Le service IPv4 multicast pour les sites RAP

Description : Ce document présente le service IPv4 multicast pour les sites sur RAP
Version actuelle : 1.2
Date : 08/02/05
Auteurs : NM

Version	Dates	Remarques
1.0	19/12/03	Création du document
1.1	05/02/04	Correction du document
1.2	08/02/05	Correction du document

Table des matières

1 Introduction

Ce document rappelle les concepts multicast utiles à la compréhension du service IPv4 multicast de RAP.
Il décrit ensuite le multicast de RAP : le domaine PIM-SM de RAP et les principes de mise en œuvre du service multicast pour un site sur RAP.

2 Rappels sur les concepts multicast

2.1 Adresse multicast de niveau 2 (couche liaison)

Les spécifications IEEE 802.3 définissent un bit d'indication de l'adresse de trame multicast. Normalement une carte réseau (NIC) répond soit à son adresse MAC, soit à l'adresse de broadcast de niveau 2 (0xFFFF.FFFF.FFFF). Le groupe d'adresse 0100.5E00.0000 à 0100.5E7F.FFFF est réservé pour des adresses multicast de niveau 2. Derrière ces 25 bits de préfix (01:00:5E:0-7), restent 23 bits d'adresse MAC disponibles pour faire correspondre un groupe multicast de niveau 3 (exemple : 224.90.21.1) à une adresse multicast de niveau 2. On prendra les 23 bits de niveau faible de l'adresse de niveau 3 du groupe. Exemple sur l'adresse précédente : 01:00:5E:3A:15:01 (3a:15:01 correspondant à 90.21.1). Comme les adresses de groupe multicast de niveau 3 sont sur 28 bits (32-4 (1110)), les 5 bits de niveau haut de ces adresses sont mappés sur la même adresse multicast de niveau 2. Ces 32 groupes sont le résultat d'une part des 4 premiers bits de l'octet quatre de l'adresse IP (qui donnent 16 valeurs) faisant varier cette plage de 255 à 239, d'autre part du dernier bit de l'octet trois qui oscille entre 0 et 128. On a donc 2*16=32 groupes IP qui correspondent au même groupe de niveau 2.

2.2 Adresse multicast de niveau 3 (couche réseau)

Les adresses de classe D (qui commencent par 1110) suivantes sont utilisées pour le multicast : 224.0.0.0 – 239.255.255.255. Ces adresses ne sont utilisées que pour des adresses de groupes, adresses de destination du trafic multicast. L'adresse source d'un flux multicast est quant à elle toujours une adresse unicast.
La plage 224.0.0.0 à 224.0.0.255 est réservée pour les protocoles réseaux sur le segment local, les paquets ayant cette adresse de destination ne seront jamais transmis par un routeur car ils partent avec un TTL de 1.
Par exemples :

- 224.0.0.1 : tous les systèmes multicast sur le LAN
- 224.0.0.2 : tous les routeurs multicast sur le LAN
- 224.0.0.13 : tous les routeurs PIM version 2 sur le LAN

Les adresses de 239.0.0.0 à 239.255.255.255 sont réservées pour des diffusions à portée limitée (RFC 2365).
Voir http://www.iana.org/assignments/multicast-addresses pour une liste exhaustive.

2.3 Adressage GLOP

Il permet de définir un sous réseau globalement réservé à un AS. Pour cela, la RFC 2770 propose l'utilisation des réseaux 233.0.0.0/8 pour lesquels on ajoute aux octets deux et trois son numéro d'AS. Par exemple pour l'AS de RAP (AS 2422, en hexa 0976, 09 = 9 et 76 = 118) on obtient 233.9.118.0/24 qui est globalement réservé pour l'AS de RAP.
Voir http://gigapop.uoregon.edu/glop ou http://www.shepfarm.com/multicast/glop.html pour la correspondance entre le numéro d'AS et la plage d'adresse multicast réservée.

2.4 IGMP (Internet Group Management Protocol)

Au niveau 3 sur le LAN, IGMPv2 (protocole id=2 véhiculé dans de l'IP) permet aux stations du segment local de faire connaître au routeur local multicast leur demande d'enregistrement ou de retrait de participation aux groupes multicast.
Les routeurs envoient périodiquement (toutes les minutes) un *IGMP Host Membership Query* (Type 0x11) sur le groupe 224.0.0.1 avec un TTL de 1. Tous les systèmes capables de traiter le multicast reçoivent cette « sonde » et afin qu'ils ne répondent pas tous en même temps, chacun lance un compteur aléatoire pour chaque groupe auprès duquel il est enregistré. Une fois le compteur arrivé à terme, la station envoie un *IGMP Host Membership Report* (TTL 1 Type 0x12) à l'adresse multicast du groupe auquel il participe et pour lequel le compteur vient de se terminer. Comme le rapport est envoyé au groupe multicast, les autres stations du segment éventuellement enregistrées sur le même groupe recevront ce rapport et pourront ainsi abandonner le compteur et annuler leur intention d'envoyer un même rapport au routeur local. Ce dernier continuera de transmettre le flux multicast d'un groupe dès l'instant qu'il y a au moins une réponse d'une station concernant le groupe. Il n'a pas besoin de savoir combien de stations demandent ce flux ni lesquelles. Quand une station quitte un groupe, elle envoie un *IGMP Leave Group* (Type 0x17) envoyé au groupe 224.0.0.2 et non au groupe pour lequel il ne désire plus recevoir de flux, ceci afin d'éviter d'encombrer quelques minutes de plus la bande passante par le flux d'un groupe pour lequel il n'y a plus de stations intéressées. En effet, un routeur arrête de transmettre le flux d'un groupe seulement s'il ne reçoit pas de réponse après 3 IGMP *Host Membership Query* consécutifs.

2.5 Arbres de distribution

2.5.1 Arbre Source (« Source Tree »)

L'arbre source est un arbre du plus court chemin (« Shortest Path Tree ») dont la racine est la source et les nœuds sont les routeurs multicast sur le trajet entre la source et les récepteurs. Chaque routeur créé un enregistrement (S,G) concernant le flux en provenance de la source S et à destination du groupe G. Il y a aura autant d'enregistrements (S,G) et de SPT, dans les routeurs qu'il y a de sources (émetteurs) pour le groupe.

2.5.2 Arbre Partagé (« Shared Tree »)

Contrairement aux arbres source dont la racine est la source, les arbres partagés ont une racine centralisée, partagée, un point de rendez-vous (RP). Les sources envoient leur flux à la racine (le RP), depuis laquelle le trafic est transmis vers les récepteurs. Comme toutes les sources utilisent un arbre partagé commun, on retrouvera dans les routes la notation (*,G). L'arbre de distribution est dynamique, des branches seront crées ou coupées en fonction des abonnements/désabonnements de récepteurs. L'implémentation fera en sorte après utilisation de cet arbre partagé, de permettre de rejoindre un arbre SPT afin d'optimiser le chemin entre une source et une destination. Ce qui est important, c'est que le RP doit être informé de la présence d'une source et d'un groupe de façon à pouvoir annoncer cette source à d'autres domaines multicast via MSDP.

2.6 Seuil (« Threshold »)

Le champ TTL en multicast a deux fonctions : décrémentation de 1 à chaque passage d'un routeur et vérification de seuil si le TTL est inférieur à la valeur du seuil multicast de l'interface (*ip multicast ttl-threshold 16* par exemple) alors le paquet n'est pas transmis. Cependant les adresses privées et locales, 239.0.0.0 à 239.255.255.255, n'utilisent pas le TTL pour définir la portée, mais leur zone d'adresse.

2.7 PIM-SM (Protocol Independent Multicast - Sparse Mode)

PIM se base sur les tables de routage existantes (OSPF, RIP, statique, …).
En Sparse Mode, seuls les récepteurs ayant fait un *Join* sur un groupe feront partie de l'arbre de distribution. Quand un récepteur rejoint un groupe multicast son routeur émet un *PIM Join* vers le RP. Le routeur construit un arbre de distribution partagé, les sources envoient au RP et les récepteurs font une demande explicite auprès de ce RP pour accéder à la source. Une fois l'arbre partagé établi, le récepteur peut faire remonter une demande de SPT qui en passant outre le RP peut trouver le chemin le plus court entre une source particulière et lui-même.

2.8 RPF (Reverse Path Forwarding)

RPF est l'algorithme qui permet de construire les arbres de distribution. Le contrôle RPF utilise les tables de routage existantes pour déterminer l'interface par laquelle doit arriver les paquets multicast. Le contrôle RPF réussit si le routeur reçoit le paquet par l'interface qui amène par le plus court chemin vers la source. Dans ce cas, le paquet est transmis sur les interfaces de sortie présentes dans la table de routage autrement il est détruit afin d'éviter les boucles.

2.9 RP (Rendez-vous Point)

Une source prévient de son existence à un RP et un récepteur vient chercher l'existence de nouvelles sources auprès du RP. Un routeur PIM peut être un RP pour plusieurs groupes. Un groupe est associé à un seul RP qui peut résulter d'une élection entre plusieurs candidats RP.

2.10 BSR (BootStrap Router)

Le BSR permet de minimiser les tâches de configuration des routeurs PIM. Cette technique ne fonctionne qu'en PIM version 2. Elle permet aux routeurs de prendre connaissance du ou des RP présents le domaine PIM-SM et cela automatiquement et dynamiquement.
Dans cette méthode, le RP pour un groupe multicast est issu d'une élection parmi plusieurs candidats RP. Tous les candidats à la fonction de RP pour ce groupe multicast font acte de candidature auprès du BSR par des messages *Advertissements*. Le BSR va collecter les informations en provenance des candidats RP pour les différents groupes multicast puis il diffuse en direction de tous les routeurs PIM un message *BSR* qui contient la liste des candidats RP. Pour un groupe multicast, chaque routeur PIM du domaine exécute l'algorithme du BootStrap pour calculer le RP choisi parmi l'ensemble des candidats RP pour ce groupe multicast donné.

2.11 BR (Border Router)

Les BR sont des routeurs frontières du domaine PIM dont le rôle est d'empêcher d'acheminer les messages BSR donc les informations sur la connaissance des candidats RP.

2.12 MSDP (Multicast Source Discovery Protocol)

Les RP ne peuvent pas par défaut communiquer entre différents domaines. Les routeurs PIM SM communiquent entre différents domaines PIM-SM par MSDP pour trouver les sources actives multicast qui se trouvent dans d'autres domaines. Ce protocole transfère donc des informations (S,G) entre RP de différents domaines PIM-SM. Ces informations sont véhiculées par des *MSDP SA-messages*, les RP faisant l'échange sont appelés des *MSDP peers*. Quand une source émet dans un domaine PIM-SM, le PIM « Designated Router » directement connecté à cette source envoie un *PIM Register message* au RP du domaine. Ce dernier construit un message « Source-Active » (SA) et l'envoie à son MSDP peer. Ce message parcourt l'arbre multicast entre MSDP peers suivant une méthode de « peer-RPF

flooding ». Les RP cachent localement les messages SA. Tant que la source émet, le RP envoie toutes les 60 secondes les message SA. Les messages MSDP sont encapsulés dans une connexion TCP. Le peer MSDP qui a la plus grande adresse IP sera le « listener » et écoutera sur le port TCP 639.

2.13 MBGP (Multiprotocol BGP)

MBGP permet de distinguer quels préfixes de route sont utilisés pour réaliser la vérification du RPF. L'avantage du protocole est de permettre d'avoir des topologies unicast différentes de celles multicast, ou en cas de topologies identiques d'appliquer des politiques différentes dans les deux mondes.

2.14 SDR (Session DiRectory)

SDR est une application multicast qui utilise le protocole SDP/SAP afin de disséminer les noms et les propriétés des sessions de diffusion via la session ayant pour groupe 224.2.127.254 pour les sessions à portée globale et 239.255.255.255 pour les sessions à portée locale.
SDP décrit les sessions à annoncer.
SAP envoie périodiquement un paquet multicast aux groupes définis ci-dessus suivant la portée. L'annonce est faite avec la même portée que celui de la session. L'intervalle des annonces est fonction de la taille de l'annonce, de la fréquence des annonces précédentes (ou de l'horloge à la première annonce) ainsi que de la bande passante du réseau. Plus exactement, l'intervalle est égal à la valeur maximum entre la valeur 300 et la valeur 8*nbre d'annonce*taille/BP.

3 Le multicast sur RAP

3.1 Introduction

Le service multicast de RAP met en oeuvre le protocole PIM-SM version 2.
C'est aussi le protocole de routage multicast mis en oeuvre par RENATER.
Il y a 2 modes de raccordement multicast possibles pour un site sur RAP :

- Le raccordement multicast simplifié pour un site est celui où le site ne gère pas de domaine PIM-SM interne. Le site s'intègre dans le domaine PIM-SM de RAP. MSDP et les fonctions de « Border Router » ne sont pas mis en œuvre entre RAP et le site.
- Le raccordement multicast préconisé pour un site est celui où le site gère un service multicast en interne avec les adresses de groupe 239.x.x.x. Le site a son propre domaine PIM-SM. MSDP et les fonctions de « Border Router » sont mis en œuvre entre RAP et le site.

MBGP n'est pas mis en oeuvre entre RAP et le site sauf si le site fait déjà du BGP avec RAP et qu'il souhaite avoir plus d'une route multicast vers l'extérieur, l'une avec RAP et au moins une autre avec un autre AS. Les annonces de réseaux pour le multicast (NLRI de type 2 annoncés en MBGP) ne sont pas détaillées dans ce document, les sites souhaitant disposer de multihoming BGP pour le multicast doivent contacter CORAP pour les modalités spécifiques de mise en œuvre de ce service.

3.2 Le domaine PIM de RAP

La figure suivante montre l'architecture du domaine PIM-SM de RAP et le rôle de chaque composante.

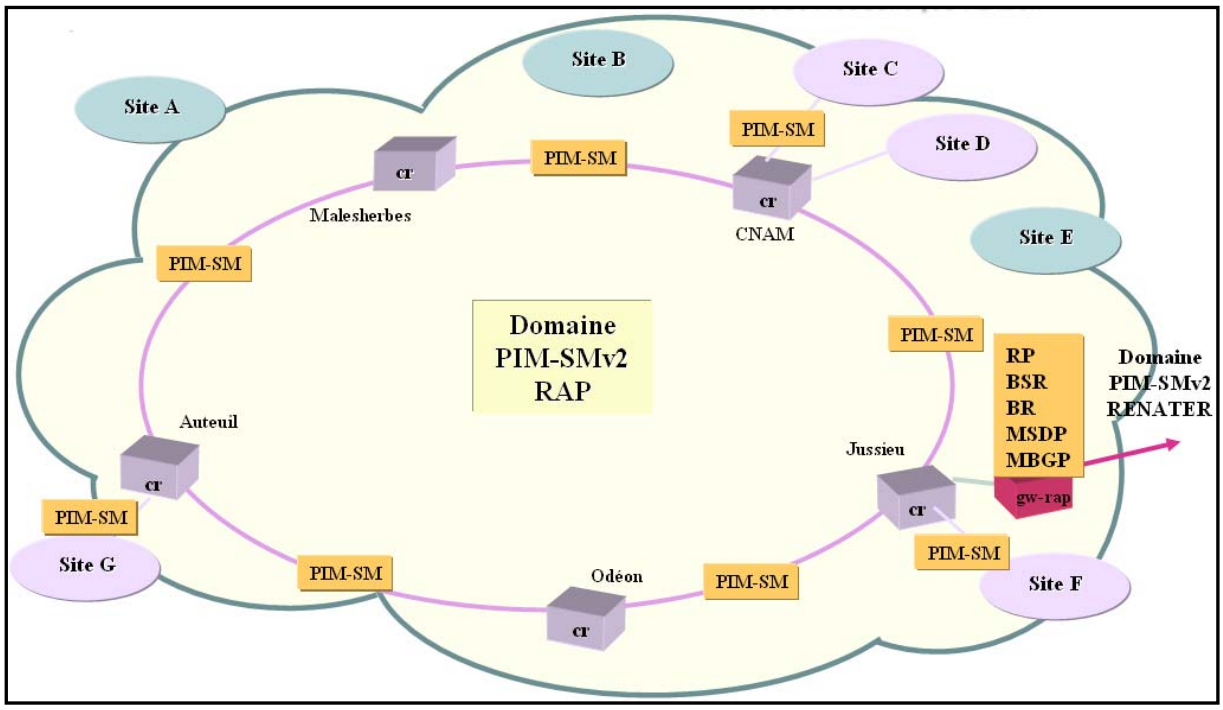

Figure 1 – Domaine PIM-SM de RAP

3.2.1 PIM-SM

Le protocole PIM-SM version 2 est activé sur toutes les interfaces des équipements du cœur de RAP, entre les commutateurs-routeurs. Il est aussi activé sur les interfaces avec les sites mais uniquement pour les sites qui en ont fait la demande.

3.2.2 RP

GW-RAP est défini comme l'unique RP du domaine PIM de RAP pour tous les groupes multicast en dehors des groupes privés (adresses 224.0.0.0 à 239.255.255.255).
Le CR à Odéon est défini comme le RP du domaine PIM de RAP pour les groupes privés permettant d'avoir un service multicast interne à RAP. Ces adresses de groupe sont réservées pour un usage local et privé et ne sont pas routés en dehors du domaine PIM de RAP.

3.2.3 BSR

GW-RAP est défini comme unique BSR pour le domaine PIM de RAP.

3.2.4 BR

Afin de délimiter la frontière du domaine PIM de RAP, l'interface entre le routeur de peering de RAP et le routeur RENATER face à lui est déclarée *ip pim border*. Au-delà de cette frontière, GW-RAP n'envoie jamais de message BSR. Le domaine PIM de RENATER utilise son propre réseau de RP et le RP de RAP reste local à son domaine.

3.2.5 MSDP

Pour que les récepteurs du domaine PIM de RAP puissent recevoir les sources des domaines extérieurs, un peering MSDP est établi entre le routeur GW-RAP et le routeur RENATER face à lui. Un filtrage de SA est mis en œuvre sur le peering MSDP pour le contrôle de la cohérence des annonces de SA. Les SA ayant comme sources des adresses RFC 1918 sont refusés, de même que les SA portant sur des groupes de type locaux (239.0.0.0/8), SSM (232.0.0.0/8) réservés à RAP.

3.2.6 MBGP

Un peering MBGP est établi avec le routeur GW-RAP et le routeur RENATER face à lui. MBGP est le protocole de routage inter-domaine qui permet au domaine PIM de RAP de prendre connaissance des routes des sources actives des autres domaines, et aux autres domaines PIM de l'Internet de prendre connaissance des routes des sources actives du domaine PIM de RAP.
MBGP permet d'utiliser des liens différents pour le trafic unicast et multicast. Dans le cas de RAP, la topologie est congruante, c'est-à-dire que le peering pour le routage multicast est le même que le peering pour le routage unicast. Le lien BGP utilisé pour le trafic unicast et multicast avec RENATER est une liaison Gigabit Ethernet.

3.3 Raccordement des sites

3.3.1 Raccordement multicast simplifié

C'est une manière simple et rapide pour un site de se raccorder au multicast de RAP.
Le site ne gère pas de domaine PIM-SM interne. Il n'y a pas de RP local sur le site et le RP de RAP formate les messages PIM pour l'enregistrement ou l'abonnement des stations du site sur les groupes multicast. Le protocole MSDP et les fonctions de Border Router ne sont pas mis en œuvre entre RAP et le site.
Il suffit que le routeur IP du site face à RAP supporte PIM-SM version 2. PIM-SM est activé sur l'interface de RAP face au site à la demande du site.
Le site ne doit pas déclarer l'adresse du RP sur ses routeurs PIM puisque le mécanisme de BootStrap Router de PIM version 2 permet aux routeurs d'apprendre dynamiquement et automatiquement que le RP est GW-RAP pour les groupes 224.0.0.0 à 238.255.255.255.
Les routeurs du site dont le multicast est actif ne doivent en aucun cas être déclarés candidat BSR ni candidat RP.
Dans le site, PIM-SM doit être activé sur chacune des interfaces multicast et IGMP doit être disponible sur chaque poste de travail multicast.
La validation du service peut se faire avec des outils qui génèrent du trafic multicast (mcast, mcaster) ou des applications de diffusions vidéo (l'application VLC ou les outils du MBone).

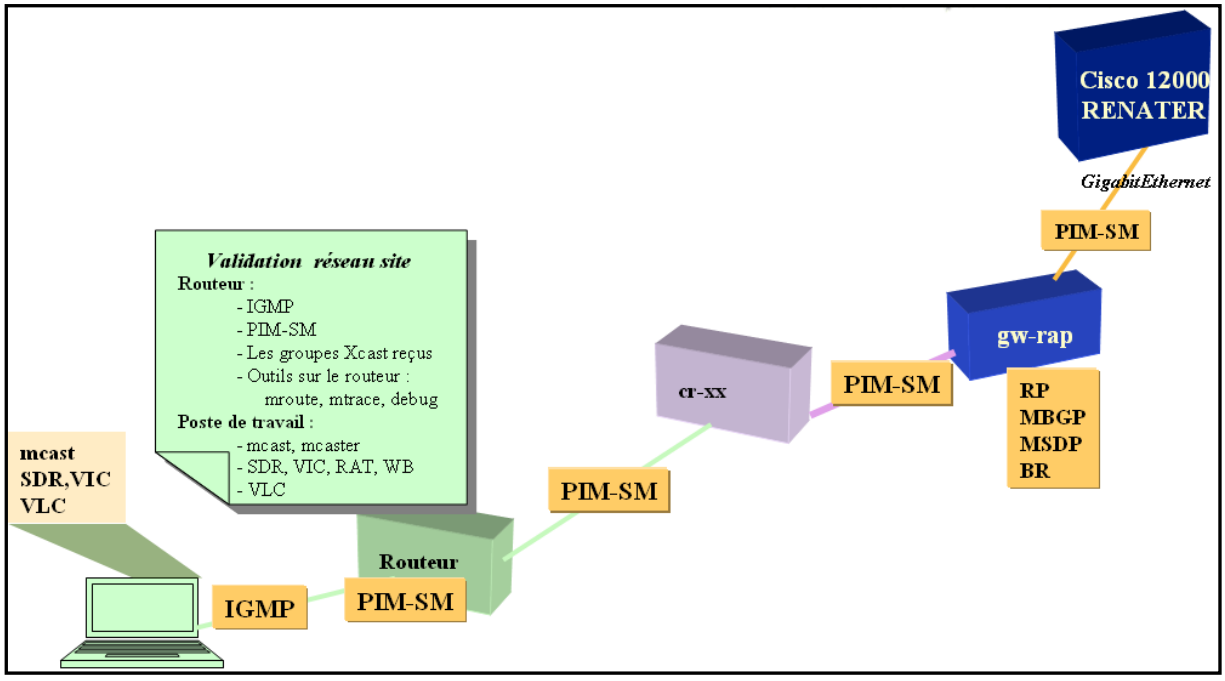

Figure 2 – Raccordement muticast simplifié

3.3.2 Raccordement multicast préconisé

Le site gère son propre domaine PIM-SM et utilise son propre réseau de RP.
Au moins un routeur du site doit être déclaré RP pour tout ou une partie du trafic multicast.
Le RP du site reste local à son domaine et est indépendant de celui de RAP pour l'accès à ses sources multicast.
Ce mode de raccordement permet au site de gérer en interne un service multicast avec ses groupes privées et locaux (239.0.0.0 à 239.255.255.255).
Le routeur IP du site face à RAP doit bien évidemment supporter PIM-SM version 2.
PIM-SM version 2 doit être activé de chaque côté de l'interface entre le routeur de RAP et le routeur du site face à lui.
Dans le site, PIM-SM doit être activé sur chacune des interfaces multicast et IGMP doit être disponible sur chaque poste de travail multicast.
Pour limiter la frontière entre les 2 domaines PIM (celui de RAP et celui du site), le routeur du site face à RAP doit supporter la fonction « Border Router ». La fonction Border Router n'est pas supportée par les Black Diamond 6808 (les CR), aussi il n'est pas possible de définir la frontière de domaine PIM du côté de RAP, elle doit donc l'être obligatoirement du côté du site, sur l'interface avec RAP. Si cela n'est pas fait, des messages BSR pourraient s'échangés entre les 2 domaines.
MSDP doit être mis en œuvre entre un routeur du site e
GW-RAP pour que les échanges des message SA puissent se faire entre les 2 domaines PIM.
La validation du service peut se faire avec des outils qui génèrent du trafic multicast (mcast, mcaster) ou des applications de diffusions vidéo (l'application VLC ou les outils du MBone).

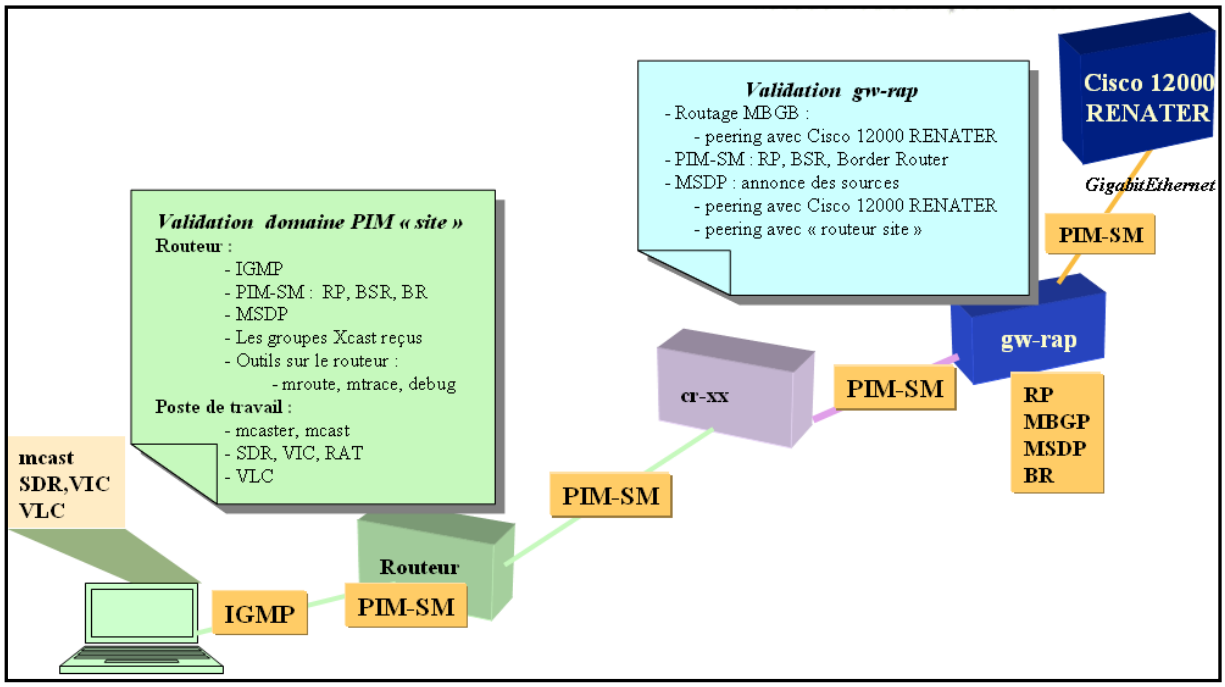

Figure 3 – Raccordement multicast préconisé

3.4 Accès au service

Tout site de RAP peut accéder au multicast. Il s'agit d'un accès multicast IPv4.
L'accès au service suppose que le site dispose du matériel nécessaire avec le niveau de logiciel adéquat : support du routage multicast PIM-SM version 2 pour l'accès au multicast natif et support du protocole MSDP et de la fonction Border Router pour la création d'un point de rendez vous local au site.
Le site doit faire une demande de service auprès de rap-ds@rap.prd.fr.

RAP peut attribuer des adresses de groupe multicast aux sites qui n'en disposent pas.
Pour cela il suffit de faire une demande écrite à rap-ds@rap.prd.fr.

3.5 Métrologie

RAP maintient une matrice site à site de diverses statistiques d'état de performances sur le multicast qui s'appuie sur le logiciel Beacon (http://dast.nlanr.net/projects/beacon). Beacon est une sonde logicielle permettant de surveiller l'ensemble de la qualité du multicast et de détecter les incidents. C'est un outil essentiel pour le bon fonctionnement du multicast.
Pour une meilleure gestion du service, CORAP demande aux sites d'installer un client beacon sur un poste multicast, de l'interfacer au serveur Beacon de RAP et de le faire fonctionner en permanence. Pour s'intégrer dans la grille de métrologie multicast de RAP, veuillez consulter http://www.rap.prd.fr/services/supervisionXcast.php

4 Acronymes

AS : Autonomous System
BGP : Border Gateway Protocol
BR : Border Router
BSR : BootStrap Router
IGMP : Internet Group Management Protocol
MAC : Media Access Control
MBGP : Multiprotocol BGP
MSDP : Multicast Source Discovery Protocol
NIC : Network Interface Card
NLRI : Network Layer Reachable Information
OSPF : Open Shortest Path First
PIM : Protocol Independent Multicast
PIM-SM : Protocol Independent Multicast - Sparse Mode
RAP : Réseau Académique Parisien
RIP : Routing Information Protocol
RP : Rendez-vous Point
RPF : Reverse Path Forwarding
SA : Source Active
SAP : Session Announce Protocol
SDP : Session Description Protocol
SDR : Session DiRectory
SPT : Shortest Path Tree
VLC : VideoLAN Client