

# Multicast Routing

(C:\Users\bcousin\Desktop\Multicast-routing.3.fm- 21 October 2009 16:29)

## OUTLINE

- Introduction
- IGMP Protocol
- DVMRP Protocol
- MOSPF Protocol
- PIM Protocol

## Bibliographie

- S. Paul, "Multicasting on the Internet", Kluwer academic publishers, 1998
- C. Huitema, "Routing in the Internet". Prentice Hall, 1999
- W. Stallings, "High Speed Networks", Prentice Hall, 1998
- C. Comer, "TCP/IP: architectures, protocoles, applications", InterEditions, 1998

## 1. Introduction

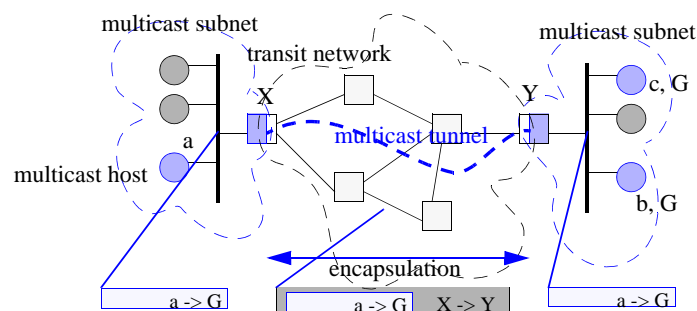
A set of representative multicast routing protocols:

- multicast membership between hosts and routers: IGMP
- multicast routing in dense mode: DVMRP
- extension of a well known unicast routing protocol: MOSPF
- multicast routing in sparse mode: PIM

### 1.1. The Mbone

**Multicast Backbone:** interconnection of multicast IP subnets

- the multicast subnets process multicast packets
  - all hosts, routers and links in the subnet support multicasting
- the multicast subnets are interconnected through a non-multicast (transit) network
  - multicast tunnels



**Datagram encapsulation:** multicast packet in unicast packet (IP in IP) :

- encapsulation at the border router between the source subnet and the transit network
- dis-encapsulation at the border router between the transit network and the destination subnet

## 2. IGMP Protocol

### 2.1. Introduction

Protocol is used by the hosts to **inform the multicast routers of the active groups.**

- IGMP : “Internet Group Management Protocol”
- rfc 1112 : “Host extensions for IP multicasting”

Definition :

- a group is active at a router interface if at least there is a member of this group on this interface.

IGMP message are encapsulated in IP datagrams:

- *Protocol* field of IP datagram = 2
- nota, with IPv6, IGMP is integrated into ICMP:
  - . Multicast Lister Discovery (MLD)

## 2.2. Principle

### 2.2.1 Group subscription

When a **host subscribes to a group** (with multicast IP address G):

- it adds the corresponding MAC group address to the network card
- it sends an IGMP *report* message
  - the *group address* field is the multicast address of the group.
  - the IGMP *report* message is encapsulated in an IP datagram with the *Destination address* is the multicast address and the *TTL* is equal to 1
    - => local multicast routers listen all multicast packets

### 2.2.2 Group Discovery

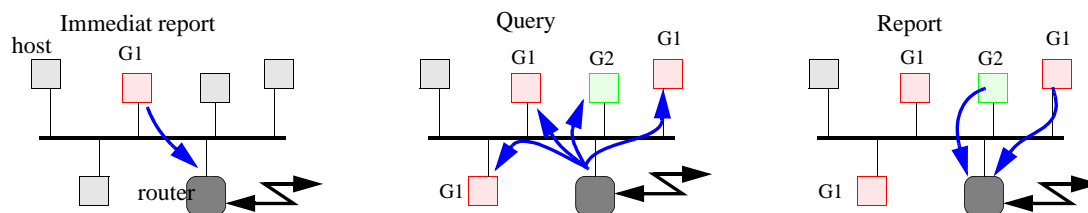
Multicast routers **monitor the active groups**:

- routers send IGMP *query* messages
  - periodically : by default, every 125 s
    - . not too often to limit the overhead
    - . not too seldom to have a good picture of the membership
- the IGMP *query* message is encapsulated in an IP datagram with the *Destination address* 224.0.0.1 and the *TTL* is equal to 1.

No action is required to monitor the group 224.0.0.1: it is supposed to always be active.

Each member of a group should **respond to the query**:

- it sends an IGMP *report* message after a **random delay** (default value: [0 - 10 s])
- the IGMP *report* message is encapsulated in an IP datagram where *Destination address* is the multicast address and the *TTL* is equal to 1.
- if, during the delay, another group member replies, the *report* message is not sent
  - => in general, one reply message is sent for each group and for each period (the traffic is minimized)



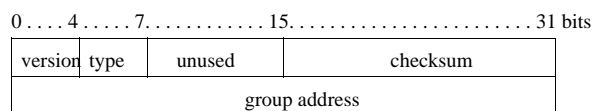
### 2.2.3 Group Leaving

A host **leaves a group**:

- the host ceases all IGMP transmission about this group
- during the next query phase, if the router does not received any report about this group, the group is declared inactive
  - it was the last host of the group on this router interface
    - => multicast packet delivery could be suboptimal (useless) for the remaining part of the period

### 2.3. General Format of IGMP Message

IGMP message size : 8 bytes



Version :

- Historical version = 1 : rfc 1112 (1989), old version = 0 : rfc 988

Two types de messages IGMP :

- “Host membership **query**” = 1
- “Host membership **report**” = 2

Checksum computation:

- one's complement of the one's complement sum of the 16-bits words of the header
- same computation as TCP, UDP or IP

**Group address** :

- The multicast IP address which identifies the group

## 2.4. Designated Router

When a subnet has several routers, one is designated to manage multicast membership on the subnet:

- The designated router (DR) : the one with the smallest IP address
- When a router receives an IGMP *query* message sent by another smallest router
  - . it stops sending IGMP *query* message
  - . it monitors the periodic sending of IGMP *query* messages (by the DR)

Nota: DR election process is defined since IGMP version 2.

In IPv6 : the DR is called the Querier.

## 2.5. Failure Management

### Packet Corruption

- Datagrams with incorrect checksum or incorrect destination address are silently discarded.

### Packet loss

- IGMP message losses are resolved by message periodic repetition :
  - => temporarily, multicast packet delivery can be incorrect

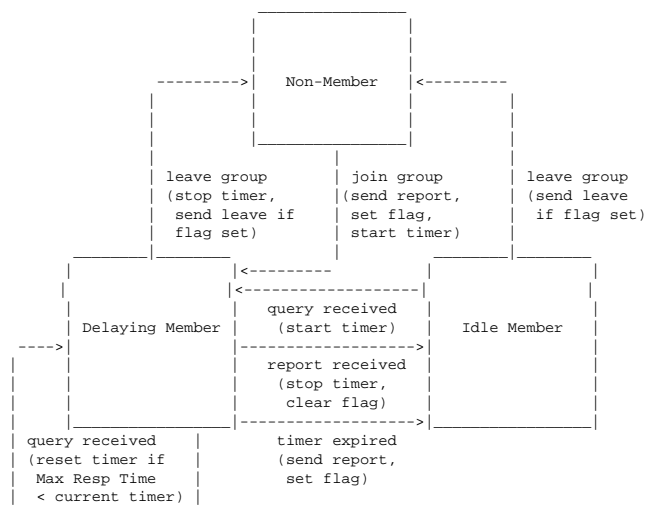
### Host failure

- - !

### Router failure

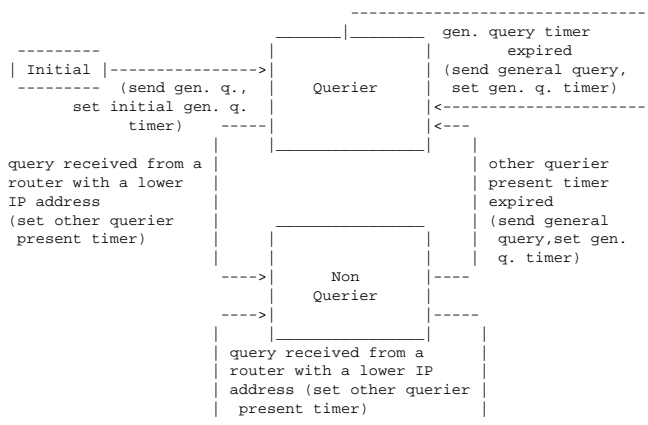
- If the failing router is the last or only one which connects hosts to the internet then the hosts are disconnected else one of the redundant routers will be elected as DR.

### 2.6. IGMPv2 Host State Diagram



- "set flag" that we were the last host to send a report for this group.
- "clear flag" since we were not the last host to send a report for this group.

### 2.7. IGMPv2 Router State Diagram

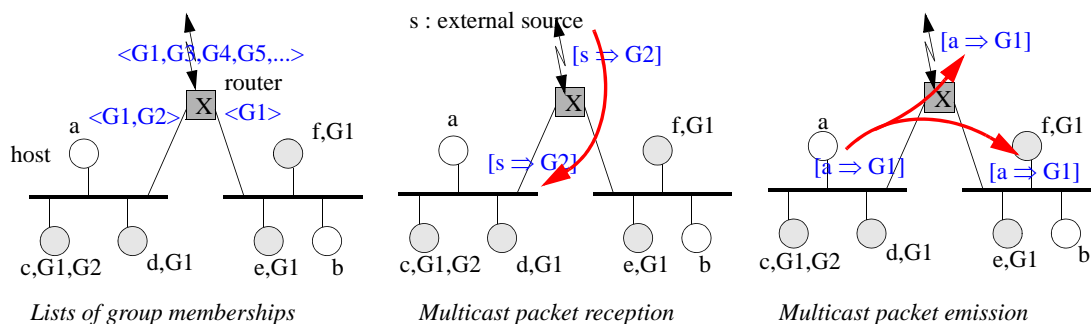




## 2.9. Forwarding of multicast datagrams

When a router receives a multicast packet, it forwards the packet on every interface where a member is active (except on the interface on which the datagram has been received).

A router knows that a group is active on an interface thanks to IGMP.



## 3. DVMRP protocol

### 3.1. Presentation

Distance vector multicast routing protocol: rfc 1075 (1988)

- similar to the routing algorithm: **distance vector**
- use **RPF with pruning**
- DVMRP is experimental:
  - DVMRP messages are extensions of IGMP messages
- used by the Mbone, when most of the Internet routers were multicast incompatible:
  - management of multicast tunnels

DVMRPv3 : T. Pusateri, "DVMRP", draft-ietf-idmr-dvmrp-v3.txt, 1999.

Hierarchical DVMRP : A. Thyagarajan, S. Deering, "H-DVMRP", SIGCOMM, 1995.

### 3.2. Principe

Every routers exchange DV routing messages toward their neighbors:

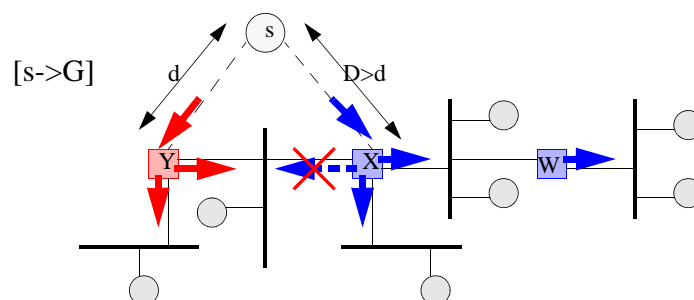
- a list of couples <destination, distance>
  - destination: address et subnet mask of a potential multicast source!
  - distance: the number of hops between the destination and the local router

At the end, each router knows for every destination

- the interface which gives access to the shortest path toward each destination (we called this interface, the upstream interface of this destination)

### 3.3. Reverse Path Multicasting

- When a multicast router receives a multicast packet on the upstream interface associated with the source of the multicast packet:
  - it forwards the multicast packet **all its interfaces** except those which leads to a router which is on a shortest path toward the source of the multicast packet
  - else the multicast packet is discarded



- This algorithm computes a (total) **spanning tree** whose root is the multicast source
- This tree has two properties:
  - use the shortest path from the multicast member to the multicast source
  - the tree depends of the source
    - => for the same multicast group, the traffic load from the different sources can be spread

### 3.4. RPM pruning

Routers should not send multicast packets belonging to a group on a subnet where no group member is active.

- Group membership is monitored by the IGMP protocol

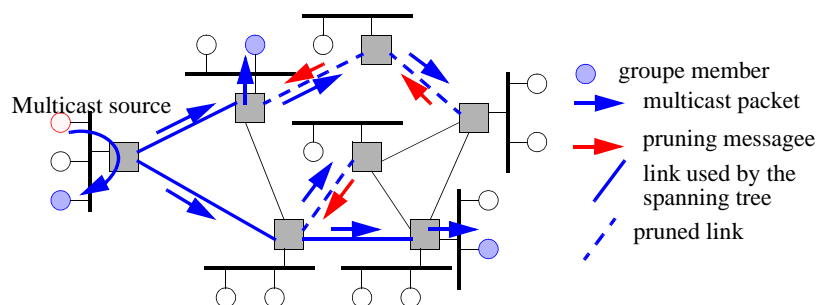
Pruning of the useless branches of the spanning tree:

- a router sends a pruning message for a group:
  - when all its interfaces are inactive, for this group
- toward the upstream router, for this group

Recursively

nota : the downstream/upstream direction is determined by the multicast data flow direction

- Example :



After convergence, the tree contains only the branches which have a group member.

Multicast routing data are put in a **memory cache for a while** in the routers:

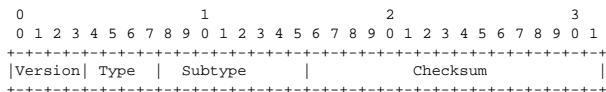
- periodically, for each group, multicast data packets are flooded on every router interface where there is a pruned downstream router
  - if there is **no group member**, a pruning message is sent back by the downstream router
  - if there are **new group members**, they received the multicast data packets
- if a multicast source stops
  - the multicast state in the router memory cache is clean up, naturally

### 3.5. Format of DVMRP Messages

Structure of DVMRP messages: a DVMRP header followed by commands.

The first word of DVMRP messages is compatible with IGMP messages.

#### 3.5.1 DVMRP Message Header



The version is 1.

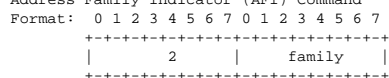
The type for DVMRP is 3.

The subtype is one of:

- 1 = Response; the message provides routes to some destination(s).
- 2 = Request; the message requests routes to some destination(s).
- 3 = Non-membership report; the message provides non-membership report(s). ("prune message")
- 4 = Non-membership cancellation; the message cancels previous non-membership report(s). ("graft message")

#### 3.5.2 DVMRP Commands

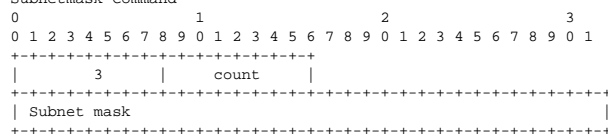
##### 1. Address Family Indicator (AFI) Command



Values for family:

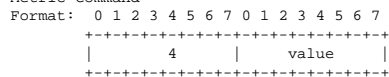
- 2 = IP address family, in which addresses are 32 bits long. Default: Family = 2.

##### 2. Subnetmask Command

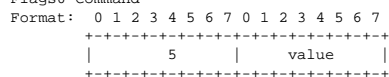


Count is 0 or 1.

##### 3. Metric Command



##### 4. Flags0 Command



Meaning of bits in value:

- Bit 7: Destination is unreachable.
- Bit 6: Split Horizon concealed route.

5. Destination Address (DA) Command

```

0          1          2          3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+
|          7          | count          |
+-----+-----+-----+-----+
| Destination Address1 |
+-----+-----+-----+-----+
| Destination Address2 |
+-----+-----+-----+-----+
etc.

```

6. Requested Destination Address (RDA) Command

```

0          1          2          3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+
|          8          | count          |
+-----+-----+-----+-----+
| Requested Destination Address1 |
+-----+-----+-----+-----+
| Requested Destination Address2 |
+-----+-----+-----+-----+

```

7. Non Membership Report (NMR) Command

```

0          1          2          3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+
|          9          | count          |
+-----+-----+-----+-----+
| Multicast Address1   |
+-----+-----+-----+-----+
| Hold Down Time1     |
+-----+-----+-----+-----+
| Multicast Address2   |
+-----+-----+-----+-----+
| Hold Down Time2     |
+-----+-----+-----+-----+

```

### 3.6. DVMRP Message Examples

- DVMRP *response* for two Internet destinations 128.2.251.231 and 128.2.236.2 with distance = 2, and subnet mask = 255.255.255.0:

```

Subtype 1,
AFI 2, Metric 2, Infinity 16, Subnet Mask 255.255.255.0
{2} {2} {4} {2} {6} {16} {3} {1} {255} {255} {255} {0}

DA Count=2 [128.2.251.231] [128.2.236.2]
{7} {1} {128} {2} {251} {231} {128} {2} {236} {2}

```

- DVMRP *request* for any destination:

```

Subtype 2, AFI 2, RDA Count = 0
{2} {2} {8} {0}

```

- DVMRP *pruning* for groups 224.2.3.1 et 224.5.4.6 with an hold down time = 20 seconds, and for the group 224.7.8.5 with an hold down time = 40 seconds.

```

Subtype 3,
AFI 2, NMR Count = 3 [224.2.3.1, 20]
{2} {2} {10} {3} {224} {2} {3} {1} {0} {0} {20}

[224.5.4.6, 20] [224.7.8.5, 40]
{224} {5} {4} {6} {0} {0} {0} {20} {224} {7} {8} {5} {0} {0} {0} {40}

```

### 3.7. Management of Tunnels

Multicast data packets are encapsulated in unicast packet with a "loose source routing" option:

| Field          | Value                      |
|----------------|----------------------------|
| -----          | -----                      |
| src address    | = src gateway address      |
| dst address    | = dst gateway address      |
| LSRR pointer   | = points to LSRR address 2 |
| LSRR address 1 | = src host                 |
| LSRR address 2 | = multicast destination    |

Each router manages its tunnels :

- the IP address of the distant end router
- the cost of the tunnel ( $\geq 1$ )
- a **threshold**

The DVMRP routers forward a packet on an tunnel, iff

- its **packet TTL** is higher than the tunnel threshold
- limit the broadcasting scope of the multicast packets
  - conventional threshold values of a border tunnel are:
    - . for a company site = 32
    - . for a world region = 64
    - . for a continent = 128
- similar to the IPv6 *scope* field

## 4. MOSPF protocol

### 4.1. Presentation

DVMRP and tunnels are inappropriate, when multicast transmissions are generalized:

- Tunnel management is difficult:
  - Too large number of tunnels
    - . for instance, a complete mesh between N multicast subnets has  $N^2/2$  tunnels
- DVMRP is inefficient in sparse network:
  - flooding + pruning ... then flooding + pruning ...

MOSPF : **multicast OSPF** ("Open short path first")

- Rfc 1584 : "Multicast extensions to OSPF", march 1984
- Multicast routing inside a routing domain (AS: "autonomous system")
- Definition of a new OSPF record:
  - distant routers can know the location of group members

## 4.2. Principe

### MOSPF:

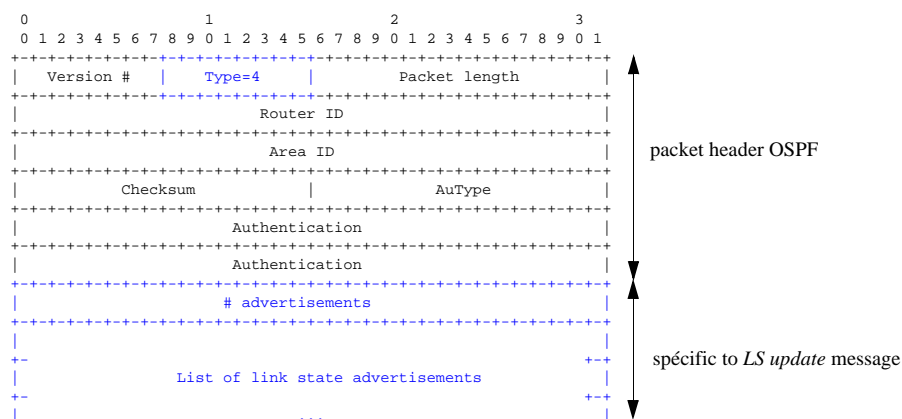
- an extension of "OSPF protocol version 2" (rfc 1583)
- **Link-state protocol:**
  - Every router sends to all network routers the state of its links
    - . each router gets a complete topology of the network
  - Every router computes the best route towards each destination using SPF algorithm
- Definition of a **new OSPF record** (Link-State-advertisement):
  - routers know the location of the group members, for all groups in the network
- Every router could compute a shortest path tree **using SPF algorithm**

When a MOSPF router received a multicast packet from S for group G:

- The router computes, **on demand**, a tree (from the source S to all members of group G)
  - the shortest path tree
    - . using the "shortest path first" algorithm [Dijkstra]
- a different tree can be computed for each metric (CoS: Class of service)
  - => the multicast data packet is then forwarded

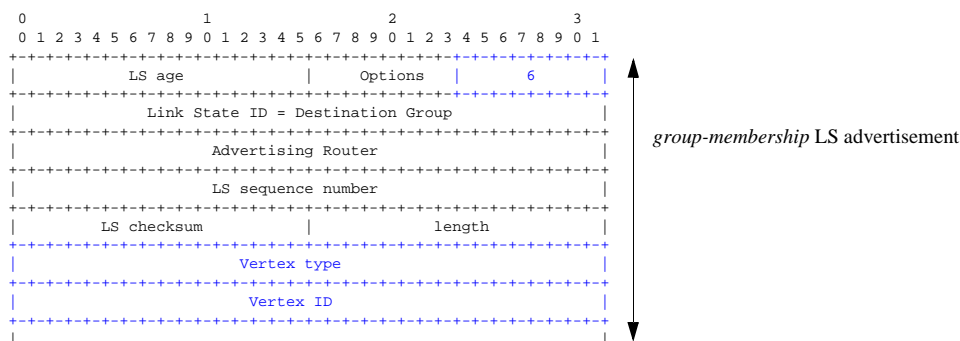
## 4.3. MOSPF Messages

### Link-state update OSPF message type



### 4.4. Group-membership LSA

Code value of a *group-membership* LSA is 6

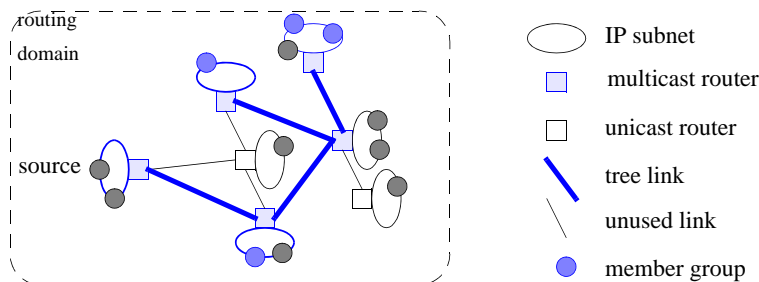


- standard header of a LSA : 20 bytes
  - Link State ID = multicast address of the group
- list of vertices associated with the multicast address
  - Each vertex is associated with:
    - . Vertex type : 1 = router, 2 = transit network
    - . Vertex ID : OSPF ID of the router or IP address of the transit network DR
- a router sends a group-membership LSA iff the router is the DR of a subnet where it exists a group member

### 4.4.1 Heterogeneous Networks

Routing domain can be heterogeneous from a multicast point of view:

- The multicast routers can be identified by a specific bit of the *option* field in the OSPF messages
  - multicast paths are only computed over multicast routers

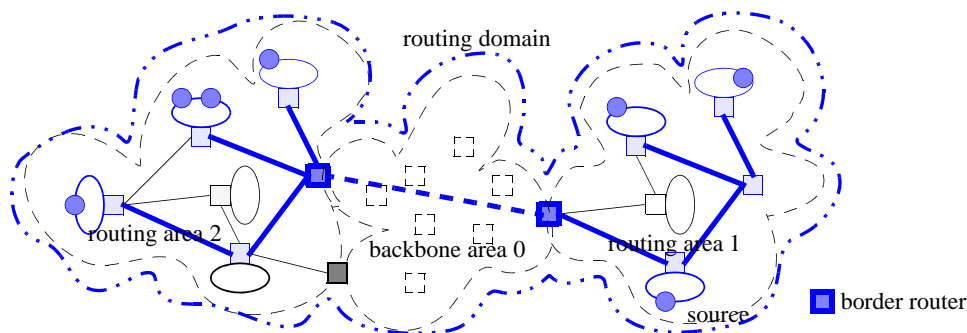


### 4.4.2 Routing Areas

The routing domain is split in several routing areas. They are interconnected by the backbone area (area #0).

- The border routers collect the group membership in their area
  - they advertize the active group in the backbone area
- It may exist several border routers in a routing area

- a distance metric determines the router on the shortest path, it is selected
- a tie-breaker is used to separate the ex-aequo (the lowest address router is selected)
- A border router is, by default, a **group member for any group** in its routing area:
  - it receives a copy of all multicast packets sent by a local source, and forward them on the backbone area.
- A border router becomes, in its area, **the root of a group** when there exists at least one group member in the area and the source is distant



## 5. PIM Protocol

### 5.1. Presentation

When the network is very large **MOSPF is not well suited**

- The computation complexity is the square of the number of network nodes
- The cost of the OSPF message broadcasting sharply increases:
  - message sending frequency should be greater than the join and leave frequency of group member (which much higher than the frequency of topology change)
- MOSPF introduces large delay
  - paths are computed on demand, i.e. when the first multicast data packet arrived at a router

⇒ PIM (**Protocol independent multicast**) protocol

Two modes of PIM protocol:

- PIM dense mode
- PIM sparse mode
  - depends of the **node density** in the network

PIM protocol makes the assumption that somehow the unicast routing table is available:

- every multicast router knows its next hop toward any destination
- PIM is **independent of the unicast routing protocol** used to produce this routing table

S. Deering & al., "PIM-SM : protocol specification", rfc 2362, June 1998.

A. Adams, J. Nicholas, W. Siadak, "Protocol Independent Multicast - Dense Mode : protocol specification", rfc 3973, January 2005.

**PIM dense mode:**

- RPF algorithm with pruning (cf. S. Deering)
- similar to DVMRP,
  - but do not broadcast distance vector messages:
    - . uses any available unicast routing table, instead
  - the paths have symmetrical performances:
    - . the shortest path is the same on one way and on the other

The two PIM modes are compatible, they share the same message format

## 5.2. PIM sparse mode protocol

For sparse groups:

- The number of group members is small against the number of network hosts
- The group members are widely spread

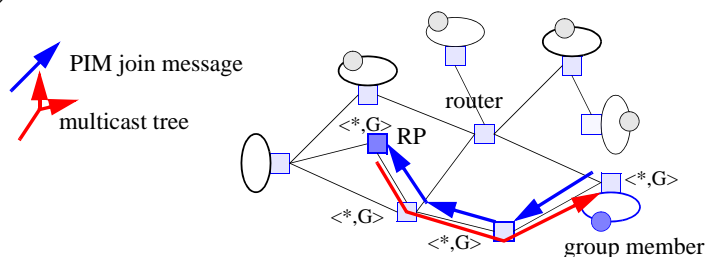
This protocol can be deployed for multicast routing between routing domains (AS).

## 5.3. PIM-SM Principle

- One router is chosen for the rendez-vous point (**RP**), for each group
  - The location of the RP is determinant to produce an efficient multicast tree

### 5.3.1 First Multicast Destination

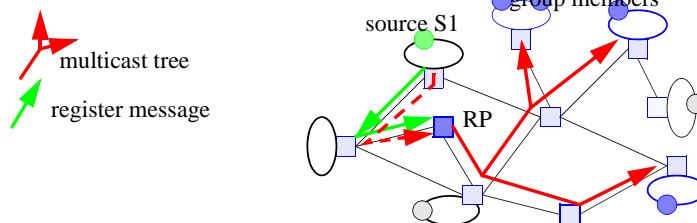
- A router with at least an active group member sends periodically:
  - a PIM *join* message ( $\ast, G$ )
  - toward the group RP, following the shortest unicast path
    - => the path becomes a branch of the multicast group tree (the tree of reverse shortest paths!)



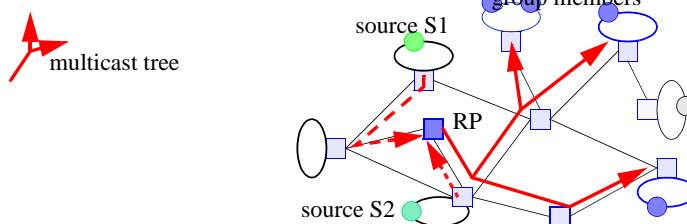
- When a router receives a PIM *join* message, it creates a multicast entry ( $\ast, G$ ) associated with the address of the neighbor router which has sent the join message as the next hop

### 5.3.2 Multicast sources

- When a source sends a multicast data packet to a group:
  - The designated router for this source unicasts a PIM *register message* toward the RP
  - which encapsulates the multicast data packet



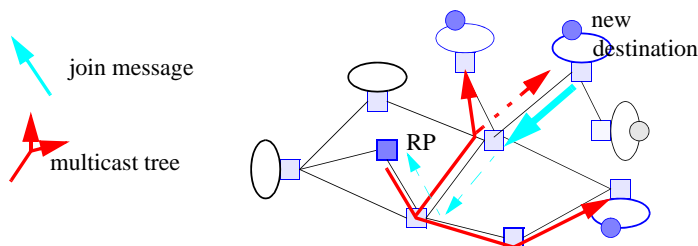
- On reception of the register message, the RP forwards the multicast data packet on the multicast tree, when it exists
- When there are several sources, their data flows convergent toward the RP



### 5.4. PIM Optimizations

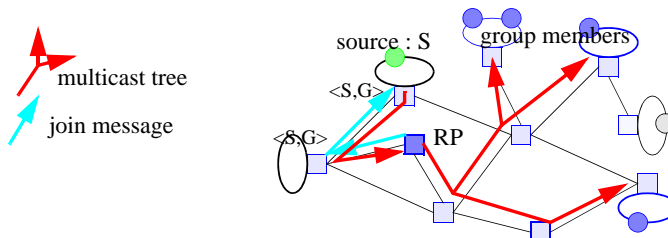
A PIM message sent to the RP can be processed by a on-the-path router which already belongs to the multicast tree:

- the *join* message does not need to be forwarded toward the RP



When a source sends numerous multicast data packets, the RP may decide to build a specific multicast branch for this multicast source:

- The RP sends a *join* message (S,G) and an <S,G> state is created in the on-the-path routers



Until now, with PIM,

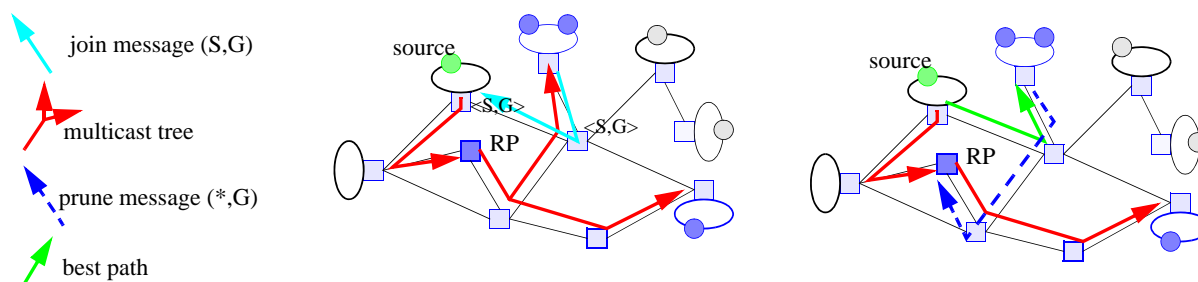
- The multicast data packets
  - sent by a source S to a group G follow
    - . between S and the RP, **the reverse shortest path**
    - . then between the RP and each group member, the reverse shortest path
- PIM computes a partial spanning tree which produces
  - a minimal number of data packet copies for each multicast data packet
  - few additional PIM control messages
    - . a join message for each group member on the exclusive part of the multicast tree

However:

- The concatenation of two shortest path is not a shortest path
  - => The router receiving a multicast data packet from a source for a group may decide to use **a best path** to receive the multicast data packets from that specific source

Computation of a best path:

- the router detects it receives multicast data packets from a source S on an interface which is not the shortest path
- the router, for this group, sends to the multicast source S
  - a **join message (S,G)**
- the source sends the multicast data packets toward the designated router
- when the router receives multicast data packets from the source on the new interface, it sends :
  - a pruning message **(\*,G)** toward the RP



Nota : all the sources keep sending multicast data packets toward the RP for the others group members

## 5.5. Multicast Data Transmission Halt

### 5.5.1 Group leave

A router with no activity for a group, leaves the group

- explicitly sending a *prune* (.,G) message
- implicitly without any new sending of *join* (.,G) message

The routing state of the routers are clean

- implicitly when it is not refreshed
- explicitly when the *prune* message is received

### 5.5.2 PIM *register-stop* message

An RP sends *register-stop* message to a source when

- the RP has no group members and has received a multicast data packet for that group (encapsulated in a *register* message)
- the RP receives simultaneously native multicast data packets and encapsulated multicast data packets for the same group.

## 5.6. RP election

Routers, which are RP-candidate, send a message to the Bootstrap Router (BSR) :

- PIM Candidate-RP-Advertisements unicast message
- a router may be candidate for a certain subset of groups
- a priority may be associated to each router

PIM messages are broadcast periodically by the BSR to all PIM routers :

- PIM *Bootstrap* Message
- list all the RP-candidate routers
- election of the bootstrap router (BSR)
  - . the router with the highest priority et address
  - . similar to the spanning tree protocol

At each router, a function, for a group, select the appropriate RP-candidate:

- the router with the highest priority et address



### 5.7.4 PIM join or prune messages

