

DEPLOYMENT OF NATIVE IP MULTICAST ROUTING SERVICES ON THE ITALIAN ACADEMIC AND RESEARCH NETWORK

TIZIANA FERRARI

*Italian National Institute for Nuclear Physics (INFN-CNAF)
v.le Bertini 6/2, I-40127 Bologna ITALY
E-mail: Tiziana.Ferrari@cnafl.infn.it*

ANTONIO PINIZZOTTO, MARCO SOMMANI

*Institute for Telematic Application – Italian National Research Council (IAT-CNR)
via Alfieri 1, I-56010 Ghezzano – Pisa ITALY
E-mail: Antonio.Pinizzotto@iat.cnr.it, Marco.Sommani@iat.cnr.it*

DAMIR POBRIC

*Consorzio Pisa Ricerche (CPR)
p.za D'Ancona 1, I-56127 Pisa ITALY
E-mail: Damir.Pobric@iat.cnr.it*

In 1993 the GARR (Italian Academic and Research Network) was connected to MBONE (worldwide IP multicast enabled network backbone). Since then, and especially in the last few years, this network has been used to test and develop new IP multicast routing protocols and applications. However it was a tunnel-based solution, not suitable for taking advantage of the new potential of IP multicast nor for use by a large community.

This paper describes the work done for a migration to a native IP multicast routing deployment, highlighting the solutions adopted when dealing with the implementation problems and the complex wide area network management. Aim of this work is also to ensure a native IP multicast connection with the other Research Networks and the rest of Internet.

1 Introduction

IP multicast services were introduced on the Internet in 1991, when a set of IP tunnels were set off among a group of hosts (m routers) enabled to act as multicast routers. This structure, known as Mbone, allowed the use of multicast applications by computers placed on the subnets adjacent to an m router. In 1993 the first m router of GARR (Italian Academic and Research Network) was connected to MBone. Since then the Italian Mbone structure has grown constantly, becoming a multicast network composed of 40 m routers. In recent years Mbone has been used to test and develop IP multicast routing protocols, transport protocols and application programs [1].

The original MBone structure allowed the testing of multimedia services in a small Internet community but, due to various limitations, it could not be extended to a larger community of users and Internet on the whole. Due to the growing interest in multicast technology, in commercial as academic areas as well, many research networks and some Internet Service Providers have begun to enable multicast on their devices, using technical solutions which differ from those of the original Mbone [6].

It is essential for the GARR network to ensure that their customers will receive the same multicast services as those already provided in many Internet networks. In order to improve performances, multicast and unicast routing must be done on the same production routers of the GARR network [3].

2 Mbone limits and solutions

Past years' experience has confirmed the advantages and potentials of IP multicast, while revealing the limits of the original Mbone model.

The main limits were as follows:

1. At the beginning, the multicast routing process was implemented on workstations (mrouters), generally equipped with a single network interface. On this same physical interface numerous tunnels were defined causing a high traffic load on the subnet and excessive bandwidth utilization.
2. In most cases a change in network topology requires multicast tunnels to be reconfigured.
3. The multicast routing protocol initially used on Mbone (DVMRP, [2]) creates distribution trees by using the flood and prune technique. It implies multicast packets to be periodically delivered to every corner of the network including the branches with no listeners.
4. Multicast routers use Reverse Path Forwarding (RPF) check to guarantee that the distribution tree is loop free. A router will forward a multicast packet only if it comes from the upstream neighbor, the one leading back to the source. This check is based on a forwarding table. DVMRP builds this forwarding table using a distance vector unicast protocol similar to RIP, unsuitable for a wide area multicast network.

In order to eliminate the disadvantages described in points 1, 2 and 4, a native IP multicast must be used, that is to run production IP multicast on the common GARR-B infrastructure with other services.

At present nearly all the manufacturers have introduced routing multicast functionality on their routers, using one of the routing protocols in the following list:

- DVMRP
- PIM, which is independent of unicast routing protocol and does not build its own table. It supports dense (or push) and sparse (or pull) modes:
 - a) PIM-DM uses the "flood and prune" technique, like DVMRP;
 - b) PIM-SM [11] is designed to support sparse multicast groups. It uses an explicit join model and delivers the data using both source and shared distribution trees. Shared trees require, for each multicast group, a router with the function of "rendezvous point" (RP), the point where the group's receivers and senders meet.
- MOSPF, an extension of the OSPF protocol in which a new type of Link State Advertisement is introduced and used by a router to advise others of the presence of listeners on their adjacent subnets and distribution tree creation. MOSPF is protocol (OSPF) dependent and, by definition, is not suitable as a multicast protocol for the whole Internet.

PIM-SM is the only protocol, of those listed above, which does not use the “flood and prune” technique and is independent of the unicast protocol chosen. For this reason it is an optimal protocol for enabling native multicast routing.

3 Implementation

3.1 PIM-SM routing protocol

PIM-SM is a multicast routing protocol designed to efficiently create the multicast distribution tree for groups scattered over the Internet. The PIM-SM features are:

- PIM-SM is receiver-initiated, that is the distribution tree is created as a consequence of the explicit join requests sent from the PIM designated routers (DR) having active members of a multicast group on their adjacent networks.
- For each multicast group there is a router, Rendezvous Point (RP), where the receivers and senders “learn” about each other.

Receivers join RP – when a DR detects the presence of members of a group on an adjacent network it sends a join request to the RP router. Each upstream router intercepts the join request, along its path towards the RP, thus creating a shared distribution tree (RPT).

Senders register to RP - A DR that receives a multicast packet, originated on adjacent networks, must send it to the RP for distribution down the shared tree.

- If a router does not join the distribution tree it will not receive any multicast traffic for that group.
- Beside shared distribution trees, PIM-SM provides for shortest-path tree (SPT) also. Switchover to SPT is performed if shared tree is not an optimal path between the source and the receiver.
- For different traffic sources of the same group different types of distribution tree can be chosen.

3.2 The Backbone

The first step in the service implementation is the configuration of IP multicast on the backbone routers. PIM-SM protocol must be configured on all the routers’ interfaces. PIM-SM uses the unicast routing tables, already present on the router, for the following functions: multicast packet forwarding, RPF check and the transmission of control messages. The unicast routing model of the GARR network is simple and designed according to the existing backbone network topology. It is hierarchically split in two parts and composed of two types of routers: transport router (TR) and concentration router (CR).

National and international ISPs and the concentration routers are connected to the transport routers, which are connected in a complete ATM PVC mesh. On CR routers there are only user network access links through the user routers (UR). In terms of routing the entire backbone is a single Autonomous System. All the routing information, except for the inner networks and backbone, is carried in BGP protocol. TR BGP tables contain all the

reachable destinations, both internal (user network) and external. CR routers know only the user networks connected to them. OSPF protocol is configured on TR and CR routers; its only aim is to determine the “next-hop” router address used by BGP. Paths among backbone routers are multiple but the BGP protocol always chooses the best one. Therefore, there is no risk of multiple parallel paths; this would be very difficult to manage by multicast routing because of the RPF check, and the multicast packet forwarding needs no particular unicast routing functionality.

The user networks are “single homed” and routing is configured in two ways: static and by BGP protocol. For both, multicast and unicast topologies coincide and generally multicast does not require non-standard unicast routing solutions.

3.3 Rendezvous Point RP and peering MSDP

PIM-SM protocol specification assumes a single active RP router for each multicast group. For this reason, in a wide area network, the location of the RP router can cause several problems such as: traffic concentration, lack of backup RP and dependence on distant RP. These problems can be resolved using logical RP [10]. This technique, based on MSDP protocol [9], allows an arbitrary number of RP thus adapting to specific network topology. The mechanism works as follows: all the RP routers, for a group or a set of groups, are configured with a unique IP address (normally the logical loopback interface is used). Group joins or the transmission of register packets, from designated routers, will reach the topologically closest RPs. In this way multiple distribution trees are created. To connect these trees or domains each RP must establish a MSDP peering with other RPs. By means of MSDP messages, RPs can discover the active sources on other tree/RP and, if they have any group members for that group, can generate a join message toward the source. In this way a delivery of multicast packets from a certain source is guaranteed to all members of a specific group.

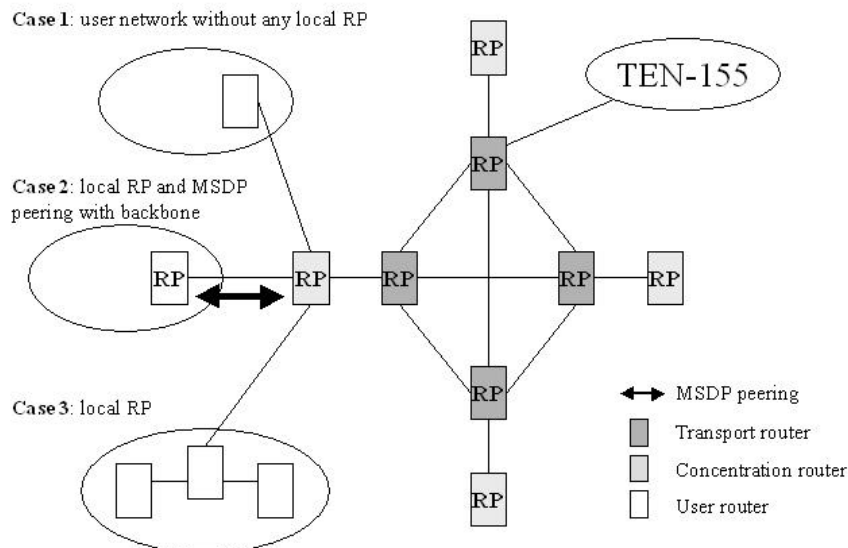


Figure 1. RP overall location and user type link.

All the CR routers are configured as RP (internal RP). Vice versa, only some of the TR routers are configured as RP (boundary RP): the ones connected to ISPs with which a multicast traffic exchange is desired (multicast peering).

Figure 1 shows the overall RP deployment scheme and the three different link types for the user routers.

3.3.1 Internal RP

CR routers act as RP for the user networks; given their vicinity to the user networks they are the most suitable point. By default these RP serve all the multicast groups (see multicast boundaries). However, it is advisable to configure an internal user network RP for scoped multicast addresses (Administratively Scoped Multicast Space [12]) destined for local transmissions. In this way “local” multicast traffic does not load the access link nor cause any overhead in the creation of CR router states.

3.3.2 Border RP

At present, PIM-SM protocol, in conjunction with MSDP, is also used for inter-domain routing. To facilitate the multicast peering with the national and international service providers, the border routers (TR) must also serve as RPs. As well as the MSDP peering with all the internal RP, the border RP must establish an MSDP peering with the corresponding routers in other autonomous systems. The border RPs must be configured in such a way as to announce only the local multicast sources (GARR); on the contrary some Internet service providers could use GARR as a transit network. In certain circumstances, for instance for experimental purposes or in the case that an ISP uses GARR as a multicast traffic feed, this rule may be relaxed.

3.3.3 RP configuration

All designated routers must know the IP address of RP for a particular multicast group. RP configuration can be static or dynamic. The static method is discouraged, as any configuration errors could cause multicast routing and connectivity problems.

PIM protocol provides an automatic mechanism to announce information about group to RP mapping (Auto-RP). The advantages of this method are:

- guaranteed RP information consistency and correctness;
- independence of possible changes of RPs or their IP addresses.

3.3.4 User RP

The users’ network manager can decide to configure the UR routers as RPs. These RP can serve either only the scoped groups or all multicast groups. If the RP must serve only the scoped groups it is particularly important to correctly configure the range of scoped addresses. An error could produce the overlapping of RPs for some global multicast address ranges and they would be then cut off from the rest of the GARR net. If, instead, the RP must serve all the multicast addresses for its domain, it must establish an MSDP peering with a backbone RP router.

3.4 Administrative boundaries

In order to safely use the same multicast address in different administrative domains the propagation of packets for these groups must be limited. The mechanism consists of the creation of administrative boundaries. On the boundary, the passage of packets addressed to a scoped range is not allowed. The boundaries (for the entire range 239.0.0.0 – 239.255.255.255) must be defined on all the interfaces of an RP router, but not on those used to reach the user networks (UN).

In this way the use of the inner RP is allowed for scoped addresses, utilized by the user networks, but the propagation of “scoped” multicast packets towards the backbone is halted.

It is worth pointing out that the user networks can create administrative domains only if the related groups are served by their own RP (inside the user network). Otherwise, if a CR is used as an RP, the whole scoped address range is “open” to all the users connected to that RP (CR).

The same principle is adopted to avoid the propagation of PIM auto-rp packets from the backbone to the user networks and vice versa. This is prerequisite when auto-rp is used to distribute the mapping between RPs and multicast groups inside the user network.

3.5 Shared RP-Tree and source-based STP

PIM-SM provides for two types of distribution trees:

- Shared or RP tree (RP-tree):
the path chosen is the optimal one from the point of view of the shared root, RP, with respect to the receivers. Each source sends the traffic to the RP which then forwards it along the shared tree.
- Source or shortest path tree (SP-tree):
the path chosen is the optimal one from the point of view of the router adjacent to the source with respect to the receivers. Traffic is distributed along a tree with its root on a router adjacent to the source.

PIM-SM initially creates a shared distribution tree for every multicast group. The disadvantages of a shared tree are that it tends to concentrate all the traffic on a limited number of links and that the paths between the source and receivers might not be the optimal thus introducing some latency in packet delivery.

PIM-SM allows the routers either to continue to receive the multicast traffic through the shared tree or to switch over to a shortest-path tree.

Switching from RP-tree to SP-tree is configurable. The parameter that controls this change is the threshold of traffic received from one source. At this time, on CISCO routers, the default value of this threshold is 0 indicating that the passage to SPT occurs when the first packet is received from the source.

3.6 Bandwidth limitations

Multicast and unicast traffic shares the common network infrastructure. To avoid an excessive bandwidth consumption and performance degradation of other network services a limit on multicast traffic must be imposed. This limit is specified as the maximum value, in

bits per second, that the multicast traffic can get on a network line. The proposed value for this threshold is one third of the line bandwidth. The same limit is imposed both on the backbone lines and on the user access lines. If required, it is possible to define a different threshold on the user access link. The maximum threshold cannot be greater than the lowest of the threshold values configured on the backbone.

3.7 *User network*

User networks should implement PIM-SM to comply with the backbone. For small user networks, with few multicast networks/hosts, it is advisable to use the backbone RPs. In this case the RP serves all multicast groups, including scoped ones.

For user networks in which a heavy local multicast traffic is expected, it is advisable to configure one or more local RPs (routers on the user network). In this way local multicast traffic is confined inside the user network and does not need to go through the CR router. A secondary effect of this choice is a greater control over used scoped addresses and a better guarantee of privacy for local multicast data traffic.

For global multicast addresses the reference RP is always the CR router.

If the local user RP has to serve all the multicast groups it must establish a MSDP peering with the backbone RP (CR router).

A local RP configuration is essential if the user wishes to establish multicast PIM-SM peering with networks external to GARR. This is necessary because these networks are connected through private networks (backdoors) and by unicast routing not visible to the GARR backbone routers. In this case the user must opportunely configure the multicast routing in order to avoid irregular traffic flows between GARR and non GARR networks (obviously the same rule must be applied to unicast traffic).

The choice of the RP configuration type is left to the user network managers. The various modalities are depicted in Figure 1.

3.7.1 *Multiple user access lines*

Some user networks are connected to the backbone by more than one line. Load splitting across multiple (equal cost) lines, between UR and CR, can be easily achieved for unicast traffic. Due to RPF check, load balancing of multicast traffic is nearly impossible. There were two approaches for this:

- to use only one line for multicast traffic;
- to define tunnels and allow underlying unicast mechanisms to perform load splitting;

There are drawbacks to both of those. The first one requires manual router configuration and does not provide automatic backup. The second approach needs close attention when configuring and adds an overhead on line utilization while decreasing routers' performance.

In the latest Cisco IOS, a new feature regarding multicast traffic load balancing has been introduced. It allows different RPF interfaces to be used for the same unicast route prefix. Of course, multiple paths must be of equal cost in order for RPF check to succeed.

3.8 *Inter-domain routing*

Since, at present, there is no multicast inter-domain routing protocol, in order to join two or more PIM-SM domains (belonging to different Internet service providers) the same MSDP-

based technique, applied inside the GARR network, can be used. This technique has been widely accepted by the ISPs as a temporary solution for the multicast inter-domain routing problem. However, if different unicast and multicast routing policies have to be applied, that is when the two topologies diverge, it is necessary to use MBGP protocol [7]. MBGP is a multi-protocol BGP extension that allows labeling multicast routing information and creation of two sets of routes: one for unicast and one for multicast. The second is used by PIM-SM to create the distribution trees and correctly deliver multicast packets. It is worth pointing out that using MBGP is independent from MSDP. This combination is necessary only when multicast traffic with national ISPs, international ISPs or GARR user networks must follow a different routing path (for instance through multicast dedicated links).

Even if GARR unicast and multicast topologies are congruent, it is advisable to immediately use MBGP with the user networks that already use BGP for unicast routing. This choice allows easy management of multicast routing between GARR and ISP.

3.9 *Management*

Management tools are indispensable for IP multicast deployment and management [4]. Over the last few years many tools have been developed to help locate and resolve problems related to multicast routing [5]. With the support of SNMP the multicast routing management can be integrated into Network Management tools already present at NOC. This would allow NOC operators to perform monitoring, configuration and data collection still using familiar tools.

3.9.1 Simple tools

The commands available on the router and Mbone tools are particularly useful for performing configuration checking and for identifying and tracing problems or inconsistencies in multicast routing. They show the state of interface, the presence of adjacent multicast routers, the active multicast groups, the delivery table, the multicast traffic, the mapping between multicast groups and RP, etc.

3.9.2 SNMP tools

Both the complexity and the distributed nature of IP multicast make its management particularly difficult. Analysis of the state of the topology and identification of active multicast groups, localization of networks and interfaces belonging to a distribution tree and traffic volume control for a group or for the aggregation on a link, are all difficult without the support of sophisticated tools. HP OpenView, with multicast extensions¹, is a tool suitable for this purpose. This tool does not simply collect SNMP variables from routers but allows sophisticated processing of collected data with the aid of graphic visualizations.

¹ HP OpenView multicast module has not yet been officially released but it is possible to obtain a test preview version

3.9.3 Multicast Routing Monitor

Multicast Routing Monitor (MRM, [8]) is a new protocol designed for multicast routing monitoring aid and, in particular, for locating routing anomalies and connectivity problems. Protocol operation is based on communication and coordination among its three components:

- **Manager:** provides an interface for configuring and running tests and then collecting and presenting results.
- **Test Sender:** generates multicast traffic based on Manager requests.
- **Test Receiver:** can join a group or perform passive monitoring of multicast group traffic.

MRM can be implemented both on routers and on workstations and any network node can generate traffic. Its unique features, lacking in other tools, are:

- Real-time monitoring (passive) and error logging.
- *Pro active* test (impact analysis and service provisioning control).

MRM is already implemented on CISCO routers.

3.10 International multicast links

The multicast routing architecture described in this paper – based on PIM-SM, MBGP and MSDP protocols – allows the natural integration of the GARR multicast infrastructure with those of neighboring ISPs.

The European multicast infrastructure (TEN-155) has already migrated from a single DVMRP-based multicast domain to a new hierarchical multicast routing model based on PIM-SM. Almost all the National Research Networks (NRN) have adopted this architecture.

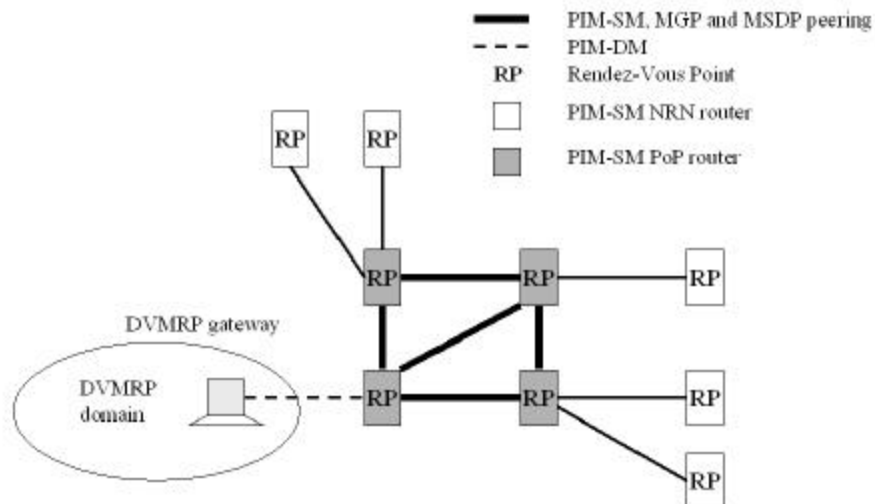


Figure 2. Logical model of present European multicast network

At present multicast connections between TEN-155 and NRN can be established:

- by tunneling PIM-SM protocol;
- by means of a native multicast connection based on ATM PVC between the PoP router and the NRN gateway router.

The logical network scheme, with MBGP and MSDP peerings, is depicted in Figure 2.

References

1. A.B. Bonito, D. Pobric "Introduzione al multicast e alla comunicazione multimediale su Internet/Intranet", Technical Report CNUCE-B4-1998-003, CNUCE-CNR, 1998.
2. T.Pusateri, "Distance Vector Multicast Routing Protocol", 08/09/2000 <draft-ietf-idmr-dvmrp-v3-10.txt, .ps>.
3. T. Ferrari, A. Pinizzotto, D. Pobric, M. Sommani, "Rete GARR-B: Piano di Routing IP-Multicast", Technical Report IAT-B4-2000-002, IAT-CNR, 2000.
4. K. Almeroth, "Managing IP Multicast Traffic: A First Look at the Issues, Tools and Challenges", A White Paper from the IP Multicast Initiative, January 1999.
5. D.Thaler, B. Aboba, "Multicast Debugging Handbook", 05/08/2000 <draft-ietf-mboned-mdh-04.txt>.
6. T.A.Maufer, "Deploying IP Multicast in the Enterprise", Prentice-Hall: Upper Saddle River, NJ, 1998.
7. T. Bates, R. Chandra, D.Katz, Y. Rekhter, RFC2283, "Multiprotocol Extensions for BGP-4", February 1998.
8. L. Wei, D. Farinacci, "Multicast Routing Monitor (MRM)", 02/01/1999 <draft-ietf-mboned-mrm-00.txt>.
9. D.Farinacci, Y. Rekhter, D. Meyer, P. Lothberg, H. Kilmer, J. Hall, "Multicast Source Discovery Protocol (MSDP)", 07/19/2000, < draft-ietf-msdp-spec-06.txt >.
10. D. Kim, H. Kilmer, D. Farinacci, D. Meyer, "Using MSDP to create Logical RPs", 03/24/1999, <draft-ietf-mboned-logical-rp-00.txt>.
11. D. Estrin, D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson, C. Liu, P. Sharma, L. Wei, RFC 2362, "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification", June 1998.
12. D. Meyer, RFC2365, "Administratively Scoped IP Multicast", July 1998.