

Les protocoles de routage multicast

(Z:\Polys\Multicast\Le_multicast.routage.fm- 17 octobre 2007 13:16)

PLAN

- Introduction
- Le protocole IGMP
- Le protocole DVMRP
- Le protocole MOSPF
- Le protocole PIM
- le protocole BGMP
- Conclusion

Bibliographie

- S. Paul, "Multicasting on the Internet", Kluwer academic publishers, 1998
- C. Huitema, "Le routage dans l'Internet", Eyrolles, 1995
- W. Stallings, "High speed networks", Prentice Hall, 1998
- C. Comer, "TCP/IP: architectures, protocoles, applications", InterEditions, 1998

1. Introduction

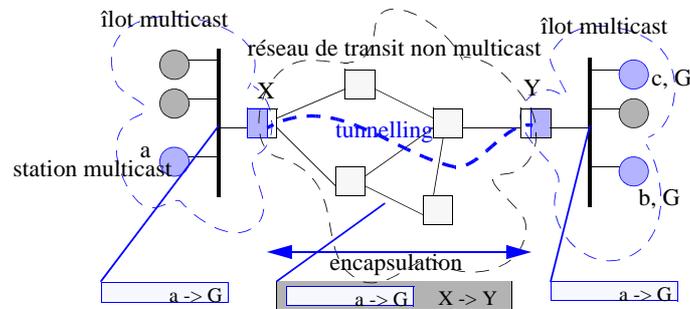
Nous allons étudier un ensemble représentatif de protocoles et de techniques de routage multicast :

- routage multicast entre stations et routeurs : IGMP
- routage multicast en mode dense : DVMRP
- extension d'un protocole de routage unicast : MOSPF
- routage multicast en mode clairsemé : PIM
- routage multicast inter AS : BGMP

1.1. L'infrastructure du Mbone

“**Multicast Backbone**” : interconnexion de réseaux IP multicasts

- chaque réseau multicast forme un **flot**
- interconnectés par des “**tunnels**”
 - liaisons virtuelles entre flots multicasts
- utilise un réseau de transit non multicast : l'Internet actuel



Encapsulation des datagrammes multicasts dans des datagrammes unicasts (IP in IP) :

- encapsulation au routeur d'interface entre réseau multicast local et réseau de transit
- désencapsulation à l'autre routeur d'interface entre réseau de transit et réseau multicast distant

2. Le protocole IGMP

2.1. Introduction

Protocole utilisé par les stations pour **informer les routeurs multicasts des groupes actifs**.

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

Définition :

- un groupe est actif localement à un réseau IP si au moins une station de ce réseau IP appartient au groupe.

Les messages IGMP sont encapsulés dans des datagrammes IP :

- champ Protocol du datagramme = 2
- note : pour IPv6, ICMP est intégré à IGMP

2.2. Principe

2.2.1 Adhésion

Pour **adhérer à un groupe** :

- une station émet un message “IGMP report”
 - le champ “group address” contient l’adresse du groupe auquel la station veut adhérer.
- le message “IGMP report” est encapsulé dans un datagramme dont le champ “Destination address” est l’adresse du groupe et le TTL égal à 1
 - => les routeurs multicasts sont à l’écoute de tous les datagrammes multicasts
- cette première transmission est répétée après un délai aléatoire
 - => lutte contre les pertes

2.2.2 Liste des groupes actifs

Les routeurs multicasts **surveillent les groupes actifs** :

- les routeurs envoient un message “IGMP query”
 - périodiquement (mais pas trop souvent pour limiter la surcharge) : par défaut = 125 s
 - permettent à un routeur de rafraîchir sa connaissance
 - => panne de station
- le message “IGMP query” est encapsulé dans un datagramme dont le champ “Destination address” = 224.0.0.1 et le champ TTL = 1.

Toute machine multicast est abonnée au groupe 224.0.0.1. Ce groupe est considéré comme toujours actif. Aucune action de surveillance n’est nécessaire pour ce groupe.

Chaque station appartenant à un groupe **répond à l’interrogation** :

- en émettant un message “IGMP report” après un **délai aléatoire**, par défaut : [0 - 10 s] !
- le message IGMP est encapsulé dans un datagramme dont le champ “Destination address” est l’adresse du groupe et le TTL =1.
- si avant l’écoulement du délai une autre station appartenant au même groupe répond l’émission est annulée
 - => en général une seule réponse par groupe (minimisation du trafic)

2.4. Election du "designated router"

Lorsqu'un sous-réseau possède plusieurs routeurs, un seul doit avoir la charge de gérer l'activité des groupes multicast:

- Le "designated router" (DR) : celui dont l'adresse est la plus petite
- lorsqu'un routeur constate qu'un autre routeur émet des messages "IGMP query"
 - . il cesse d'en émettre
 - . il surveille l'émission régulière des messages "IGMP query"

Note : cette élection du DR n'est définie que depuis la version 2 d'IGMP.

2.5. Gestion des erreurs

Corruption

- Les datagrammes dont le checksum est erroné ou dont l'adresse de destination du datagramme ne correspond pas au champ "Group address" sont détruits silencieusement.

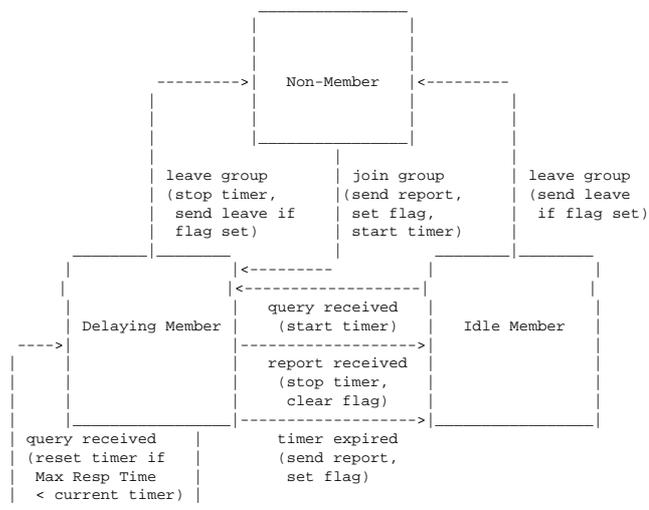
Perte

- Les pertes de messages IGMP sont traitées par répétition :
 - => transitoirement la distribution des messages multicasts peut être incorrecte

Panne

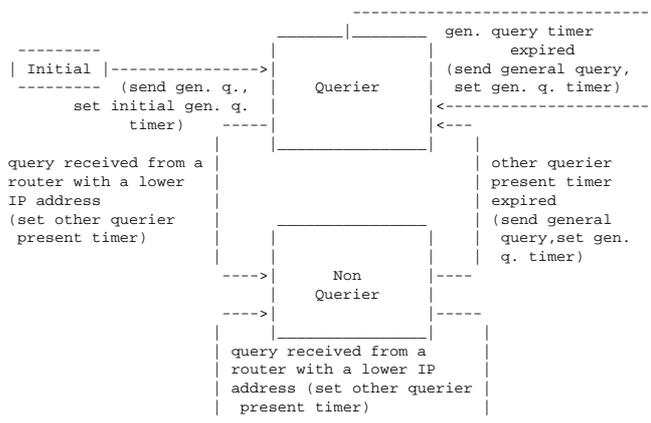
- La panne d'un routeur multicast peut isoler les stations d'un groupe s'il n'y a pas de routeurs supplémentaires, sinon celui le remplacera.

2.6. L'automate des stations IGMP



"flag" : la station a envoyé le dernier message "IGMP report"

2.7. L'automate des routeurs IGMP

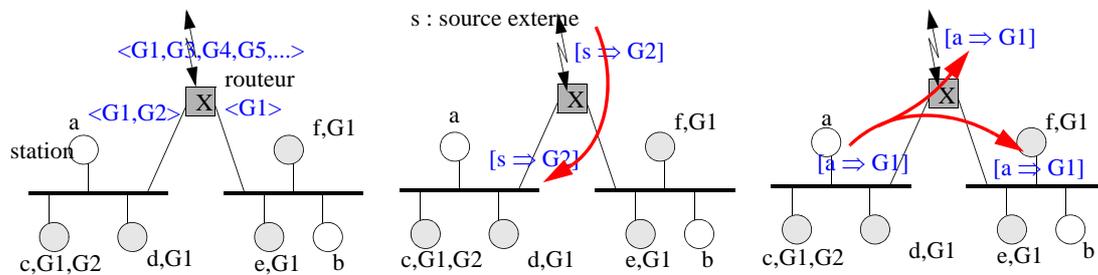


2.9. Routage local des datagrammes multicasts

Un routeur recevant un datagramme muni d'une adresse multicast identifiant un groupe le diffuse sur tous les interfaces où le groupe est actif (sauf celui d'où provient le datagramme).

Un routeur multicast n'a besoin de ne connaître que l'activité d'un groupe. Il ne mémorise que cette seule information.

=> chaque routeur pour chacune de ses interfaces maintient la liste des groupes actifs.



3. Le protocole DVMRP

3.1. Présentation

“Distance vector multicast routing protocol” : rfc 1075 (1988)

- même principe que l'algorithme de routage : “**distance vector**”
- utilise l'algorithme du **RPF avec élagage**
- DVMRP est expérimental :
 - les messages DVMRP sont des extensions des messages IGMP
- utilisé par leMbone, dans les premiers temps où les routeurs de l'Internet n'étaient pas multicast :
 - gestion des tunnels

Version 3 de DVMRP : T. Pusateri, "DVMRP", draft-ietf-idmr-dvmrp-v3.txt, 1999.

Hierarchical DVMRP : A. Thyagarajan, S. Deering, "H-DVMRP", SIGCOMM'95.

3.2. Principe

Les routeurs échangent avec leurs voisins des messages de mise-à-jour du vecteur de distance :

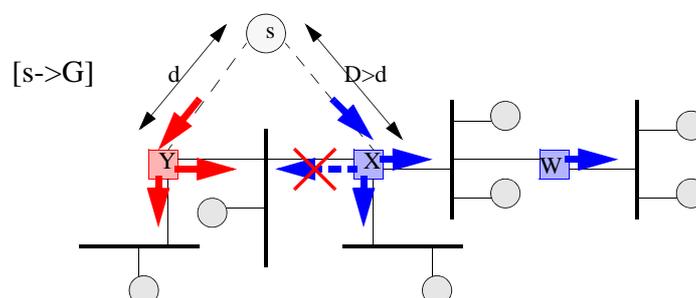
- liste de couples <destination, distance>
 - destination : l'adresse et le masque de sous-réseau de l'émetteur multicast !
 - distance : nombre de routeurs intermédiaires entre la destination et le routeur local

A l'issue de l'algorithme, chaque routeur connaît pour chaque destination

- l'interface qui donne accès au plus court chemin vers cette destination (c'est à dire l'émetteur multicast)

3.3. L'algorithme RPM ("Reverse path multicasting")

- Quand un routeur multicast reçoit un paquet multicast sur une interface il vérifie si cette interface est celle du (plus court) chemin vers l'émetteur du paquet multicast:
 - si ce n'est pas le cas le paquet est détruit
 - sinon il relaye alors le paquet sur toutes les interfaces sauf celles qui hébergent un routeur proposant un plus court chemin vers l'émetteur



- Cet algorithme construit un arbre recouvrant total dont la racine est l'émetteur
- Cet arbre a 2 propriétés :
 - utilise le chemin le plus court de la source à chaque destination
 - l'arbre dépend de la source
 - => pour un même groupe la charge des différentes sources est répartie

3.4. l'algorithme RPM avec élagage

Dans la version précédente tout paquet multicast est diffusé à tous les routeurs du réseau mondial si les seuils le permettent.

=> risque de congestion

- les routeurs terminaux, s'ils constatent qu'il n'y a aucun abonné sur leurs réseaux locaux, détruiront le paquet

On se propose d'élaguer les branches inutiles :

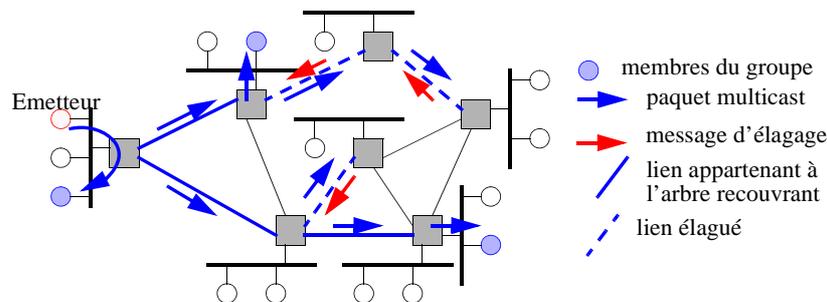
- le routeur terminal (aval) lors de la réception d'un paquet multicast va envoyer un **message d'élagage** ("pruning") :
 - "n'envoyez plus de paquets multicasts pour cet source multicast, pour ce groupe multicast et sur l'interface considérée"
- au routeur amont (qui mène à la source multicast)

De manière récursive :

- si le routeur amont reçoit sur toutes ses interfaces aval un message d'élagage
- il envoie un message d'élagage à son propre routeur amont

note : la notion aval/amont est déterminée par le sens naturel d'écoulement des paquets multicast

- Exemple :



Après convergence, l'arbre ne contiendra que les seules branches qui mènent à un abonné.

Ces informations ne sont **mémorisées que durant un moment** au sein des routeurs :

- périodiquement des paquets multicasts de données du groupe seront propagées
 - s'il n'y a toujours pas d'abonné, un message d'élagage est reçu en retour
 - s'il y a de **nouveaux abonnés**, il reçoivent alors les paquets multicasts
- si **l'émetteur se tait**
 - l'état des routeurs sera nettoyé naturellement

3.5. Les messages DVMRP

Structure du message DVMRP : une entête suivie des commandes.

Le premier mot est compatible avec celui des messages IGMP.

3.5.1 L'entête des messages DVMRP

```

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
+-----+-----+-----+-----+
|Version| Type | Subtype   |          Checksum          |
+-----+-----+-----+-----+

```

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 Les commandes DVMRP

1. Address Family Indicator (API) Command

Format: 0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7

```

+-----+-----+
|          2          | family |
+-----+-----+

```

Values for family:

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

2. Subnetmask 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
+-----+-----+-----+-----+
|          3          | count |
+-----+-----+-----+-----+
| Subnet mask          |
+-----+-----+-----+-----+

```

Count is 0 or 1.

3. Metric Command

Format: 0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7

```

+-----+-----+
|          4          | value |
+-----+-----+

```

4. Flags0 Command

Format: 0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7

```

+-----+-----+
|          5          | value |
+-----+-----+

```

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. Exemples de messages IGMP

- Transmet une route vers les destinations 128.2.251.231 et 128.2.236.2 ayant une distance de 2, l'infini valant 16 et un subnetmask de 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}

```

- Demande une route pour toutes les destinations possibles.

```

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

```

- Message d'élagage pour les groupes 224.2.3.1 et 224.5.4.6 avec un "hold down time" de 20 secondes, et pour le groupe 224.7.8.5 avec un "hold down time" de 40 secondes.

```

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. La gestion des tunnels par DVMRP

Les paquets multicast sont transformés en paquets unicast avec l'option "loose source routing" :

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

Chaque routeur gère ses tunnels :

- l'adresse du routeur de l'autre extrémité du tunnel
- le coût du tunnel (≥ 1)
- un **seuil**

Les routeurs DVMRP relayent un paquet sur une interface

- seulement si son **TTL est plus élevé que son seuil**
 - limite la portée des paquets multicasts
 - valeur conventionnelle des seuils des routeurs en frontière :
 - . d'une organisation = 32
 - . d'une région du monde = 64
 - . d'un continent = 128
- fonction similaire au champ "scope" de IPv6

4. Le protocole MOSPF

4.1. Présentation

La technique précédente (DVMRP et tunnels) est inadaptée si on envisage la généralisation des transmissions multicasts :

- La multiplicité des tunnels :
 - maillage complet des liaisons virtuelles ($N^2/2$)
 - gestion spécifique et lourde
 - induit des répétitions multiples des mêmes paquets sur le même interface
- technique précédente peu efficace:
 - inondation + élagage ... puis inondation + élagage puis ...

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

- rfc 1584 : "Multicast extensions to OSPF", mars 1984
- routage à l'intérieur d'un domaine de routage (AS : "autonomous system")
- définit un nouvel enregistrement :
 - permet aux routeurs de prendre connaissance de la localisation des groupes actifs

4.2. Principe

MOSPF, c'est :

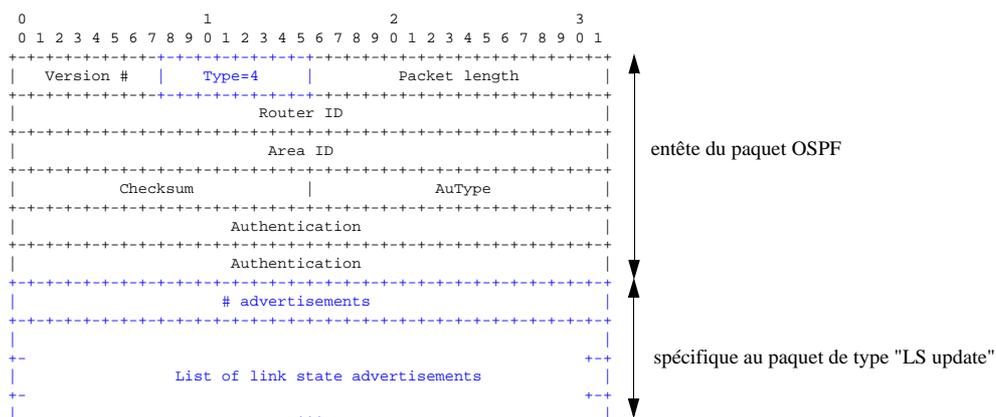
- Extension du protocole OSPF : version 2 (rfc 1583)
- **Protocole à état des liaisons** ("link-state protocol") :
 - chaque routeur diffuse à tous les routeurs l'état des liaisons
 - . chaque routeur obtient donc une connaissance de la totalité de la topologie
 - chaque routeur peut calculer localement les meilleures routes
- Définit un **nouvel enregistrement** :
 - chaque routeur prend connaissance des groupes actifs
- Calcul local d'**arbres recouvrants**

Lorsqu'un routeur MOSPF reçoit un paquet de donnée multicast (s -> G) :

- le routeur utilise sa connaissance de la topologie du réseau et de la répartition des abonnés au groupe G pour **construire** ("on demand") **l'arbre ayant pour racine s**
 - l'arbre des plus courts chemins
 - . l'algorithme du "shortest path first" [Dijkstra] entre la source et chaque routeur
- un arbre différent peut être construit pour chaque ToS
 - => le paquet de données multicast est relayé vers les bons interfaces

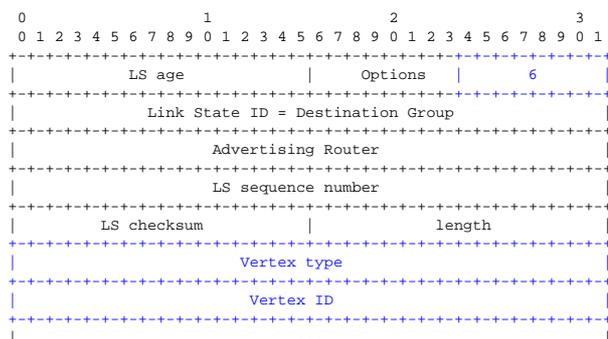
4.3. Les messages MOSPF

Format des paquets OSPF de type "Link-state update"



4.4. Le LSA de type "group-membership"

Code du LSA de type "group-membership" = 6



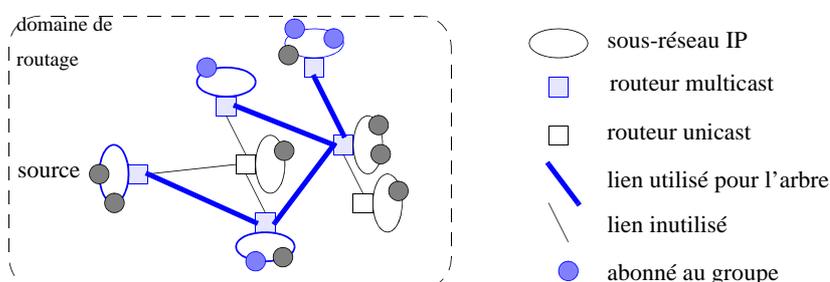
- l'entête standard d'un LSA : 20 octets
 - Link State ID = l'adresse multicast du groupe
- liste de sommets associés avec l'adresse multicast.
 - A chaque sommet est associé :
 - . Vertex type : 1 = router, 2 = transit network.
 - . Vertex ID : l'ID OSPF du routeur ou l'adresse IP du DR du réseau de transit
- un routeur émet un tel message ssi il est le routeur désigné d'un sous-réseau où une station est membre d'un groupe

4.5. Quelques raffinements

4.5.1 Hétérogénéité

Les zones de routage peuvent être hétérogènes :

- les routeurs multicasts ou non seront désignés par un bit spécifique du champ *option* des messages OSPF
- les chemins multicasts seront construits de telle sorte qu'ils ne passeront que par des routeurs multicasts

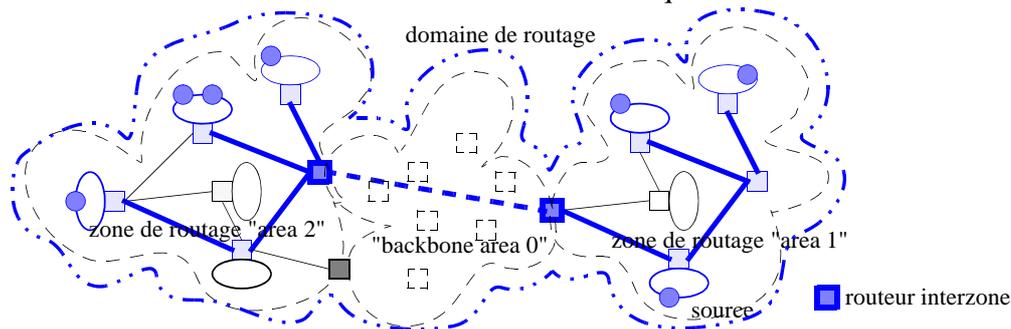


4.5.2 Zones de routage

Le domaine de routage est structuré en plusieurs zones de routage, interconnectées par une zone

centrale (de transit) appelée “backbone area”.

- Les routeurs interzones collectionnent tous les abonnés de leur zone
 - ils annoncent sur le “backbone area” les groupes actifs de leur zone
- Il peut y avoir plusieurs routeurs interzones pour une même zone
 - la métrique permet de n'utiliser que le routeur situé sur le meilleur chemin
 - un procédé permet de séparer les ex-aequo (celui de plus petite adresse)
- Un routeur interzone est toujours **noeud de l'arbre local** :
 - il reçoit une copie de tous les paquets multicasts émis localement, qu'il propage sur le “backbone”
- Un routeur interzone est **racine de l'arbre local** lorsque l'émetteur est externe



5. Le protocole PIM

5.1. Présentation

Si on envisage un réseau très large (l'Internet) **MOSPFP n'est pas adapté**

- le nombre de calculs croît avec le carré de la taille de la zone
- trop de diffusion de messages
- même si on lance les calculs “à la demande”, c-à-d. quand arrive le 1^{er} paquet de données
 - => le protocole PIM (“Protocol independent multicast”)

Deux versions du protocole PIM :

- PIM en mode dense (“dense mode”)
- PIM en mode clairsemé (“sparse mode”)
 - dépend de la probabilité de trouver un membre du groupe dans une zone

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

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

PIM en mode dense :

- implémente le routage RPF avec élagage (cf. S. Deering)
- ressemble à DVMRP mais n'utilise pas des tables de routage spécifiques :
 - on utilise les tables de routage du protocole de routage point-à-point
 - on fait l'hypothèse que les chemins sont symétriques :
 - => le chemin le plus court est le même à l'aller et au retour

Les 2 modes de PIM sont compatibles, ils partagent le même format de message

5.2. Le protocole PIM en mode clairsemé

Pour les groupes clairsemés :

- le nombre de AS membres du groupe est très petit devant le nombre total de AS dans l'Internet
- cependant l'étendue du groupe est très vaste

Un routeur est considéré comme un routeur terminal pour un groupe

- s'il est connecté directement à un réseau qui héberge au moins un membre du groupe

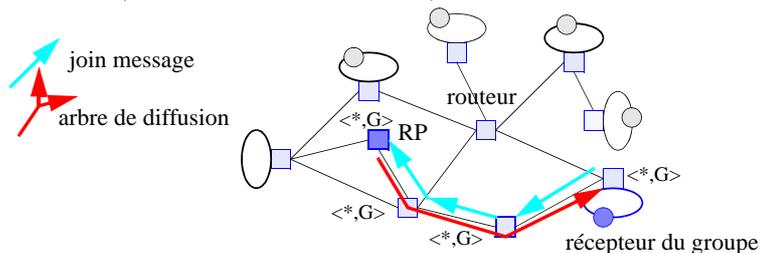
Protocole pouvant être déployé pour le routage entre domaines de routage (AS).

5.3. Principe de fonctionnement de PIM-SM

- Un routeur est choisi comme le **RP** (“rendez-vous point”) du groupe
 - les RP devraient être répartis judicieusement

5.3.1 Les récepteurs

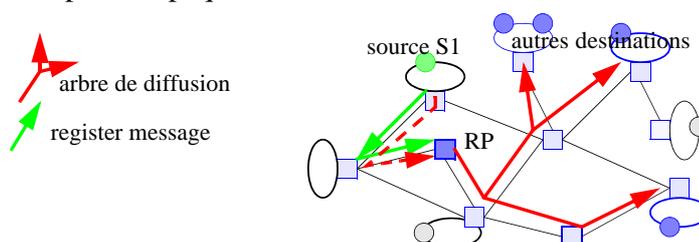
- Un routeur terminal ayant au moins un récepteur pour un groupe envoie périodiquement :
 - un message d'**inscription au groupe** (“**join message** (*,G)”)
 - vers le RP du groupe, en utilisant le chemin unicast le plus court
 - => le chemin suivi deviendra la branche de l'arbre de diffusion des paquets de données multicasts (utilisée en sens inverse !)



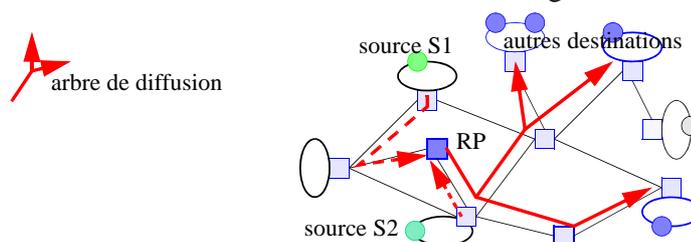
- Tout routeur le long du chemin du "join message" crée un état (*, G) dans sa table de routage multicast associant l'adresse du groupe à l'interface d'entrée et la liste des interfaces de sortie

5.3.2 Les émetteurs

- Lorsqu'une source veut diffuser un paquet de données vers un groupe :
 - le routeur d'extrémité, responsable de la source, envoie en unicast un message d'**enregistrement** (“**register message**”)
 - dans lequel il encapsule le paquet de données

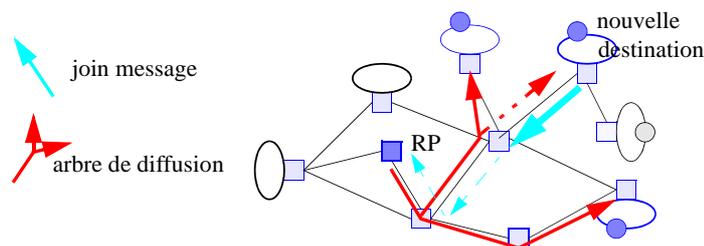


- Le RP, après désencapsulation, retransmet le paquet de données en le diffusant sur l'arbre
- Lorsqu'il y a plusieurs sources, leur flux de données convergent vers le RP



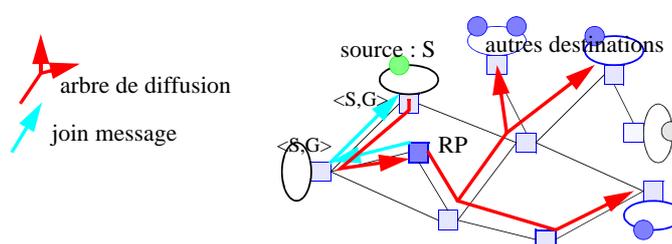
5.4. Optimisations

Les messages adressés au RP peuvent être traités par un routeur situé sur le chemin et qui fait déjà parti de l'arborescence du groupe



Si le flux de données en provenance d'une source S est important, le RP peut décider de contruire une branche spécifique vers cette source :

- émission d'un "join message (S,G)" et création d'un état par les routeurs intermédiaires



Jusqu'à maintenant,

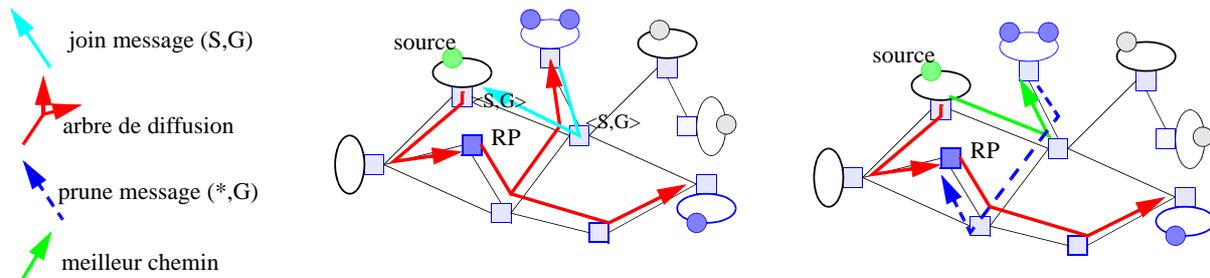
- Les paquets de données multicasts :
 - émis par un émetteur S pour un groupe G suivent
 - . sur la première partie du chemin entre S et le RP le plus court chemin inverse (de RP vers S)
 - . puis sur la deuxième partie du chemin entre le RP et un membre de G le plus court chemin inverse (du membre vers le RP)
 - utilisent un arbre de diffusion partiel ce qui minimise le nombre de duplications du paquet
- Les paquets de routage multicast:
 - sont en nombre limité puisque seuls les routeurs terminaux en émettent

Il persiste un inconvénient :

- Le chemin suivi par les données n'est pas le plus court
 - => le routeur terminal peut décider d'utiliser le meilleur chemin pour recevoir les paquets multicasts d'un groupe

Construction d'un meilleur chemin :

- le routeur terminal prévient l'émetteur
 - envoie un "join message (S,G)" directement à l'émetteur
- L'émetteur émet les paquets de données à la fois vers le RP et le routeur terminal
- Le routeur terminal dès qu'il reçoit des données provenant du meilleur chemin :
 - envoie une message d'élagage (*,G) vers le RP
 - => les paquets multicasts utilisent alors uniquement le chemin le plus court entre l'émetteur et le routeur terminal



Note : toutes les sources continuent d'émettre leurs données vers le RP et donc pour les autres destinations continuent d'utiliser l'arbre de diffusion

5.5. L'arrêt de la transmission de données

5.5.1 Le désabonnement d'un groupe

Un routeur terminal ayant plus de membre actif se désabonne d'un groupe G

- soit explicitement par l'envoi d'un message "prune (.,G)"
- soit implicitement en ne réémettant plus de message "join (.,G)"

L'état de routage d'un routeur disparaît

- implicitement lorsqu'il n'est plus rafraîchi
- explicitement lors de la réception de message "prune"

5.5.2 Le message "register-stop"

Ce message est envoyé par un RP vers une source lorsque

- le RP n'a pas de membre pour un groupe et qu'il reçoit des données de cette source pour ce groupe encapsulé dans un message "register"
- le RP reçoit de cette source pour ce groupe simultanément des données multicasts natives et des données encapsulées dans un message "register".

5.6. L'élection des RP

Les routeurs qui sont candidats à être RP, envoient un message au "Bootstrap Router" (BSR) :

- les messages PIM de "Candidate-RP-Advertisements" (message unicast)
- un routeur peut être candidat pour un certain sous-ensemble de groupes
- une priorité est associée à chaque routeur

Des messages sont envoyés périodiquement par le BSR de proche en proche à tous les routeurs du domaine PIM :

- les messages PIM de "Bootstrap"
- permet à l'ensemble des routeurs candidats à être RP d'être connu de tous les routeurs
- permet l'élection du routeur de bootstrap (BSR)
 - . le routeur avec la plus haute priorité et adresse est élu
 - . processus similaire au protocole d'élection du "spanning tree"

Au niveau de chaque routeur, une fonction de hachage associe une valeur pour chaque routeur candidat RP pour un groupe :

- celui choisi est celui de plus haute priorité avec la plus forte valeur

5.7. Les messages

5.7.1 Utilisation des messages PIM

Les messages PIM sont encapsulés dans des paquets IP :

- numéro du champ "protocol" d'IP : PIM = 103.

Les messages PIM sont transmis par

- soit unicast : Register, Register-Stop
- soit multicast hop-by-hop avec l'adresse "ALL-PIM-ROUTERS" = 224.0.0.13 : Join/Prune, Asserts, etc.

5.7.2 L'entête des messages PIM

```

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
+-----+-----+-----+-----+
| Ver=2 | Type | Reserved | Checksum |
+-----+-----+-----+-----+

```

PIM Types are:

- 0 = Hello
- 1 = Register
- 2 = Register-Stop
- 3 = Join/Prune
- 4 = Bootstrap
- 5 = Assert
- 6 = Graft (used in PIM-DM only)
- 7 = Graft-Ack (used in PIM-DM only)
- 8 = Candidate-RP-Advertisement

6. Le protocole BGMP

6.1. Présentation

Les protocoles précédents (DVMRP, MOSPF, PIM) sont inadaptés pour la transmission multicast entre plusieurs domaines de routage :

- DVMRP utilise la diffusion
- OSPF (donc MOSPF) perd son efficacité au-delà de 200 routeurs
- la localisation du RP de PIM peut être inappropriée :
 - domaine sans membre, domaine ayant une faible connectivité

6.1.1 BGMP

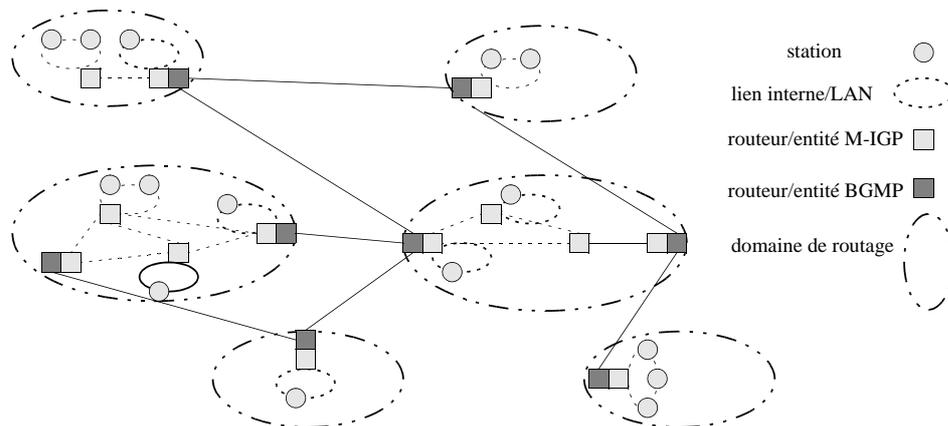
- Utilise un arbre bidirectionnel associé à chaque groupe, arbre qui relie les routeurs de bordure des domaines
- Le choix du RP dans PIM est aléatoire, mais dans BGMP le choix de la racine est soumis à la politique déterminée par l'administrateur :
 - en se basant sur le préfixe de l'adresse soit multicast : (*, prefix-G), soit de la source : (S-prefix, G).
 - Actuellement BGMP utilise des "unicast-prefix-based multicast address".
 - Le protocole MASC ([RFC 2909] : "Multicast Address-Set Claim") pourrait être utilisé.
- Pas d'état de routage pour un groupe multicast dans les zones du réseau où il n'y a pas de membres de ce groupe
- Compatible avec les protocoles de routage multicast intra-domaine (DVMRP, MOSPF, PIM, etc.)
- rfc 3913, "Border Gateway Multicast Protocol (BGMP): Protocol Specification." D. Thaler, September 2004.
- Attention à ne pas confondre avec :
 - MBGP : rfc 2858, "Multiprotocol Extensions for BGP-4". T. Bates, Y. Rekhter, R. Chandra, D. Katz. June 2000.

6.1.2 Localisation de la racine

Dans le cas d'utilisation d'adresses de type "unicast-prefix-based multicast", la racine de l'arbre associée à un groupe est déterminée par l'adresse du groupe.

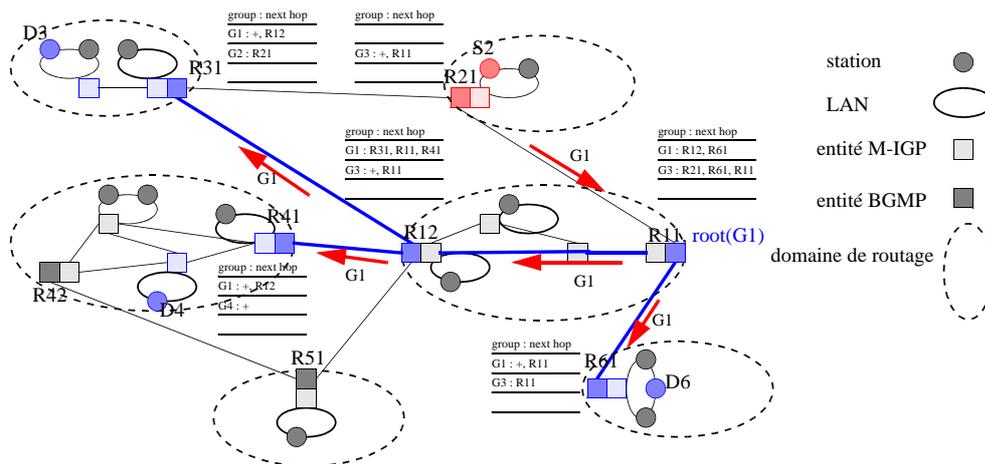
- Par exemple :
 - si l'adresse IPv6 de groupe est ff2e:0100:1234:5678:9abc:def0::xyz,
 - alors le préfixe unicast est : 1234:5678:9abc:def0/64,
 - et donc l'adresse de la racine devrait être : 1234:5678:9abc:def0::
 - . c'est l'adresse du routeur du sous-réseau correspondant au préfixe

6.2. Architecture

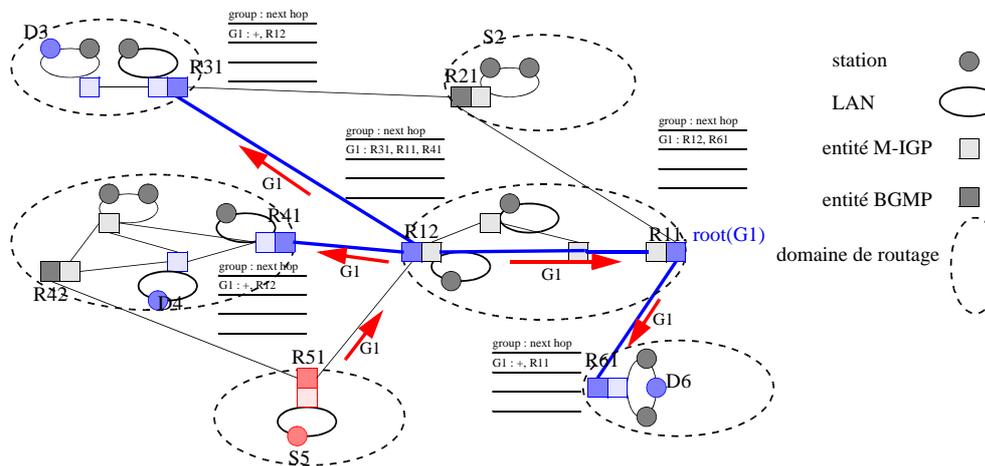


- pairs BGMP : "internal peer", "external peer"
- interconnexion par TCP : port 264 :
 - une transmission fiable :
 - . segmentation/réassemblage
 - . protection contre les erreurs et les pertes
 - . reséquencement

6.3. Transmission des paquets de données multicast



- les paquets de données multicast sont transmis vers la racine du groupe, puis suivent l'arbre bi-directionnel associé au groupe concerné
 - notation : le signe + signifie qu'il existe au moins un membre du groupe dans le domaine de routage local au pair BGMP. Le pair BGMP transmet le paquet au protocole M-IGP qui le transmet localement en conséquence.

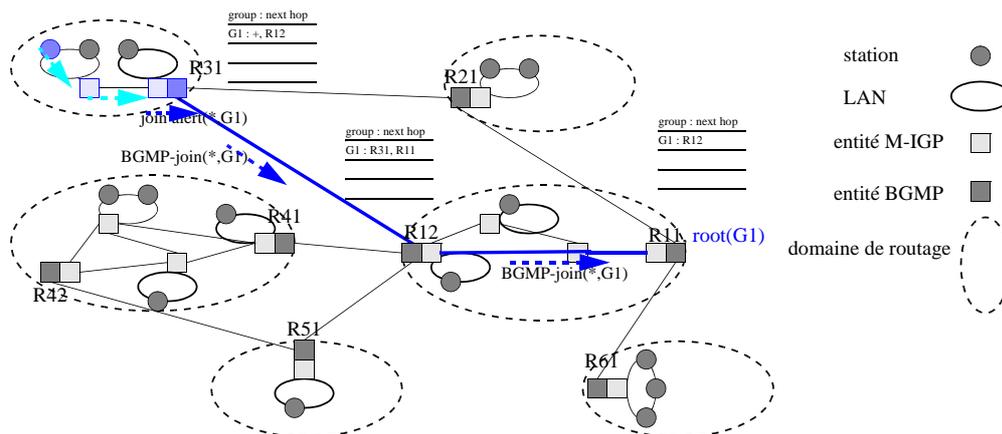


- Dès que le paquet de données multicast atteint l'arbre, il suit toutes ses branches.

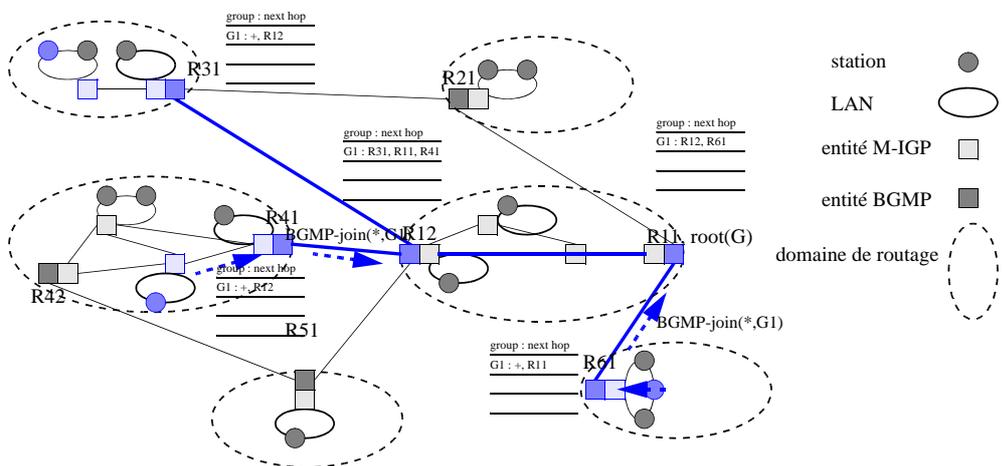
En résumé, lorsqu'un routeur reçoit un paquet de données multicast pour un groupe (S,G) ou (*,G)

- s'il possède une entrée associée soit à (*,G) soit à (S, G) alors le paquet multicast est routé en conséquence
- sinon il est routé vers la racine associée à G

6.4. Construction d'un arbre



- les messages BGMP-join sont transmis vers la racine du groupe multicast
- une entrée est créée dans la table de routage multicast <group, list of next-hop peers> dans les routeur BGMP le long de ce chemin.



- dès que le message BGMP-join arrive sur une BGMP-routeur qui possède une entrée pour le groupe concerné, la propagation du message s'arrête

En résumé, un routeur BGMP lorsqu'il reçoit un message BGMP-join

- s'il possède déjà une entrée pour le groupe, il se contente d'ajouter une branche
- sinon, il en crée une, la met à jour, puis retransmet le message vers la racine

6.5. Les messages BGMP

Gestion des arbres des groupes multicast

- BGMP-join(S, G)
 - des arbres spécifiques à une source peuvent être construits (SSM). Pour éviter les confusions, les arbres partagés ont priorité sur les arbres spécifiques à une source
- BGMP-prune(S, G)
 - permet d'élaguer les branches sans membre
- BGMP-poison-reverse(S,G)
 - notification de changement de route : l'ancienne route n'est plus bonne
- BGMP-forward-preference(S,G)
 - indique les préférences qu'à un pair concernant un groupe

Gestion de la connexion BGMP :

- "Open message" [1] : ouverture et paramétrisation de la connexion (version du protocole, identification des entités BGMP (@), authentification, temporisateur, etc.)
- "Update message " [2] : contient les messages de gestion des arbres multicasts, notamment "prune" ou "join"
- "Keepalive message" [4] : maintien de la connectivité TCP
- "Notification message" [3] : message d'erreur

7. Conclusion

Le service de multicast est un service utile.

Le procédé d'acheminement et l'adressage IP permet aisément à un paquet d'atteindre ses destinataires, pour peu que les tables de routage soient correctement configurées.

De nombreux protocoles de routage multicast sont chargés de cette mise à jour :

- local entre stations et routeurs terminaux : IGMP
- au sein d'un AS :
 - de type "distance vector" : DVMRP, PIM-DM; ou de type "link state" : MOSPF; ou de type arbre partagé : CBT, PIM-SM, etc.
- entre AS : BGMP ou MBGP.

Actuellement le réseau multicast se développe doucement.

Autres protocoles :

- de gestion des groupes
- de transport de données fiables (TCP est inadapté au multicast)