that packets are delivered to all participants when sending from one country to all the others, the number of appearances of a packet on all physical circuits of TEN-155 is as follows when sending from:

AT - 14 times; CH - 12 times; DE - 8 times; FR - 12 times; IT - 14 times; NL - 10 times; SE - 12 times; UK - 14 times

The increase is caused by the fact that most of the ATM PVCs span more than one physical line. This leads to an average value of 12. The ratio of packet appearances on physical lines when comparing both cases above becomes 7 : 12 which is much more favourable than the first view on the problem, which was: "There is seven times more multicast packets than necessary when sending from AT or IT to all other participants."

8. TEN-155 Ghosts

It has been noticed that when a workstation in the DVMRP cloud joins and subsequently leaves a multicast group, the forwarding state remains in the multicast-forwarding table at a router. DANTE named this phenomenon "ghost", because these entries persisted in the network several weeks. The forwarding state is regenerated from the neighbours so

even a router reload, IOS upgrade or whatever else does not help to remove the entry from the table. This does not affect the multicast functionality, but creates a potential problem considering the number of end systems using RTCP based applications around the world. In addition it has similar effects to that of the non-pruners on DVMRP. Cisco debugged the problem quite extensively and the explanation in summary is as follows: When the WS became active, MSDP at DE2 (which is responsible to generate MSDP messages on behalf of the DVMRP cloud) announced a new data source to the neighbours. When the WS left the group DE2 erased the state, but then joins arrived from the neighbours, still having the forwarding state. The problem was that DE2 re-announced this source in MSDP SA message again. Since the IOS version 12.0.6S MSDP originates SA messages only for sources, which are registered (a specified PIM procedure) at the RP. To fully solve the problem an upgrade to proper IOS version is still necessary on all nodes using MSDP.

The other effect related to forwarding states, which are kept unnecessarily in DM domains, is continuos forwarding of data even over MSDP / PIM-SM connections to TEN-155. It seemed to be directly related to the problem discussed above, but the forwarding continues to DFN for example even after upgrade of their MSDP nodes (DFN is mixture of DM and SM domains) and is observed also on other connections (RENATER, CESNET, POL-34 for example). This issue is demonstrated on the graphs below and needs further investigation.

The first one (Figure 7) is purely a PIM-SM domain (GRNET), monitored on the GRNET router. The input from TEN-155 (dark areas) behaves correctly - there is really traffic only when there are receivers participating in some group. The lines above dark areas show traffic sent from GRNET to TEN-155.

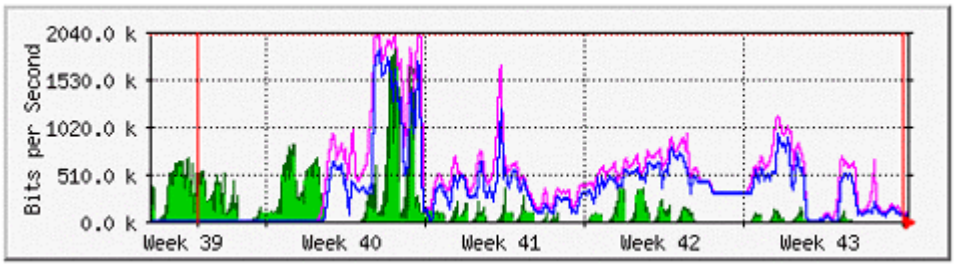

Figure 7: Multicast-only traffic on multicast-enabled access of GRNET (Greece) to TEN-155

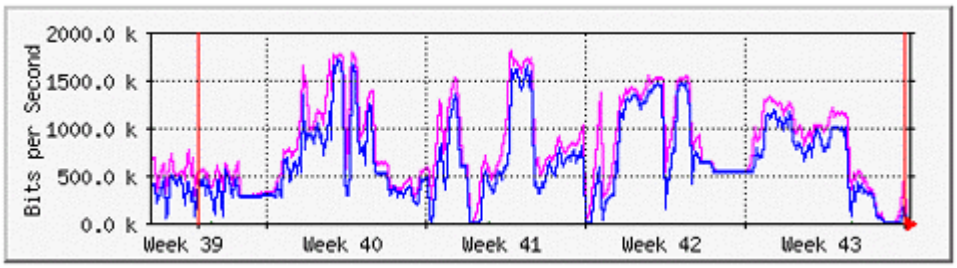

Figure 8: Multicast traffic on dedicated multicast PVC to CESNET (Czech Republic)

The second graph (Figure 8, CESNET) shows the behaviour of a non-pruning PIM-DM domain behind a MSDP / PIM-SM connection to TEN-155. The graph shows multicast traffic from TEN-155 (e.g. attracted by CESNET) to CESNET. During this period (weeks 39 to 43 of 1999) there was no traffic from CESNET to TEN-155. The figure demonstrates that Sparse Mode behaviour occurs only after router reboots or connection outages, until someone joins a group. After that, traffic grows in steps, as other users join other groups, but almost never decreases until the next reboot / outage (week 41, beginning of the week 42), long enough for the whole multicast forwarding table to time out.

9. The DVMRP Neighbour

The DE2 test router has a direct DVMRP neighbour, connected in native mode over a local Ethernet. There are several problems related to this, which prevent DANTE from offering DVMRP connections to TEN-155 routers:

1) The DVMRP connections are typically non-pruning. All the TEN-155 routers are configured as being PIM-SM RPs and run MSDP e.g. they know about all the sources and traffic in the multicast network. Regardless of the quality of PIM-SM to DVMRP interoperability implementation, this always causes all traffic to be forwarded to the DVMRP cloud. This behaviour is documented on the traffic graph below taken from the live TEN-155 network. There are problems with non-pruning PIM-DM domains connected over standard TEN-155 PIM-SM / MSDP connection, but the traffic rates (steady "stream" part of the rate, which does not change even during off-peak hours) are never as bad as on this DVMRP connection, as shown by figure 9.

2) When Cisco IOS discovers a DVMRP neighbour in native mode (this is not valid for tunnelled DVMRP connections) , it automatically adds the corresponding interface to the list of outgoing interfaces in all multicast forwarding table entries. This enables interoperability, but has two serious side effects:

a) excessive traffic load, even if the DVMRP domain behaves and prunes correctly
b) excessive PIM process CPU load - the PIM process takes up to 10% of CPU capacity, because of the "virtual" joins originated on behalf of the DVMRP neighbour. This effect is probably even multiplied by the ghost entries described above.

3) Cisco IOS sets the C flag in all multicast forwarding table entries which means that the receivers are directly connected. This is done again for the sake of easy interoperability and enables Cisco to use IGMP to register receivers from the DVMRP domain and to forward data to the DVMRP domain. The side effect of

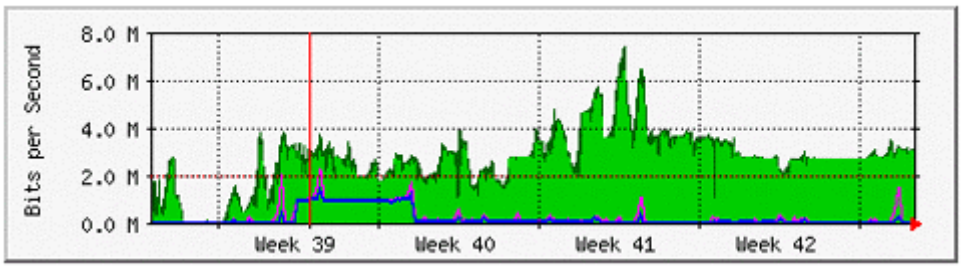

Figure 9: Traffic from DE2 TEN-155 router towards the DVMRP domain

this is incorrect responses to Unix mtraces - Cisco router fails to respond, because it believes all sources are directly connected. Therefore most mtrace forwarding attempts, which are essential for debugging purposes, end with a Wrong interface error message as shown in the output bellow:

```
Mtrace from 150.203.20.40 to
193.63.211.1 via group 224.2.140.220
  Querying full reverse path... * switching
to hop-by-hop:
 0 sun.dante.org.uk (193.63.211.1)
 -1 * * atm.ws1.de.ten-155.net
(212.1.193.233) DVMRP threshold 0 83 s
  Wrong interface [default]
```

This does not influence the forwarding itself, but complicates debugging and causes a lot of questions from the NRN side about the proper routing.

10. The TEN-155 Multicast Service

The result of all the work done throughout 1999 will be the introduction of a production multicast service on TEN-155 in January 2000. This means a detailed service specification will be included in the contracts between DANTE and NRNs, operational procedures with TEN-155 IP NOC will be defined and network monitoring will be set-up. In order to reach these objectives, DANTE plans to upgrade all TEN-155 routers to 12.0 IOS, which enables the gathering of per ATM PVC statistics and use of IPmrouteMIB to separate multicast traffic from unicast in the backbone. The outlines of the service specification are available at http://www.dante.net/MBONE/mcast99/mphase2.html. The European part of the multicast service is currently technically operational and stable, apart from the problems outlined previously. There are still operational issues to be solved regarding the US connectivity, which is not as stable as required.

11. Future Developments

DANTE plans to set-up multicast at it's New York PoP also and to terminate the UUNET multicast connection there, preferably using native mode over a separate ATM PVC. This New York router will be used also for multicast connection to the US research networks. Currently an unicast connection to Abilene is in place and DANTE has started to arrange a multicast connection as well. In the near future DANTE intends to enable TTL and administrative scoping on all TEN-155 interfaces to external peers.

12. Conclusion

Multicast is one of the rare technologies where the designers paid an amazing attention to the highest efficiency possible. Multicast data are not broadcasted even at the last hop, e.g. local area networks, where a lot of cheap bandwidth could be anticipated.
The backside of this is an increased network complexity to fulfil these goals. The main purpose of this work is to achieve the transparency of this complexity for the end users on a network of an international scale and to enable the many-to-many communication in an effective way. This has already been done nationally in some connected NRNs. The countries like France, Germany and United Kingdom have relatively broad range of users of their national multicast service for operational meetings, teaching purposes or conferences on the national level. Other countries, like Belgium, Netherlands and NORDunet connected countries seem to concentrate more on the content providing side of the multicast network usage, e.g. multicast of radio and TV transmissions and of some important international events.

The challenge, which still remains to be faced, is to traverse several management domains and make it fully and easily operational on an international level. With respect to this, Cisco provides stable and fully functional software to reach these goals.

Acknowledgements

DANTE and the authors would like to thank Roberto Sabatino (DANTE), Patrick De Muynk (BELNET), Havard Eidnes (NORDunet), Bernard Tuy (Renater), Dimitrios Kalogeras (GRnet), Duncan Rogerson (TEN-155 IP NOC) and Steffen Baur (DFN) for their interest in this work and for many comments and ideas provided. DANTE would also like to thank all NRNs engineers, who dedicated part of their time to set-up new connections and debug them.

The authors especially acknowledge the enthusiasm and excellent work of Ijsbrand Wijnands (Cisco) in debugging problems related to this project. Cisco also supported this project in the startup phase through a router on loan.

References

[1] Quality Network Technology for User-Oriented Multi-media,
http://www.dante.net/quantum

[2] Quantum Test Programme,
http://www.dante.net/tf-tant

[3] Trans European Network, 155 Mbit/s,
http://www.dante.net/ten-155

[4] Delivery of Advanced Networking Technology to Europe Ltd,
http://www.dante.net

[5] Trans European Network - 34 Mbit/s,
http://www.dante.net/ten-34

[6] "DANTE in Print 36",
http://www.dante.net/pubs/dip/36/36.html

[7] Initial TEN-155 migration plan,
http://www.dante.net/MBONE/mcast99/migration.html

[8] Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification - RFC2362

[9] Multiprotocol Extensions for BGP-4 - RFC2283

[10] Multicast Source Discovery Protocol (MSDP) -

[11] DANTE in Print 37,
http://www.dante.net/pubs/dip/37/37.html