Groupe de travail Réseau
Request for Comments : 5015
Catégorie : Sur la voie de la normalisation

Traduction Claude Brière de L'Isle

M. Handley, UCL
I. Kouvelas, Cisco
T. Speakman, Cisco
L. Vicisano, Digital Fountain
octobre 2007

Diffusion groupée bidirectionnelle indépendante du protocole (BIDR-PIM)

Statut du présent mémoire
Le présent document spécifie un protocole de l'Internet sur la voie de la normalisation pour la communauté de l'Internet, et appelle à des discussions et suggestions pour son amélioration. Prière de se référer à l'édition en cours des "Protocoles officiels de l'Internet" (STD 1) pour voir l'état de normalisation et le statut de ce protocole. La distribution du présent mémoire n'est soumise à aucune restriction.
(La présente traduction incorpore l'erratum 3762)

Résumé
Le présent document discute de PIM bidirectionnel (BIDIR-PIM, *Bidirectional PIM*) une variante de PIM en mode épars qui s'appuie sur des arborescences bidirectionnelles partagées connectant plusieurs sources et receveurs. Les arborescences bidirectionnelles sont construites en utilisant un mécanisme d'élection résistante à l'échec d'un transmetteur désigné (DF, *Designated forwarder*) qui fonctionne sur chaque liaison d'une topologie de diffusion groupée. Avec l'aide du DF, les données en diffusion groupée sont naturellement transmises des sources au point de rendez-vous (RP, *Rendezvous-Point*) et donc le long de l'arborescence partagée jusqu'aux receveurs sans exiger d'état spécifique de la source. L'élection du DF a lieu au moment de la découverte du RP et donne le chemin au RP, éliminant donc l'exigence d'événements de protocole s'appuyant sur les données.

Table des Matières

1. Introduction

Le présent document spécifie PIM bidirectionnel (BIDIR-PIM) une variante de PIM en mode épars (PIM-SM) [RFC4601] qui construit des arborescences partagées bidirectionnelles connectant des sources et des receveurs de diffusion groupée.

PIM-SM construit des arborescences partagées unidirectionnelles qui sont utilisées pour transmettre des données des envoyeurs aux receveurs d'un groupe de diffusion groupée. PIM-SM permet aussi la construction d'arborescences spécifiques de source, mais cette capacité est sans rapport avec le protocole décrit dans le présent document.

L'arborescence partagée pour chaque groupe de diffusion groupée a sa racine à un routeur de diffusion groupée appelé le point de rendez-vous (RP). Différents groupes de diffusion groupée peuvent utiliser des RP séparés au sein d'un domaine PIM.

Dans PIM-SM unidirectionnel, il y a deux méthodes possibles pour distribuer les paquets de données sur l'arborescence partagée. Elles diffèrent par la façon dont les paquets sont transmis d'une source au RP :

- Initialement, quand une source commence à transmettre, son routeur de premier bond encapsule les paquets de données dans des messages de contrôle spéciaux (Registers) qui sont en envoi individuel au RP. Après avoir atteint le RP, les paquets sont désencapsulés et distribués sur l'arborescence partagée.

- Une transition à partir du mode de distribution ci-dessus peut être faite à un stade ultérieur. Ceci est réalisé en construisant un état spécifique de la source sur tous les routeurs le long du chemin entre la source et le RP. Cet état est alors utilisé pour transmettre nativement les paquets à partir de cette source.

Ces deux mécanismes connaissent des problèmes. L'encapsulation résulte en un traitement, une consommation de bande passante, et de frais généraux de délais, significatifs. La transmission utilisant l'état spécifique de source a des exigences de protocole et de mémoire supplémentaires.

PIM bidirectionnel dispense de l'encapsulation et de l'état de source en permettant que les paquets soient nativement transmis d'une source au RP en utilisant l'état d'arborescence partagée. À la différence de PIM-SM, ce mode de transmission n'exige aucun événement piloté par les données.

La spécification de protocole du présent document suppose la familiarité avec la spécification PIM-SM [RFC4601]. Des portions du fonctionnement du protocole BIDIR-PIM qui sont identiques à celui de PIM-SM sont seulement définies par référence.

2. Terminologie

Les mots clés "DOIT", "NE DOIT PAS", "EXIGE", "DEVRA", "NE DEVRA PAS", "DEVRAIT", "NE DEVRAIT PAS", "RECOMMANDE", "PEUT", et "FACULTATIF" en majuscules dans ce document sont à interpréter comme décrit dans la [RFC2119] et indiquent les niveaux d'exigence pour les mises en œuvre conformes de BIDIR-PIM.

2.1 Définitions

La présente spécification utilise un certain nombre de termes pour se référer aux rôles des routeurs qui participent à BIDIR-PIM. Les termes suivants ont une signification particulière pour BIDIR-PIM :

Base de données d'informations d'acheminement de diffusion groupée (MRIB, *Multicast Routing Information Base*) : tableau de la topologie de diffusion groupée, qui est normalement déduite du tableau d'acheminement en envoi individuel, ou de protocoles d'acheminement tels que BGP multiprotocoles (MBGP) [RFC4760] qui portent les informations de topologie spécifiques de la diffusion groupée. Elle est utilisée par PIM pour établir l'interface de RPF (utilisée dans les règles de transmission). Dans PIM-SM, la MRIB est aussi utilisée pour prendre des décisions sur où transmettre les messages Join/Prune, tandis que dans BIDIR-PIM, elle est utilisée comme source des métriques d'acheminement pour le processus d'élection de DF.

Adresse de point de rendez-vous (RPA) : une RPA est une adresse utilisée comme racine de l'arborescence de distribution pour une gamme de groupes de diffusion groupée. La RPA doit être acheminable à partir de tous les routeurs dans le domaine PIM. La RPA n'a pas besoin de correspondre à une adresse d'une interface d'un routeur réel. À cet égard,

BIDIR-PIM diffère de PIM-SM, qui exige qu'un routeur réel soit configuré comme point de rendez-vous (RP). Les messages joints des receveurs pour un groupe BIDIR-PIM se propagent bond par bond vers la RPA.

Liaison de point de rendez-vous (RPL) : une RPL pour une RPA particulière est la liaison physique à laquelle appartient la RPA. Dans BIDIR-PIM, tout le trafic en diffusion groupée aux groupes qui se transposent en une RPA spécifique est transmis sur la RPL de cette RPA. La RPL est particulière dans un domaine BIDIR-PIM car elle est la seule liaison sur laquelle une élection de transmetteur désigné n'a pas lieu (voir la définition de DF ci-dessous).

En amont : vers la racine (RPA) de l'arborescence. Direction utilisée par les paquets voyageant des sources à la RPL.

En aval : en s'éloignant de la racine de l'arborescence. Direction dans laquelle les paquets voyagent de la RPL aux receveurs.

Transmetteur désigné (DF, *Designated Forwarder*) : le protocole présenté dans ce document est largement fondé sur le concept d'un transmetteur désigné (DF). Un seul DF existe pour chaque RPA sur chaque liaison au sein d'un domaine BIDIR-PIM (cela inclut des liaisons multi-accès et point à point). La seule exception est la RPL sur laquelle aucun DF n'existe. Le DF est le routeur sur la liaison qui a le meilleur chemin pour la RPA (déterminé en comparant les métriques fournies par la MRIB). Un DF pour une RPA donnée est chargé de transmettre le trafic en aval sur sa liaison, et de transmettre le trafic en amont de sa liaison vers la RPL. Il fait cela pour tous les groupes bidirectionnels qui se trransposent sur la RPA. Le DF sur une liaison est aussi responsable du traitement des messages Join provenant des routeurs en aval sur la liaison ainsi que de s'assurer que les paquets sont transmis aux receveurs locaux (découverts par un mécanisme d'adhésion locale comme MLD [RFC2710] ou IGMP [RFC3376]).

Interface de transmission sur le chemin inverse (RPF, *Reverse Path Forwarding)* : l'interface RPF d'un routeur par rapport à une adresse est l'interface que la MRIB indique comme devant être utilisé pour atteindre cette adresse. Dans le cas d'un groupe de diffusion groupée BIDIR-PIM, l'interface RPF est déterminée en cherchant la RPA dans la MRIB. Les informations de la RPF déterminent l'interface du routeur qui va être utilisée pour envoyer des paquets vers la RPL pour le groupe.

Voisin de RPF : le voisin de RPF d'un routeur par rapport à une adresse est le voisin que la MRIB indique comme devant être utilisé pour atteindre cette adresse. Noter que dans BIDIR-PIM, le voisin de RPF pour un groupe n'est pas nécessairement le routeur sur l'interface de RPF où les messages Join pour ce groupe vont être dirigés (les messages Join sont seulement dirigés sur le DF sur l'interface RPF pour le groupe).

Base de données d'information d'arborescence (TIB, *Tree Information Base*) : c'est la collection des états à un routeur PIM qui ont été créés en recevant des messages PIM Join/Prune, des messages d'élection de DF PIM, et des informations IGMP ou MLD de la part des hôtes locaux. Elle mémorise essentiellement l'état de toutes les arborescences de distribution de diffusion groupée à ce routeur.

Base de données d'information de transmission en diffusion groupée (MFIB, *Multicast Forwarding Information Base*) : la TIB contient tout l'état nécessaire pour transmettre les paquets en diffusion groupée à un routeur. Cependant, bien que la présente spécification définisse la transmission en termes de TIB, transmettre réellement des paquets en utilisant la TIB est très inefficace. Une mise en œuvre de routeur réel va plutôt normalement construire une MFIB efficace à partir de l'état de la TIB pour effectuer la transmission. Comment cela est fait est spécifique de la mise en œuvre, et n'est pas discuté dans ce document.

2.2 Notation en pseudo code

On utilise la notation des ensembles en plusieurs endroits de cette spécification.

A (+) B est l'union de deux ensembles, A et B.

A (-) B sont les éléments de l'ensemble A qui ne sont pas dans l'ensemble B.

NULL est l'ensemble ou liste vide.

De plus, on utilise une syntaxe de style C :

= note l'allocation d'une variable.

== note une comparaison pour égalité.

!= note une comparaison pour inégalité.

Les accolades { et } sont utilisées pour grouper.

3. Spécification du protocole

La spécification de BIDIR-PIM est divisée en plusieur parties :
Le paragraphe 3.1 détaille l'état de protocole mémorisé.
Le paragraphe 3.2 définit les extensions BIDIR-PIM au mécanisme de découverte de voisin PIM-SM [RFC4601].
Le paragraphe 3.3 spécifie les règles de transmission de paquet de données.
Le paragraphe 3.4 spécifie les règles de génération et de traitement des Join/Prune BIDIR-PIM.
Le paragraphe 3.5 spécifie l'élection du transmetteur désigné (DF).
Le paragraphe 3.6 récapitule les temporisateurs BIDIR-PIM et donne leurs valeurs par défaut.
Le paragraphe 3.7 spécifie les formats de paquets PIM.

3.1 États du protocole BIDIR-PIM

Ce paragraphe spécifie tout l'état de protocole qu'une mise en œuvre de BIDIR-PIM devrait maintenir afin de fonctionner correctement. On appelle cet état la base de données d'informations d'arborescence (TIB, *Tree Information Base*) car il contient l'état de toutes les arborescences de distribution de diffusion groupée à ce routeur. Dans la présente spécification, on définit les mécanismes PIM en termes de TIB. Cependant, seulement une mise en œuvre très simple ferait les opérations de transmission de paquet réelles dans les termes de cet état. La plupart des mises en œuvre vont utiliser cet état pour construire un tableau de transmission de diffusion groupée, qui devrait alors être mis à jour quand l'état pertinent change dans la TIB.

Bien qu'on spécifie précisément l'état à conserver, cela ne signifie pas qu'une mise en œuvre de BIDIR-PIM ait besoin de détenir l'état dans cette forme. C'est en fait une définition d'état abstrait, qui est nécessaire afin de spécifier le comportement du routeur. Une mise en œuvre de BIDIR-PIM est libre de conserver tout état interne qu'elle exige, et va encore être conforme à la présente spécification pour autant qu'il en résulte le même comportement de protocole visible de l'extérieur qu'un routeur abstrait détenant l'état suivant.

On divise l'état de TIB en deux sections :
état de RPA : état qui conserve les informations d'élection de DF pour chaque RPA ;
état de groupe : état qui conserve l'arborescence spécifique d'un groupe pour les groupes qui se transposent en une certaine RPA.

L'état qui devrait être conservé est décrit ci-dessous. Bien sûr, les mises en œuvre vont seulement conserver l'état quand il est pertinent pour les opérations de transmission - par exemple, l'état "NoInfo" pourrait être supposé à partir du manque d'autres informations d'état, plutôt que d'être tenu explicitement.

3.1.1 État d'objet général

Un routeur détient l'état suivant qui n'est pas spécifique d'une RPA ou d'un groupe :

État de voisin :
pour chaque voisin :
l'identifiant général du voisin
le temporisateur de vie de voisin (NLT, *Neighbor Liveness Timer*)
d'autres informations provenant du Hello du voisin.

Pour en savoir plus sur les informations de Hello, voir au paragraphe 3.2 ainsi que dans la spécification PIM-SM [RFC4601].

3.1.2 État RPA

Un routeur tient une transposition de groupe de diffusion groupée en RPA, qui est construite par configuration statique ou en utilisant un mécanisme de découverte automatique de RP comme BSR ou AUTO-RP (voir la Section 4). Pour chaque RPA BIDIR-PIM, un routeur garde l'état suivant :

RPA (adresse actuelle)
État du transmetteur désigné (DF) :
pour chaque interface de routeur :
    Informations sur le DF agissant :
        Adresse IP du DF
        Métrique du DF
    Informations d'élection :
        état d'élection
        temporisation d'élection de DF (DTF, *DF Election Timer*)
        compte de messages (MC, *Message Count*)
    Meilleure offre actuelle :
        adresse IP du routeur de la meilleure offre
        métrique du routeur de la meilleure offre

L'état du transmetteur désigné est décrit au paragraphe 3.5.

3.1.3 État de groupe

Pour chaque groupe G, un routeur garde l'état suivant :
État de groupe :
pour chaque interface :
Membres locaux :
    o état : un de {"NoInfo", "Include"}
État PIM Join/Prune :
    o État : un de {"NoInfo" (NI), "Join" (J), "PrunePending" (PP)}
    o Temporisateur Élagage en cours (PPT, *PrunePendingTimer*)
    o Temporisateur Join/Prune expiré (ET)
Non spécifique de l'interface :
    o Temporisateur Join/Prune vers l'amont (JT)
    o Dernière RPA utilisée

Membres locaux est le résultat du mécanisme d'adhésion locale (comme IGMP [RFC3376]) fonctionnant sur cette interface. Cette information est utilisée par la macro pim_include(*,G) décrite au paragraphe 3.1.4.

État PIM Join/Prune est le résultat de la réception de messages Join/Prune (*,G) PIM sur cette interface, et est spécifié au paragraphe 3.4.1. L'état est utilisé par les macros qui calculent la liste des interfaces sortantes au paragraphe 3.1.4, et dans la macro JoinDesired(G) (définie au paragraphe 3.4.2) qui est utilisée pour décider si un Join(*,G) devrait être envoyé vers l'amont.

Le temporisateur Join/Prune vers l'amont est utilisé pour envoyer des messages périodiques Join(*,G) et pour outrepasser les messages Prune(*,G) provenant d'homologues sur une interface de LAN en amont.

La dernière RPA utilisée doit être mémorisée parce que si la transposition de groupe en RPA change (voir les changements d'ensemble de RP dans la [RFC4601]) l'état doit alors être supprimé et reconstruit pour les groupes dont la RPA change.

3.1.4 Macros de récapitulation d'états

En utilisant cet état, on définit les "macros" suivantes qu'on va utiliser dans les descriptions des automates à états et le pseudo code dans les paragraphes suivants.

    olist(G) = RPF_interface(RPA(G)) (+) joins(G) (+) pim_include(G)

RPF_interface(RPA) est l'interface que la MRIB indique comme devant être utilisée pour acheminer les paquets à la RPA. olist(G) est la liste des interfaces sur lesquels les paquets pour le groupe G doivent être transmis.

La macro pim_include(G) indique les interfaces auxquelles le trafic peut être transmis parce que des hôtes sont des membres locaux sur cette interface.

pim_include(G) = { toutes les interfaces I telles que : I_am_DF(RPA(G),I) ET local_receiver_include(G,I) }

La clause "I_am_DF(RPA,I)" est VRAI si le routeur est dans les états Win ou Backoff dans l'automate à états de l'élection de DF (décrite au paragraphe 3.5) pour la RPA donnée sur l'interface I. Autrement, elle est FAUX.

La clause "local_receiver_include(G,I)" est vraie si le module IGMP, le module MLD, ou un autre mécanisme d'adhésion local a déterminé qu'il y a des membres locaux sur l'interface I qui désirent recevoir le trafic envoyé au groupe G.

L'ensemble "joins(G)" est l'ensemble de toutes les interfaces sur lesquelles le routeur a reçu des Joins(*,G) :

joins(G) = { toutes interfaces I telles que I_am_DF(RPA(G),I) ET DownstreamJPState(G,I) est Joined ou PrunePending }

DownstreamJPState(G,I) est l'état de l'automate à états finis du paragraphe 3.4.1.

RPF_DF(RPA) est le voisin auquel les messages Join doivent être envoyés afin de construire l'arborescence partagée de groupe dont la racine est au RPL pour la RPA donnée. C'est le transmetteur désigné sur la RPF_interface(RPA).

3.2 Découverte de voisin PIM

Les routeurs PIM échangent des messages PIM-Hello avec leurs routeurs PIM du voisinage. Ces messages sont utilisés pour mettre à jour l'état de voisin décrit au paragraphe 3.1. Les procédures pour générer et traiter les messages Hello et maintenir l'état de voisin sont spécifiées dans le document PIM-SM [RFC4601].

BIDIR-PIM introduit l'option PIM-Hello à capacité bidirectionnelle qui DOIT être incluse dans tous les messages Hello provenant d'un routeur à capacité BIDIR-PIM. L'option de capacité bidirectionnelle annonce la capacité du routeur à participer au protocole BIDIR-PIM. Le format de l'option de capacité bidirectionnelle est decrit au paragraphe 3.7.

Si un routeur BIDIR-PIM reçoit d'un de ses voisins un message PIM-Hello qui ne contient pas l''option de capacité bidirectionnelle, l'erreur doit être enregistrée chez l'administrateur du routeur d'une manière limitée en débit.

3.3 Règles de transmission des paquets de données

Pour les groupes qui se transposent en une certaine RPA, les responsabilités suivantes sont allouées de façon unique au DF pour cette RPA sur chaque liaison :
- o le DF est le seul routeur qui transmet des paquets voyageant vers l'aval sur la liaison ;
- o le DF est le seul routeur qui prenne des paquets voyageant vers l'amont sur la liaison à transmettre vers la RPL.

Les routeurs non DF sur une liaison, qui utilisent cette liaison comme leur interface de RPF pour atteindre la RPA, peuvent effectuer les actions de transmission suivantes pour les groupes bidirectionnels :
- o Transmettre les paquets provenant de la liaison vers les receveurs en aval.
- o Transmettre les paquets provenant de sources en aval sur la liaison (pourvu qu'ils soient le DF pour la liaison aval d'où le paquet a été pris).

Les règles de transmission de paquet BIDIR-PIM sont définies ci-dessous en pseudo code.
iif est l'interface entrante du paquet ;
G est l'adresse de destination du paquet (adresse de groupe).
RPA est l'adresse de point de rendez-vous pour ce groupe.

D'abord on vérifie si le paquet devrait être accepté sur la base de l'état de TIB et de l'interface sur laquelle le paquet est arrivé. Un paquet est accepté si il arrive sur l'interface RPF pour atteindre la RPA (paquet voyageant vers l'aval) ou si le routeur est le DF sur l'interface sur laquelle le paquet arrive (paquet voyageant vers l'amont).

Si le paquet devrait être transmis, on construit une liste d'interfaces sortantes pour le paquet.

Finalement, on supprime l'interface entrante de la liste des interfaces sortantes qu'on a créée, et si la liste résultante des

interfaces sortantes n'est pas vide, on transmet le paquet par ces interfaces.

```
À réception des données pour G sur l'interface iif :
 si( iif == RPF_interface(RPA) || I_am_DF(RPA,iif) ) { oiflist = olist(G) (-) iif
      transmettre le paquet sur toutes les interfaces de oiflist
  }
```

3.3.1 Transmission en amont au RP

Quand on configure un domaine BIDIR-PIM, il est possible d'allouer l'adresse de point de rendez-vous (RPA) de telle sorte qu'elle n'appartienne pas à une boîte physique mais soit simplement une adresse acheminable. Les routeurs qui ont des interfaces sur la RPL à laquelle appartient la RPA vont transmettre en amont le trafic sur la liaison. Les messages Joins provenant des receveurs dans le domaine vont se propager bond par bond jusqu'à ce qu'ils atteignent un des routeurs connectés à la RPL où ils vont terminer (car il ne va pas y avoir de DF élu sur la RPL).

Si l'administrateur choisit plutôt de configurer la RPA à être l'adresse d'une interface physique d'un routeur spécifique, rien ne change alors. Ce routeur doit quand même transmettre le trafic en amont à la RPL et se comporter de la même façon que tout autre routeur ayant une interface sur la RPL.

Pour configurer un réseau BIDIR-PIM à opérer dans un mode similaire à celui de PIM-SM où un seul routeur (le RP) agit comme racine de l'arborescence de distribution, la RPA peut être configurée à être l'interface de rebouclage d'un routeur.

3.3.2 Branches seulement de source

Les branches seulement de source de l'arborescence de distribution pour un groupe G sont des branches qui ne conduisent à aucun des receveurs, mais sont utilisées pour transmettre les paquets qui voyagent vers l'amont des sources vers la RPL. Les routeurs le long de branches seulement de source ont l'interface de RPF à la RPA dans leur olist pour G, et n'ont donc pas besoin de conserver d'état spécifique d'un groupe. La transmission vers l'amont peut être effectuée en utilisant seulement l'état spécifique de la RPA. Une mise en œuvre peut décider de conserver l'état de groupe pour les branches seulement de source pour des raisons de comptabilité ou de performances. Cependant, le faire exige des événements pilotés par les données (pour découvrir les groupes qui ont des sources actives) et donc de sacrifier un des principaux avantages de BIDIR-PIM.

3.3.3 Sources directement connectées

Un avantage majeur de l'utilisation d'un transmetteur désigné dans BIDIR-PIM comparé à PIM-SM est qu'un traitement spécial n'est plus exigé pour les sources qui sont directement connectées à un routeur. Les données provenant de telles sources n'ont pas besoin d'être différenciées des autres trafics de diffusion groupée et vont automatiquement être prises par le DF et transmises en amont. Cela supprime le besoin d'effectuer une vérification de source directement connectée pour les données aux groupes qui n'ont pas d'état existant.

3.4 Messages PIM Join/Prune

Les messages Join/Prune BIDIR-PIM sont utilisés pour construire des arborescences de distribution spécifiques de groupe entre les receveurs et la RPL. Les Joins sont générés par les routeurs de dernier bond qui sont élus comme DF sur une interface avec les receveurs directement connectés. Les Joins se propagent bond par bond vers la RPA jusqu'à ce qu'ils atteignent un routeur connecté à la RPL.

Un message Join/Prune BIDIR-PIM consiste en une liste de groupes joints et élagués. Lors du traitement d'un message Join/Prune reçu, chaque groupe joint ou élagué est effectivement considéré individuellement en appliquant les automates à état suivants. Quand on considère un message Join/Prune dont le champ Destination PIM vise ce routeur, les Joins et Prunes (*,G) peuvent affecter l'automate à états aval. Quand on considère un message Join/Prune dont le champ Destination PIM vise un autre routeur, la plupart des entrées Join ou Prune pourraient affecter l'automate à état amont.

3.4.1 Réception des messages Join/Prune (*,G)

Quand un routeur reçoit un Join(*,G) ou Prune(*,G), il DOIT d'abotd vérifier si l'adresse de RP dans le message correspond à la RPA(G) (idée du routeur de ce qu'est l'adresse de point de rendez-vous). Si l'adresse de RP dans le message ne

correspond pas à la RPA(G), le Join ou Prune DOIT être éliminé en silence.

Si un routeur n'a pas d'informations de RPA pour le groupe (par exemple, il n'a pas reçu récemment de message BSR) il PEUT alors choisir d'accepter le Join(*,G) ou Prune(*,G) et de traiter l'adresse de RP dans le message comme RPA(G). Si la RPA nouvellement découverte n'existait pas précédemment pour un autre groupe, une élection de DF doit être initiée.

Noter qu'un routeur va traiter un Join(*,G) ciblé sur lui-même même si il n'est pas le DF pour RP(G) sur l'interface sur laquelle le message a été reçu. C'est une optimisation pour éliminer le délai de Join de une période Join (t_periodic) dans le cas où un nouveau DF traite les messages Pass et Join reçus dans l'ordre inverse. La logique de transmission BIDIR-PIM va assurer que les paquets de données ne sont pas transmis sur une telle interface alors que le routeur n'est pas le DF (sauf si il est l'interface de RPF vers la RPA).

L'automate à états par interface pour la réception des messages Join/Prune (*,G) est montré ci-après. Il y a trois états :

NoInfo (NI) : l'interface n'a pas d'état Join (*,G) et aucun temporisateur en cours.

Join (J) : l'interface a l'état Join (*,G). Si le routeur est le DF sur cette interface (I_am_DF(RPA(G),I) est VRAI) l'état Join va causer la transmission des paquets destinés à G sur cette interface.

PrunePending (PP) : le routeur a reçu un Prune(*,G) sur cette interface d'un voisin en aval et il attend pour voir si le Prune va être outrepassé par un autre routeur en aval. Pour les besoins de la transmission, l'état PrunePending fonctionne exactement comme l'état Join.

De plus, l'automate à état utilise deux temporisateurs :

ExpiryTimer (ET) : ce temporisateur est redémarré quand un Join(*,G) valide est reçu. L'arrivée à expiration du ExpiryTimer cause le retour de l'état de l'interface à NoInfo pour ce groupe.

PrunePendingTimer (PPT) : ce temporisateur est lancé quand un Prune(*,G) valide est reçu. L'arrivée à expiration du PrunePendingTimer cause le retour de l'état de l'interface à NoInfo pour ce groupe.

Figure 1 : Automate à états par interface de groupe aval sous forme de tableau

		État précédent	
Événement	NoInfo (NI)	Join (J)	PrunePending (PP)
Reçoit Join(*,G)	-> état J ; lance ET	-> état J ;relance ET	-> état J ; relance ET ; arrête PPT
Reçoit Prune(*,G)	-	-> état PP ; lance PPT	-> état PP
Expiration de PPT	-	-	-> état NI ; envoie Prune-Echo(*,G)
Expiration de ET	-	-> état NI	-> état NI
Cesse d'être DF sur I	-	-> état NI	-> état NI

Les événements de transition "Reçoit Join(*,G)" et "Reçoit Prune(*,G)" impliquent de recevoir un Join ou Prune ciblé sur l'adresse de ce routeur sur l'interface de réception. Si l'adresse de destination n'est pas correcte, ces transitions d'état dans cet automate à états ne doivent pas se produire, bien que voir un tel paquet pourrait causer des transitions d'état dans d'autres automates à états.

Sur des interfaces non numérotées sur des liaisons point à point, l'adresse du routeur devrait être la même que l'adresse de source choisie pour le paquet Hello qu'il a envoyé sur cette interface. Cependant, sur les liaisons point à point, on RECOMMANDE aussi que les messages PIM avec une adresse de destination toute de zéros soient aussi acceptés.

L'événement de transition "Cesse d'être DF" implique qu'une réélection de DF a eu lieu sur cette interface de routeur pour RPA(G) et que le routeur change d'état du statut de DF actif à celui de routeur non DF (la valeur de la macro I_am_DF change à FAUX).

Quand ExpiryTimer est lancé ou relancé, il est réglé au HoldTime provenant du message Join/Prune qui a déclenché le temporisateur.

Quand PrunePendingTimer est lancé, il est réglé à J/P_Override_Interval si le routeur a plus d'un voisin sur cette interface ; sinon, il est réglé à zéro, causant son expiration immédiate.

L'action "Send PruneEcho(*,G)" est déclenchée quand le routeur cesse de transmettre sur une interface par suite d'un

élagage *(Prune)*. Un PruneEcho(*,G) est simplement un message Prune(*,G) envoyé par le routeur amont à lui-même sur un LAN. Son objet est d'ajouter une fiabilité supplémentaire afin que si un Prune qui aurait dû être outrepassé par un autre routeur est perdu localement sur le LAN, le PruneEcho puisse alors être reçu et causer l'intervention de l'outrepassement. Un PruneEcho(*,G) n'a pas besoin d'être envoyé quand le routeur a seulement un voisin sur la liaison.

3.4.2 Envoi des messages Join/Prune

Les automates à états par interface vers l'aval décrits ci-dessus gardent l'état Join provenant des routeurs PIM vers l'aval. Cet état détermine alors si un routeur a besoin de propager un Join(*,G) en amont vers la RPA. De tels messages Join(*,G) sont envoyés sur l'interface de RPF vers la RPA et sont ciblés sur le DF de cette interface.

Si un routeur souhaite propager en amont un Join(*,G) il doit aussi surveiller les messages sur son interface amont provenant des autres routeurs sur ce sous réseau, et cela peut modifier son comportement. Si il voit un Join(*,G) pour le voisin amont correct, il devrait supprimer son propre Join(*,G). Si il voit un Prune(*,G) pour le voisin amont correct, il devrait être prêt à outrepasser ce Prune en envoyant presque immédiatement un Join(*,G). Finalement, si il voit changer l'identifiant de génération (voir la spécification PIM-SM [RFC4601]) du voisin amont correct, il sait que le voisin amont a perdu l'état, et il devrait être prêt à rafraîchir l'état en envoyant presque immédiatement un Join(*,G).

De plus, des changements du prochain bond vers la RPA déclenchent un élagage de l'ancien prochain bond et une jonction vers le nouveau prochain bond. Un tel changement peut être causé par les deux événements suivants :
- o la MRIB indique que l'interface de RPF vers la RPA a changé. Dans ce cas, le DF sur la nouvelle interface de RPF devient le nouveau voisin de RPF ;
- o il y a une réélection de DF sur l'interface de RPF et un nouveau routeur émerge comme DF.

L'automate à états amont (*,G) contient seulement deux état :

Non joint : l'automate à états aval indique que le routeur n'a pas besoin de rejoindre l'arborescence de RPA pour ce groupe.

Joint : l'automate à états aval indique que le routeur aimerait se joindre à l'arborescence de RPA pour ce groupe.

De plus, un temporisateur JT(G) est tenu, utilisé pour déclencher l'envoi d'un Join(*,G) au prochain bond amont vers la RPA (le DF sur l'interface de RPF pour RPA(G)).

Figure 2 : Automate à états par interface de groupe amont sous forme de tableau

	Événement	
État précédent	JoinDesired(G) ->Vrai	JoinDesired(G) ->Faux
NonJoint (NJ)	-> état J envoie Join(*,G) ; régler tempo à t_periodic	-
Joint (J)	-	-> état NJ ; envoie Prune(*,G)

De plus, les transitions suivantes se produisent dans l'état Joint :

Dans l'état Joint (J)			
Temporisateur expire	Voit Join(*,G) à RPF_DF(RPA(G))	Voit Prune(*,G) à RPF_DF(RPA(G))	Changements de GenID de RPF_DF(RPA(G))
Envoie Join(*,G) ; règle tempo à t_periodic	Augmente tempo à t_suppressed	Diminue tempo à t_override	Diminue tempo à t_override

Dans l'état Joint (J)	
Changement de RPF_DF(RPA(G))	Changements de GenID de RPF_DF(RPA(G))
Envoie Join(*,G) au nouveau DF ; envoie Prune(*,G) au vieux DF ; règle le temporisateur à t_periodic	Diminue le temporisateur à t_override

Cet automate à états utilise la macro suivante :

```
bool JoinDesired(G) {
  si (olist(G) (-) RPF_interface(RPA(G))) != NULL
    retourne VRAI
  autrement
    retourne FAUX
}
```

3.5 Élection du transmetteur désigné (DF)

Ce paragraphe présente un mécanisme résistant à l'échec pour élire un routeur désigné par RPA sur chaque liaison dans un domaine BIDIR-PIM. On appelle ce routeur le transmetteur désigné (DF). L'élection du DF n'a pas lieu sur la RPL pour une RPA.

3.5.1 Exigences pour le transmetteur désigné

L'élection du DF choisit le meilleur routeur sur une liaison pour prendre la responsabilité de la transmission du trafic entre la RPL et la liaison pour la gamme de groupes de diffusion groupée desservie par la RPA. Différents groupes de diffusion groupée qui partagent une RPA commune partagent la même direction vers l'amont. Donc, l'élection d'un transmetteur vers l'amont sur chaque liaison n'a pas à être une décision spécifique du groupe mais peut plutôt être spécifique de la RPA. Comme le nombre de RPA est normalement petit, le nombre d'élections qui doivent être effectuées est significativement réduit par cette observation.

Pour optimiser la création d'arborescences, il est souhaitable que le vainqueur du processus d'élection soit le routeur sur la liaison qui a la "meilleure" métrique d'acheminement en envoi individuel (comme rapportée par la MRIB) pour atteindre la RPA. Quand on compare les métriques provenant de différents protocoles d'acheminement en envoi individuel, on utilise les mêmes règles de comparaison qu'utilisées par le processus d'assertion PIM-SM [RFC4601].

Le processus d'élection doit avoir lieu quand des informations sur une nouvelle RPA deviennent initialement disponibles. Le résultat peut être réutilisé lorsque de nouveaux groupes bidirectionnels qui se transposent en la même RPA sont rencontrés. Cependant, il y a des conditions dans lesquelles une mise à jour de l'élection est nécessaire :

- il y a un changement dans la métrique d'envoi individuel pour atteindre la RPA pour un des routeurs sur la liaison ;
- l'interface sur laquelle la RPA est accessible (interface de RPF) change pour une interface pour laquelle le routeur était précédemment le DF ;
- un nouveau voisin PIM qui doit participer à l'élection et doit être informé du résultat commence sur une liaison ;
- le DF élu échoue (détecté par la fin de temporisation d'informations de voisin ou un changement de RPF de MRIB chez un routeur en aval).

Le processus d'élection doit être assez robuste pour assurer avec une très forte probabilité que tous les routeurs sur la liaison ont une vue cohérente du DF. Selon les règles de transmission décrites au paragraphe 3.3, des boucles peuvent résulter de ce que plusieurs routeurs finissent par penser qu'ils devraient être responsables de la transmission. Pour minimiser la possibilité de cette occurrence, l'algorithme d'élection a été biaisé pour éliminer les informations de DF et suspendre la transmission durant les périodes d'ambiguïté.

3.5.2 Description de l'élection du transmetteur désigné

Ce paragraphe donne les généralités du processus d'élection de DF. Il ne fournit pas la spécification définitive de l'élection de DF. Si des discordances existent entre le paragraphe 3.5.3 et celui-ci, la spécification du paragraphe 3.5.3 est supposée être correcte.

Pour effectuer l'élection du DF pour une RPA particulière, les routeurs sur une liaison ont besoin d'échanger leurs informations de métrique d'acheminement en envoi individuel pour atteindre la RPA. Les routeurs annoncent leur propre métrique dans les messages Offer *(offre)*, Winner *(gagnant)*, Backoff *(retard)*, et Pass. La métrique annoncée est calculé en utilisant l'interface et la métrique de RPF pour atteindre la RPA disponible à partir de la MRIB. Quand un routeur participe à une élection de DF pour une RPA sur l'interface que sa MRIB indique comme étant l'interface de RPF, ce routeur DOIT alors toujours annoncer une métrique infinie dans ses messages d'élection. Quand un routeur participe à une élection de DF sur une interface autre que l'interface de RPF indiquée par la MRIB, il DOIT alors annoncer les métriques fournies par la MRIB dans ses messages d'élection.

Dans le protocole d'élection décrit ci-dessous, de nombreux échanges de messages sont répétés Election_Robustness fois pour la fiabilité. Dans tous ces cas, les retransmissions de message sont espacées d'un court intervalle aléatoire. Toute la description qui suit est spécifique de l'élection sur une seule liaison pour une seule RPA.

3.5.2.1 Élection d'amorçage

Initialement, quand aucun DF n'a été élu, les routeurs qui trouvent une nouvelle RPA commencent à participer à l'élection en envoyant des messages Offer. Les messages Offer incluent la métrique du routeur pour atteindre la RPA. Les offres sont périodiquement retransmises à une période de Offer_Interval.

Si un routeur entend une meilleure offre que la sienne de la part d'un voisin, il cesse de participer à l'élection pour une période de Election_Robustness * Offer_Interval, donnant donc une chance au voisin qui a la meilleure métrique d'être élu DF. Si durant cette période aucun gagnant n'est élu, le routeur recommence l'élection depuis le début. Si à tout moment durant l'élection initiale un routeur reçoit une offre décalée avec une moins bonne métrique que la sienne propre, il recommence l'élection depuis le début.

Le résultat devrait être que tous les routeurs sauf le meilleur candidat cessent d'annoncer leurs offres.

Un routeur assume le rôle de DF après avoir annoncé Election_Robustness fois sa métrique sans recevoir d'offre d'un autre voisin. À ce point, il transmet un message Winner qui déclare à tous les autres routeurs sur la liaison l'identité du gagnant et la métrique qu'il utilise.

Les routeurs qui reçoivent un message Winner cessent de participer à l'élection et enregistrent l'identité et la métrique du gagnant. Si la métrique locale est meilleure que celle du gagnant, le routeur enregistre alors l'identité du gagnant (l'acceptant comme DF agissant) mais réinitie l'élection pour essayer de l'emporter.

3.5.2.2 Changements de la métrique du perdant

Chaque fois que la métrique d'envoi individuel pour une RPA change à un routeur non DF pour une valeur qui est meilleure que celle précédemment annoncée par le DF agissant, le routeur avec la nouvelle meilleure métrique devrait prendre des mesures pour prendre finalement la responsabilité de la transmission. Quand le changement de métrique est détecté, le routeur non DF avec la maintenant meilleure métrique recommence le processus d'élection de DF en envoyant des messages Offer avec sa nouvelle métrique. Noter qu'à tout moment durant une élection si aucune réponse n'est reçue après Election_Robustness retransmissions d'une offre, un routeur assume le rôle de DF suivant la procédure usuelle d'annonce de gagnant.

À réception d'une offre qui est moins bonne que sa métrique courante, le DF va répondre par un message Winner déclarant son état et annonçant sa meilleure métrique. À réception du message Winner, l'origine du message Offer enregistre l'identité du DF et interrompt l'élection.

À réception d'une offre meilleure que sa métrique actuelle, le DF enregistre l'identité et la métrique du routeur offreur et répond par un message Backoff. Cela donne pour instruction au routeur offreur de tenir pendant une courte période pendant que l'acheminement en envoi individuel se stabilise et que les autres routeurs aient une chance d'établir leurs offres. Le message Backoff inclut la nouvelle métrique et l'adresse du routeur offreur. Tous les routeurs sur la liaison qui ont des offres en instance avec des métriques moins bonnes que celle du message Backoff (incluant l'offre du routeur original) vont tenir les autres offres pendant une période définie dans le message Backoff.

Si un troisième routeur envoie une meilleure offre durant Backoff_Period, le message Backoff est répété pour la nouvelle offre et Backoff_Period est redémarré.

Avant l'expiration de Backoff_Period, le DF agissant désigne le routeur qui a fait la meilleure offre comme nouveau DF en utilisant un message Pass. Ce message inclut les identifiants et les métriques de l'ancien et du nouveau DF. L'ancien DF cesse d'effectuer ses tâches au moment où la transmission du message Pass est faite. Le nouveau DF assume le rôle de DF aussitôt qu'il reçoit le message Pass. Tous les autres routeurs sur la liaison prennent note du nouveau DF et de sa métrique. Noter que cet événement constitue un changement de voisin de RPF, qui peut déclencher des messages Join au nouveau DF (voir au paragraphe 3.4).

3.5.2.3 Changements de la métrique du gagnant

Si la métrique d'acheminement du DF pour atteindre la RPA change pour une moins bonne valeur, il envoie un ensemble de Election_Robustness messages Winner espacés aléatoirement sur la liaison, annonçant la nouvelle métrique. Les routeurs qui reçoivent cette annonce et ont une meilleure métrique peuvent répondre par un message Offer qui résulte en la même procédure de mîse à l'écart que décrit ci-dessus. Tous les routeurs supposent que le DF n'a pas changé jusqu'à ce qu'ils voient un message Pass ou Winner qui indique le changement.

Il n'y a pas de pression pour faire rapidement ce transfert si le DF agissant a toujours un chemin avec la RPL. Le vieux chemin peut être maintenant sous optimal, mais il va quand même fonctionner pendant que la réélection est en cours.

3.5.2.4 Le gagnant perd le chemin

Si l'interface de RPF d'un routeur vers la RPA passe à être sur une liaison pour laquelle il agit comme DF, alors il ne peut plus fournir de services de transmission pour cette liaison. Il cesse donc immédiatement d'être le DF et recommence l'élection. Comme son chemin pour la RPA est à travers la liaison, une métrique infinie est utilisée dans le message Offer qu'il envoie.

3.5.2.5 Démarrage d'un routeur tardif

Un routeur tardif qui commence après l'achèvement du processus d'élection du DF ne va pas avoir une connaissance immédiate du résultat de l'élection. Par suite, il va commencer par annoncer sa métrique dans des messages Offer. Aussitôt que cela arrive, le DF actuellement élu va répondre avec un message Winner si sa métrique est meilleure que celle du message Offer, ou avec un message Backoff si sa métrique est moins bonne que la métrique du message Offer.

3.5.2.6 Mort du gagnant

Chaque fois que le DF meurt, un nouveau DF doit être élu. La vitesse à laquelle cela peut être réalisé dépend de si il y a des routeurs en aval sur la liaison.

Si il y a des routeurs en aval, normalement leur prochain bond rapporté par la MRIB avant le décès du DF va être le DF lui-même. Ils vont donc remarquer un changement dans la métrique pour le chemin vers la RPA ou un changement du prochain bond à partir du DF et peuvent recommencer l'élection en transmettant des messages Offer. Si conformément à la MRIB la RPA est maintenant accessible par la même liaison via un autre routeur en amont, une métrique infinie va être utilisée dans l'offre.

Si aucun routeur n'est présent en aval, la seule façon qu'un autre routeur en amont détecte une défaillance de DF est l'expiration du temporisateur d'informations de voisin PIM, ce qui va prendre significativement plus longtemps.

3.5.3 Spécification du protocole d'élection

Ce paragraphe donne la spécification définitive du processus d'élection de DF. Si une divergeance existe entre le paragraphe 3.5.2 et ce paragraphe, la spécification de ce paragraphe est supposée être correcte.

3.5.3.1 État d'élection

L'état de l'élection de DF est conservé par RPA pour chaque interface I à capacité de diffusion groupée sur le routeur comme indiqué au paragraphe 3.1.

L'automate à états a les quatre états suivants :

Offer : état initial d'élection. Quand il est dans l'état Offer, un routeur pense qu'il peut éventuellement devenir le vainqueur et génère périodiquement des messages Offer.

Lose : dans cet état, le routeur sait qu'il y a un vainqueur différent de l'élection ou qu'aucun routeur sur la liaison n'a de chemin pour la RPA.

Win : le routeur est le DF agissant sans contestation.

Backoff : le routeur est le DF agissant mais un autre routeur a fait une enchère pour prendre sa place.

Dans l'automate à états, un routeur est considéré être un DF agissant si il est dans les états Win ou Backoff.

Le fonctionnement du protocole d'élection utilise les variables et temporisateurs décrits ci-après :

Informations de DF agissant : utilisées pour mémoriser l'identité et annoncer les métriques du vainqueur de l'élection qui est le DF agissant actuellement.

Temporisateur d'élection de DF (DFT) : utilisé pour programmer la transmission des messages Offer, Winner, et Pass.

Compte de messages (MC, *Message-Count*) : utilisé pour tenir le nombre de fois qu'un message Offer ou Winner a été transmis.

Meilleure offre : utilisé par le DF pour enregistrer l'identité et annoncer les métriques du routeur qui a fait la dernière offre, pour l'utiliser lors de l'envoi du message Pass.

3.5.3.2 Messages d'élection

Le processus d'élection utilise les messages de contrôle PIM suivants. Le format de paquet est décrit au paragraphe 3.7:

Offer (OfferingID, Metric) : envoyé par les routeurs qui estiment avoir une meilleure métrique pour la RPA que celle qui a été offerte jusqu'à présent.

Winner (DF-ID, DF-Metric) : envoyé par un routeur quand il assume le rôle de DF ou quand il le réaffirme en réponse à de moins bonnes offres.

Backoff (DF-ID, DF-Metric, OfferingID, OfferMetric, BackoffInterval) : envoyé par le DF pour accuser réception de meilleures offres. Il donne pour instruction aux autres routeurs avec des offres égales ou moins bonnes d'attendre jusqu'à ce que le DF passe la responsabilité à l'envoyeur de l'offre.

Pass (Old-DF-ID, Old-DF-Metric, new-DF-ID, new-DF-Metric) : utilisé par l'ancien DF pour passer la responsabilité de la transmission à un routeur qui avait précédemment fait une offre. La Old-DF-Metric est la métrique actuelle du DF au moment où le Pass est envoyé.

Noter que quand un routeur participe à une élection de DF pour une RPA sur l'interface que sa MRIB indique comme Interface de RPF, ce routeur DOIT toujours annoncer une métrique infinie dans ses messages d'élection. Quand un routeur participe à une élection de DF sur une interface autre que celle indiquée par la MRIB, il DOIT alors annoncer dans ses messages d'élection les métriques fournies par la MRIB.

3.5.3.3 Événements d'élection

Durant le fonctionnement du protocole, les événements suivants peuvent avoir lieu :

réception de message de contrôle : réception d'un des quatre messages de contrôle d'élection de DF (Offer, Winner, Backoff, et Pass). Quand un message de contrôle est reçu et que des actions sont spécifiées sur une condition où les métriques sont Better *(meilleure)* ou Worse *(moins bonne)* la comparaison doit être effectuée comme suit :

- o À réception d'un message Offer ou Winner, comparer les métriques actuelles pour la RPA avec les métriques annoncées par l'envoyeur du message.
- o À réception d'un message Backoff ou Pass, comparer les métriques actuelles pour la RPA avec les métriques annoncées par la cible du message.

perte du chemin de la RPA : la perte du chemin de la RPA peut arriver de deux façons. Le première arrive quand le chemin appris par la MRIB est supprimé et que la MRIB ne rapporte plus de chemin disponible pour atteindre la RPA. Le second cas arrive quand les informations de prochain bond rapportées par la MRIB changent pour indiquer un prochain bond qui est accessible par l'interface du routeur considéré. En clair, comme le routeur utilise l'interface comme son interface de RPF, il ne peut pas offrir de transmettre des services vers la RPL aux autres routeurs sur cette liaison.

changement de la métrique rapportée par la MRIB pour accéder à la RPA : cet événement est déclenché quand la MRIB a fourni des informations qui changent pour la RPA et que les nouvelles informations fournissent un chemin pour la RPA. Si les nouvelles informations de la MRIB ne rapportent pas de chemin ou rapportent une interface de prochain bond à travers l'interface pour laquelle l'élection de DF a lieu, alors c'est plutôt l'événement "Perte du chemin de la RPA" qui déclenche. Dans des états spécifiques, l'événement peut être filtré un peu plus en spécifiant si on s'attend à ce que la métrique devienne meilleure ou moins bonne et à quelles métriques mémorisées des nouvelles informations de la MRIB elle doit être comparée. Les nouvelles informations doivent être comparées soit à la vieille métrique du routeur, soit à la

métrique de DF mémorisée, soit à la métrique de meilleure offre mémorisée.

expiration du temporisateur d'élection (DFT) : l'expiration du temporisateur d'élection DFT peut causer la transmission de messages et des transitions d'état. L'événement pourrait être plus qualifié en spécifiant la valeur du compte de messages (MC) ainsi que l'existence actuelle d'un chemin pour la RPA (comme défini ci-dessus).

Détection d'une défaillance de DF : la détection d'une défaillance du DF peut survenir par la fin de temporisation de l'état de voisin PIM.

3.5.3.4 Actions d'élection

La description des actions de l'automate à états d'élection de DF utilise la notation suivante en plus de la notation de pseudo code décrite précédemment dans cette spécification:

?= note l'opération de diminution d'un temporisateur à une nouvelle valeur. Si le temporisateur ne fonctionne pas, il est alors démarré en utilisant la nouvelle valeur. Si le temporisateur fonctionne avec un temps d'expiration plus faible que la nouvelle valeur, alors le temporisateur n'est pas altéré.

Quand une action de "régler le DF à l'envoyeur ou la cible" est rencontrée durant la réception d'un message Winner, Pass, ou Backoff, cela signifie ce qui suit :
- o à réception d'un message Winner, régler le DF à être l'origine du message et enregistrer sa métrique ;
- o à réception d'un message Pass, régler le DF à être la cible du message et enregistrer sa métrique ;
- o à réception d'un message Backoff, régler le DF à être l'origine du message et enregistrer sa métrique.

3.5.3.5 Transitions d'état d'élection

Quand l'élection d'un transmetteur désigné est initiée, l'état de démarrage est l'état Offer, le compteur de messages (MC) est réglé à zéro, et le temporisateur d'élection de DF (DFT) est réglé à OPlow (voir au paragraphe 3.6 la définition des valeurs de temporisateur).

Figure 3 : Automate à états d'élection de transmetteur désigné sous forme de tableau

	Événement		
État précédent	Meilleur Pass/Win reçu	Meilleur Backoff reçu	Meilleure offre
Offer	-> Lose DF = envoyeur ou cible ; arrête DFT	- DFT = BOperiod + OPlow ; MC = 0	- DFT = OPhigh ; MC = 0
Lose	- DF = envoyeur ou cible -> Lose	- DF = envoyeur -> Lose	-> Offer DFT = OPhigh ; MC = 0 -> Backoff
Win	DF = envoyeur ou cible ; arrête DFT	DF = envoyeur ; arrête DFT	Règle Meilleure à l'envoyeur ; envoi de Backoff ; DFT = BOperiod
Backoff	-> Lose ; DF = envoyeur ou cible ; arrête DFT	-> Lose DF = envoyeur ; arrête DFT	- Règle Meilleure à l'envoyeur ; envoi de Backoff ; DFT = BOperiod

	Événement			
État précédent	Backoff reçu pour nous	Pass reçu pour nous	Moins bon Pass/Win/Backoff reçu	Moins bonne offre reçue
Offer	- DFT = BOperiod + OPlow ; MC = 0	-> Win Arrête DFT	- Règle DF à envoyeur ou cible ; DFT ?= OPlow; MC = 0	- DFT ?= OPlow; MC = 0
Lose	-> Offer DF = envoyeur ; DFT = OPlow ; MC = 0	-> Offer DF = envoyeur ; DFT = OPlow ; MC = 0	-> Offer DF = envoyeur ou cible ; DFT = OPlow ; MC = 0	-> Offer DFT = OPlow ; MC = 0
Win	-> Offer DF = envoyeur ; DFT = OPlow ; MC = 0	-> Offer DF = envoyeur ; DFT = OPlow ;	-> Offer DF = envoyeur ou cible ; DFT = OPlow ; MC = 0	- envoi de Winner

		MC = 0		MC = 0 arrête DFT
Backoff	-> Offer DF = envoyeur ; DFT = OPlow ; MC = 0	-> Offer DF = envoyeur ; DFT = OPlow ; MC = 0	-> Offer DF = envoyeur ou cible ; DFT = OPlow ; MC = 0	-> Win envoi de Winner ; arrête DFT

Dans l'état Offer

DFT expire et MC est moins que Robustness	DFT expire et MC est égal à Robustness et on a un chemin pour la RPA	DFT expire et MC est égal à Robustness et il n'y a pas de chemin pour la RPA
-	-> Win	-> Lose
Envoi de Offer ; DFT = OPlow; MC = MC + 1	Envoi de Winner	Régler DF à Aucun

Dans l'état Offer

La métrique change et est maintenant moins bonne
DFT ?= OPlow
MC = 0

Dans l'état Lose

Détecte la défaillance du DF	La métrique change et est maintenant meilleure que celle du DF
-> Offer	-> Offer
DF = aucun ; DFT = OPlow_int ; MC = 0	DFT = OPlow_int ; MC = 0

Dans l'état Win

La métrique change et est maintenant moins bonne	Le temporisateur expire et MC est moins que Robustness	Perte du chemin de la RPA
-	-	-> Offer
DFT = OPlow ; MC = 0	Envoi de Winner; DFT = OPlow ; MC = MC + 1	Règle DF à Aucun ; MC = 0

Dans l'état Backoff

La métrique change et est maintenant meilleure que Best	Le temporisateur expire	Perte du chemin de la RPA
-> Win	-> Lose	-> Offer
Arrête le temporisateur	Envoi de Pass ; règle le DF au meilleur mémorisé	Règle le DF à aucun ; DFT = OPlow ; MC = 0

3.5.4 Améliorations de la fiabilité de l'élection

Pour le fonctionnement correct de BIDIR-PIM, il est très important d'éviter des situations où deux routeurs se considèrent comme étant les transmetteurs désignés pour la même liaison. Les deux précautions ci-dessous ne sont pas exigées pour le fonctionnement correct mais peuvent aider à diagnostiquer et corriger les anomalies.

3.5.5 Pass manquant

Après l'élection d'un DF, un routeur dont la métrique change pour devenir meilleure que celle du DF va tenter de l'emporter. Si durant la réélection le DF agissant a une condition qui lui fait perdre tous les messages d'élection (comme une surcharge de CPU) le nouveau candidat va transmettre trois offres et assumer le rôle de transmetteur, résultant en deux DF sur la liaison. Cette situation est pathologique et devrait être corrigée en réparant le routeur surchargé. Il est souhaitable qu'un tel événement puisse être détecté par un administrateur du réseau.

Quand un routeur devient le DF pour une liaison sans recevoir de message Pass du vieux DF connu, les informations de voisin de PIM pour le vieux DF peuvent être marquées à cet effet. À réception du prochain message PIM Hello du vieux DF, le routeur peut retransmettre des messages Winner pour toutes les RPA pour lesquelles il agit comme DF. L'anomalie peut aussi être enregistrée en débit limité par le routeur pour alerter l'opérateur.

3.5.6 Annonce périodique de gagnant

Un degré de sûreté supplémentaire peut être réalisé en faisant que le DF pour chaque RPA annonce périodiquement son

état dans un message Winner. La transmission de messages Winner périodiques peut être restreinte pour survenir seulement pour les RPA qui ont des groupes actifs, évitant donc le trafic de contrôle périodique dans des zones du réseau sans envoyeurs ou receveurs pour une RPA particulière.

3.6 Temporisateurs, compteurs, et constantes

BIDIR-PIM tient les temporisateurs suivants, comme exposé au paragraphe 3.1. Tous les temporisateurs sont à décompte - ils sont réglés à une valeur et décrémentés jusqu'à zéro, moment où normalement ils déclenchent une action. Bien sûr, ils peuvent tout aussi facilement être mis en œuvre comme des temporisateurs incrémentaires, où la valeur absolue du moment d'expiration est mémorisée et comparée à une horolge en temps réel, mais le langage de la présente spécification suppose qu'ils décomptent jusqu'à zéro.

Adresse par point de rendez-vous (RPA) :
par interface (I) : temporisateur d'élection de DF : DFT(RPA,I)
par groupe (G) : temporisateur de jonction en amont : JT(G)
par interface (I) :
temporisateur d'expiration de jonction : ET(G,I)
temporisateur d'élagage en cours : PPT(G,I)

Quand des temporisateur sont lancés ou relancés, ils sont réglés à leur valeur par défaut. Ces valeurs par défaut sont récapitulées ici.

Nom du temporisateur : temporisateur d'élection de DF (DFT)

Nom de valeur	**Valeur**	**Explication**
Offer_Period	100 ms	Intervalle d'attente entre les messages Offer et Winner répétés.
Backoff_Period	1 s	Période d'attente du DF agissant entre la réception d'une meilleure offre et l'envoi du message Pass pour le transfert de la responsabilité de DF.
OPlow	rand(0,5, 1) * Offer_Period	Gamme de la valeur aléatoire réelle utilisée entre messages répétés.
OPhigh	Election_Robustness * Offer_Period	Intervalle d'attente afin de donner une chance à un routeur avec une meilleure offre de devenir DF.

Nom du temporisateur : temporisateur d'expiration de jonction (ET(G,I))

Nom de valeur	**Valeur**	**Explication**
J/P HoldTime	d'après le message	Temps de garde après le message Join/Prune

Nom du temporisateur : temporisateur d'élagage en cours (PPT(G,I))

Nom de valeur	**Valeur**	**Explication**
Intervalle d'outrepassement de J/P	par défaut : 3 s	Courte période après un Join ou Prune pour permettre aux autres routeurs sur le LAN d'outrepasser le Join ou Prune.

Noter que la valeur de l'intervalle J/P Override est spécifique de l'interface et dépend des deux valeurs de Propagation_Delay *(délai de propagation)* et Override_Interval *(intervalle d'outrepassement)* qui peuvent changer quand des messages Hello sont reçus [RFC4601].

Nom du temporisateur : temporisateur de jonction en amont (JT(G))

Nom de valeur	**Valeur**	**Explication**
t_periodic	par défaut : 60 s	Période entre messages Join/Prune
t_suppressed	rand(1,1 * t_periodic, 1,4 * t_periodic)	Période de suppression quand quelqu'un d'autre envoie un message J/P de sorte qu'on a pas besoin de le faire.
t_override	rand(0, 0,9 * J/P Override Interval)	Délai aléatoire pour empêcher une explosion de réponses lors de l'envoi d'un message Join pour outrepasser le message Prune de quelqu'un d'autre.

Pour plus d'informations sur ces valeurs, se référer au document PIM-SM [RFC4601].

Nom de constante : Election_Robustness *(robustesse d'élection de DF)*

Nom de constante	**Valeur**	**Explication**
Election_Robustness	Par défaut : 3	Nombre minimum de messages d'élection qui doivent être perdus pour que l'élection échoue.

3.7 Formats de paquets BIDIR-PIM

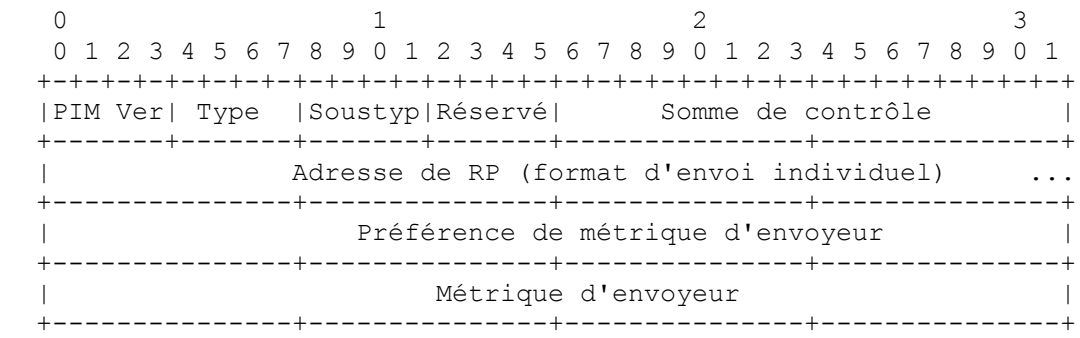
Ce paragraphe décrit les détails des formats de paquet pour les messages de contrôle BIDIR-PIM. BIDIR-PIM partage un certain nombre de messages de contrôle avec PIM-SM [RFC4601]. Cela inclut les messages Hello et Join/Prune ainsi que le format d'adresse d'envoi individuel. Pour les détails sur le format de ces paquets, on se réfèrera au document PIM-SM. On définit ici seulement les paquets supplémentaires introduits par BIDIR-PIM. Ce sont les paquets utilisés dans le processus d'élection de DF ainsi que dans l'option PIM-Hello de capacité bidirectionnelle.

3.7.1 Formats de paquets d'élection de DF

Tous les messages de contrôle PIM ont le numéro de protocole IP 103.

Les messages BIDIR-PIM sont en diffusion groupée avec le TTL 1 au groupe "ALL-PIM-ROUTERS" *(tous les routeurs PIM)*. Le groupe IPv4 "ALL-PIM-ROUTERS" est "224.0.0.13". Le groupe IPv6 "ALL-PIM-ROUTERS" est "ff02::d".

Tous les messages de contrôle d'élection de DF BIDIR-PIM partagent l'en-tête commun suivant :



PIM Ver : numéro de version PIM : 2.

Type : tous les messages de contrôle d'élection de DF PIM ont le type de message PIM de 10.

Sous type : les sous types pour les messages d'élection de DF sont :
- 1 = Offer
- 2 = Winner
- 3 = Backoff
- 4 = Pass

Réservé : réglé à zéro à l'émission, ignoré à réception.

Somme de contrôle : somme de contrôle IP standard, c'est-à-dire, le complément à un sur 16 bits de la somme des compléments à un du message PIM entier. Pour calculer la somme de contrôle, le champ Somme de contrôle est mis à zéro.

Adresse de RP : RPA bidirectionnelle pour laquelle l'élection a lieu. Le format est décrit au paragraphe 4.9.1 de la [RFC4601].

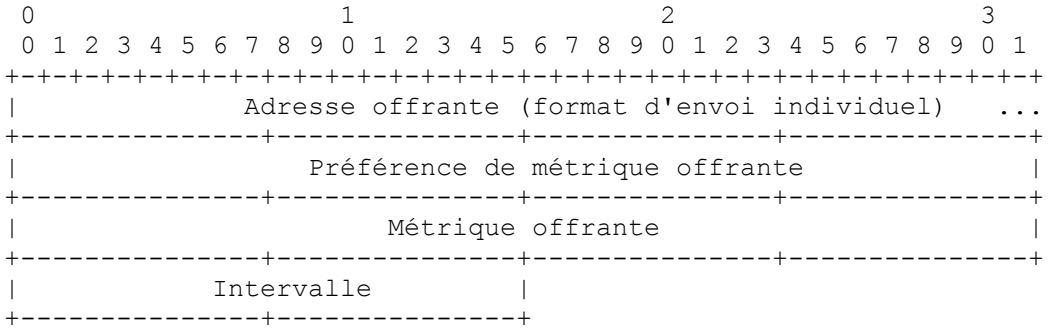
Préférence de métrique d'envoyeur : valeur de préférence allouée au protocole d'acheminement en envoi individuel que l'envoyeur du message a utilisé pour obtenir le chemin pour la RPA.

Métrique d'envoyeur : métrique du tableau d'acheminement en envoi individuel utilisée par l'envoyeur du message pour atteindre la RPA. La métrique est dans les unités applicables au protocole d'acheminement en envoi individuel utilisé.

En plus des champs définis ci-dessus, les messages Backoff et Pass ont les champs supplémentaires ci-après.

3.7.2 Message Backoff

Le message Backoff utilise les champs suivants en plus du format commun de message d'élection décrit ci-dessus.



Adresse offrante : adresse du routeur qui a fait la dernière (meilleure) offre. Le format est décrit au paragraphe 4.9.1 de la [RFC4601].

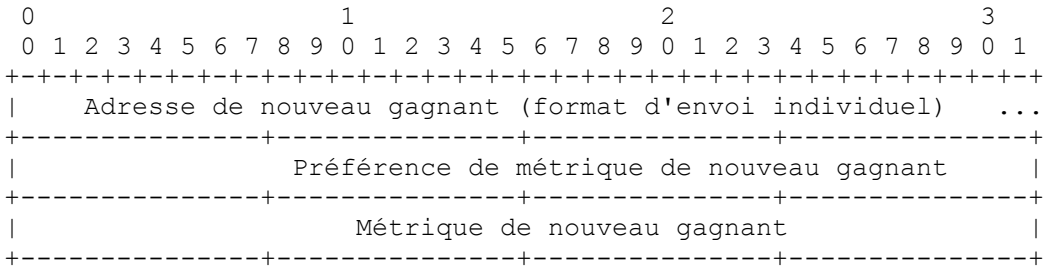
Préférence de métrique offrante : valeur de préférence allouée au protocole d'acheminement en envoi individuel que le routeur offrant a utilisé pour obtenir le chemin pour la RPA.

Métrique offrante : métrique du tableau d'acheminement en envoi individuel utilisée par le routeur offrant pour atteindre la RPA. La métrique est dans les unités applicables au protocole d'acheminement en envoi individuel utilisé.

Intervalle : intervalle de retard en millisecondes à utiliser par les routeurs de moins bonne métrique que le routeur offrant.

3.7.3 Message Pass

Le message Pass utilise les champs suivants en plus des champs commun d'élection décrits plus haut.



Adresse du nouveau gagnant : l'adresse du routeur qui a fait la dernière (meilleure) offre. Le format est décrit au paragraphe 4.9.1 de la [RFC4601].

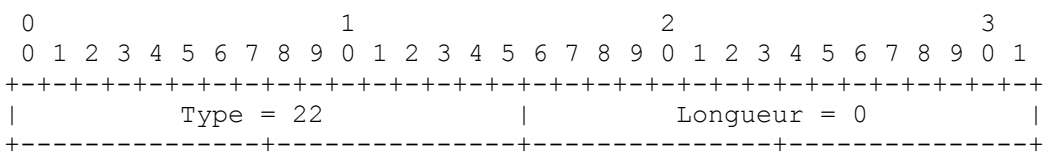
Préférence de métrique du nouveau gagnant : valeur de préférence allouée au protocole d'acheminement en envoi individuel que le routeur offrant a utilisé pour obtenir le chemin pour la RPA.

Métrique du nouveau gagnant : métrique du tableau d'acheminement en envoi individuel utilisée par le routeur offrant pour atteindre la RPA. La métrique est dans les unités applicables au protocole d'acheminement en envoi individuel utilisé.

3.7.4 Option PIM-Hello de capacité bidirectionnelle

BIDIR-PIM introduit une nouvelle option PIM-Hello.

Type d'option 22 : capacité bidirectionnelle



4. Découverte de point de rendez-vous

Les routeurs découvrent qu'une gamme d'adresses de groupe de diffusion groupée fonctionne en mode bidirectionnel, et que l'adresse du point de rendez-vous (RPA) dessert la gamme de groupes par configuration statique ou en utilisant un mécanisme automatique de découverte de RP comme le mécanisme d'amorçage PIM (BSR) [RFC5059] ou Auto-RP.

5. Considérations sur la sécurité

L'en-tête d'authentification IPsec [RFC2401] PEUT être utilisé pour assurer la protection de l'intégrité des données et l'authentification de l'origine des données au niveau du groupe des messages de protocole BIDIR-PIM. L'authentification des messages BIDIR-PIM peut protéger contre le comportement indésirable causé par des messages BIDIR-PIM non autorisés ou altérés.

5.1 Attaques fondées sur des messages falsifiés

Comme dans PIM en mode épars, l'étendue possible des dommages dépend du type de messages contrefaits acceptés. BIDIR-PIM utilise seulement des messages en diffusion groupée de liaison locale envoyés à l'adresse ALL_PIM_ROUTERS, donc les attaques peuvent seulement être effectuées par des nœuds directement connectés ou avec la complicité de routeurs directement connectés.

Certains des messages de protocole BIDIR-PIM (Join/Prune et Hello) sont identiques, en format et en fonctionnalité, aux messages respectifs utilisés dans PIM-SM. Les considérations pour la sécurité de ces messages se trouvent dans la [RFC4601]. Les autres messages (messages d'élection de DF) sont spécifiques de BIDIR-PIM et sont discutés dans les paragraphes suivants.

En falsifiant des messages d'élection de DF, un attaquant peut perturber l'élection du transmetteur désigné sur une liaison de deux façons différentes :

5.1.1 Élection d'un DF incorrect

Un attaquant peut forcer son élection comme DF en participant à une élection régulière et en annonçant la meilleure métrique pour atteindre la RPA. Un attaquant peut aussi essayer de forcer l'élection d'un autre routeur comme DF en envoyant un message Offer, Winner, ou Pass en se faisant passer pour un autre routeur. Dans certains cas (par exemple, Offer) plusieurs messages peuvent être nécessaires pour réaliser une attaque.

Dans le cas de messages Offer ou Winner, l'attaquant va devoir se faire passer pour le nœud qu'il veut faire devenir DF. Dans le cas de Pass, il va devoir se faire passer pour le DF actuel. Ce type d'attaque fait que le mauvais DF va être enregistré dans tous les nœuds à part celui qui fait l'objet de l'usurpation d'identité. Ce nœud va normalement être capable de détecter l'anomalie et, éventuellement, recommencer une nouvelle élection.

Un attaquant plus sophistiqué pourrait réaliser une attaque de déni de service concurrente sur le nœud objet de l'usurpation d'identité, afin qu'il ne soit pas capable de détecter les paquets falsifiés et/ou prendre des contre mesures.

Toutes les attaques fondées sur l'usurpation d'identité peuvent être détectées par tous les routeurs et évitées si la source des messages d'élection de DF peut être authentifiée. Quand l'authentification est disponible, les messages falsifiés DOIVENT être éliminés et un message d'avertissement en débit limité DEVRAIT être enregistré.

Un attaquant plus subtil pourrait utiliser des adresses de niveau MAC pour créer une partition de l'ensemble des receveurs de messages d'élection de DF et créer une vue de DF incohérente sur la liaison. Par exemple, l'attaquant pourrait utiliser des adresses MAC en envoi individuel pour ses messages d'élection de DF falsifiés. Pour empêcher ce type d'attaque, les routeurs BIDIR-PIM DEVRAIENT vérifier l'adresse MAC de destination des messages d'élection de DF reçus. Ceci est cependant inefficace sur des liaisons qui ne prennent pas en charge la livraison de diffusion groupé de couche 2.
'

L'authentification de la source est aussi suffisante pour prévenir cette sorte d'attaques.

5.1.2 Empêcher la convergence d'élection

En falsifiant des messages d'élection de DF, un attaquant peut empêcher l'élection de converger, perturbant ainsi l'établissement des arborescences de transmission de diffusion groupée. Il y a de nombreuses façons de réaliser cela. La plus simple est d'envoyer une séquence infinie de messages Offer (la métrique utilisée dans les messages n'a pas d'importance).

5.2 Mécanismes non cryptographiques d'authentification'

Un routeur BIDIR-PIM DEVRAIT fournir une option pour limiter l'ensemble de voisins desquels il va accepter des messages Join/Prune, Assert, et d'élection de DF. Une configuration statique des adresses IP ou une association de sécurité IPsec peut être utilisée. De plus, un routeur PIM NE DEVRAIT PAS accepter de messages de protocole d'un routeur dont il n'a pas encore reçu un message Hello valide.

5.2.1 Contrôle d'accès de base

Dans un domaine PIM-SM, quand tous les routeurs sont de confiance, il est possible de mettre en œuvre une forme basique de contrôle d'accès pour les sources et les receveurs: Les receveurs peuvent être validés par le DR de dernier bond et les sources peuvent être validées par le DR de premier bond et/ou le RP.

Dans BIDIR-PIM, c'est généralement faisable seulement pour les receveurs, car les sources peuvent envoyer au groupe de diffusion groupée sans qu'il soit besoin que les routeurs détectent leur activité et créent un état spécifique de source. Cependant, il est possible de modifier le comportement standard BIDIR-PIM d'une façon rétro compatible, pour permettre un contrôle d'accès par source. Le compromis serait entre la simplicité du protocole, la mémoire, et les exigences de traitement.

5.3 Authentification avec IPsec

Tout comme avec PIM-SM, le mode de transport IPsec [RFC2401] utilisant l'en-tête d'authentification (AH) est la méthode recommandée pour empêcher les attaques ci-dessus contre BIDIR-PIM.

Il est recommandé que l'authentification IPsec soit appliquée à tous les messages de protocole BIDIR-PIM. La spécification de la façon de le faire se trouve dans la [RFC4601]. Spécifiquement, l'authentification des messages de liaison locale PIM-SM, décrite dans la [RFC4601], s'applique aussi à tous les messages BIDIR-PIM.

5.4 Attaques de déni de service

L'attaque de déni de service fondée sur des messages Join falsifiés, décrite dans la [RFC4601], s'applique aussi à BIDIR-PIM.

6. Considérations relatives à l'IANA

L'IANA a alloué le type d'option 22 à l'option "Capacité bidirectionnelle".

7. Remerciements

La proposition bidirectionnelle du présent document est largement fondée sur les idées et le texte présentés par Estrin et Farinacci dans [BDST]. La principale différence entre les deux propositions est dans la méthode choisie pour la transmission vers l'amont.

Nous tenons aussi à remercier John Zwiebel de Cisco, Deborah Estrin de ISI/USC, Bill Fenner de AT&T Research, ainsi que Nidhi Bhaskar, Yiqun Cai, Toerless Eckert, Apoorva Karan, Rajitha Sumanasekera, et Beau Williamson de Cisco pour leurs contributions et commentaires à ce document.

8. Références normatives

[RFC2119] S. Bradner, "Mots clés à utiliser dans les RFC pour indiquer les niveaux d'exigence", BCP 14, mars 1997. *(MàJ par* RFC8174)

[RFC2401] S. Kent et R. Atkinson, "Architecture de sécurité pour le protocole Internet", novembre 1998. (*Obsolète, voir* RFC4301)

[RFC2710] S. Deering, W. Fenner et B. Haberman, "Découverte d'écouteur de diffusion groupée (MLD) pour IPv6", octobre 1999.

[RFC3376] B. Cain et autres, "Protocole Internet de gestion de groupe, IGMP version 3", octobre 2002. *(P.S.)*

[RFC4601] B. Fenner et autres, "Diffusion groupée indépendante du protocole - Mode épars (PIM-SM) : spécification du protocole (révisée)", août 2006. (*Remplace* RFC2362) (*MàJ par* RFC5059 ; *Remplacée par* RFC7761, STD83) *(P.S.)*

9. Références pour information

[BDST] Estrin, D. et D. Farinacci, "Bi-directional Shared Trees in PIM-SM", Travail en cours, mai 1999.

[RFC4760] T. Bates, R. Chandra, D. Katz et Y. Rekhter, "Extensions multi protocoles pour BGP-4", janvier 2007.

[RFC5059] N. Bhaskar et autres, "Mécanisme de routeur d'amorçage (BSR) pour la diffusion groupée indépendante du protocole (PIM)", janvier 2008. (*Remplace* RFC2362) (*MàJ* RFC4601) *(P.S.)*

Index

Adresse des auteurs

Mark Handley
Computer Science Department
University College London
mél : M.Handley@cs.ucl.ac.uk

Isidor Kouvelas
Cisco Systems
mél : kouvelas@cisco.com

Tony Speakman
Cisco Systems
mél : speakman@cisco.com

Lorenzo Vicisano
Digital Fountain
mél : lorenzo@digitalfountain.com

Déclaration complète de droits de reproduction