Groupe de travail Réseau
Request for Comments : 4391
Catégorie : Sur la voie de la normalisation
Traduction Claude Brière de L'Isle

J. Chu, Sun Microsystems
V. Kashyap, IBM
avril 2006

Transmission de IP sur InfiniBand (IPoIB)

Résumé
Le présent document spécifie une méthode pour encapsuler et transmettre des paquets IPv4/IPv6 et de protocole de résolution d'adresse (ARP, *Address Resolution Protocol*) sur InfiniBand (IB). Il décrit l'adresse de couche liaison à utiliser lors de la résolution des adresses IP dans des sous réseaux IP sur InfiniBand (IPoIB). Le document décrit aussi la transposition des adresses de diffusion groupée IP en adresses de diffusion groupée InfiniBand. De plus, ce document définit l'établissement et la configuration des liaisons IPoIB.

Table des matières

1. Introduction

La spécification InfiniBand [IBTA] se trouve à http://www.infinibandta.org. La [RFC4392] donne une brève vue d'ensemble de l'architecture InfiniBand (IBA) avec des considérations sur la spécification de IP sur les réseaux InfiniBand.

IBA définit plusieurs modes de transport sur lesquels IP peut être mis en œuvre. Le mode de transport de datagramme non fiable (UD, *Unreliable Datagram*) correspond le mieux aux besoins de IP et au besoin d'universalité décrit dans la [RFC4392].

Le présent document spécifie le mode UD de IPoIB sur IB. La mise en œuvre de sous réseaux IP sur d'autres mécanismes de transport de IB sort du domaine d'application du présent document.

Le présent document décrit les étapes nécessaires exigées pour poser un réseau IP au sommet d'un réseau IB. Il décrit tous les éléments d'une liaison IPoIB, comment configurer ses attributs associés, et comment établir pour elle les services de base de diffusion et de diffusion groupée.

Il décrit de plus la résolution d'adresse IP et l'encapsulation de paquets IP et de protocole de résolution d'adresse (ARP) dans une trame InfiniBand.

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 le BCP 14, [RFC2119].

2. IP sur mode UD

Le mode de communication de datagramme non fiable est pris en charge par tous les éléments IB qu'ils soient des routeurs adaptateurs de canal d'hôte (HCA, *Host Channel Adapter*) IB, des ou des adaptateurs de canal cible (TCA, *Target Channel Adapter*). En plus d'être la seule méthode de transmission universelle, il prend en charge la diffusion groupée, le partitionnement, et un contrôle de redondance cyclique (CRC, *Cyclic Redundancy Check*) de 32 bits [IBTA]. Bien que la prise en charge de la diffusion groupée soit facultative dans les tissus IB, l'architecture IPoIB exige que les composants participants la prennent en charge.

Toutes les mises en œuvre de IPoIB DOIVENT prendre en charge IP sur le mode de transport UD de IBA.

3. Liaison de données InfiniBand

Un sous réseau IB est formé par un réseau de nœuds IB interconnectés soit directement, soit via des commutateurs IB. Les sous réseaux IB peuvent être connectés en utilisant des routeurs IB pour former un tissu fait de plusieurs sous réseaux IB. Les nœuds résidant dans différents sous réseaux IB peuvent communiquer directement les uns avec les autres à travers les routeurs IB à la couche de réseau IB. Plusieurs sous réseaux IP peuvent être superposés à ce réseau IB.

Un sous réseau IP est configuré sur une facilité ou support de communication sur lequel les nœuds peuvent communiquer à la couche de "liaison" [RFC2460]. Par exemple, un segment Ethernet est une liaison formée par des commutateurs/concentrateurs/ponts interconnectés. Le segment est donc défini par la topologie physique du réseau. Ce n'est pas le cas avec IPoIB. Les sous réseaux IPoIB sont construits sur une "liaison" abstraite. La liaison est définie par ses membres et ses caractéristiques communes comme la P_Key, la MTU de liaison, et la Q_Key.

Deux accès quelconques utilisant le mode de communication UD dans un tissu IB ne peuvent communiquer que si ils sont dans la même partition (c'est-à-dire, ont les mêmes P_Key et Q_Key) [RFC4392]. La MTU de liaison donne la limite de taille de la charge utile qui peut être utilisée. La transmission et l'acheminement de paquet au sein du tissu IB sont aussi affectés par des paramètres supplémentaires tels que la classe de trafic *(TClass)*, la limite de bonds *(HopLimit)*, le niveau de service *(SL)*, et l'étiquette de flux *(FlowLabel)* [RFC4392]. La détermination et l'utilisation de ces valeurs pour la communication IPoIB sont décrites dans les sections suivantes.

4. Transposition de diffusion groupée

IB identifie les groupes de diffusion groupée par les identifiants mondiaux de diffusion groupée (MGID, *Multicast Global Identifier*) qui suivent les mêmes règles que les adresses de diffusion groupée IPv6. Donc, les MGID suivent les mêmes règles concernant les adresses transitoires et les bits de portée mais dans le contexte du tissu IB. L'adresse résultante ressemble donc aux adresses de diffusion groupée IPv6. Les documents [IBTA], [RFC4392] donnent une description

détaillée de la diffusion groupée IB.

La transposition de diffusion groupée IPoIB est décrite dans la Figure 1. La même fonction de transposition est utilisée pour IPv4 et IPv6 sauf pour le champ Signature IPoIB.

Sauf mention contraire explicite, toutes les adresses et champs dans les en-têtes de protocole de ce document sont mémorisés dans l'ordre des octets du réseau.

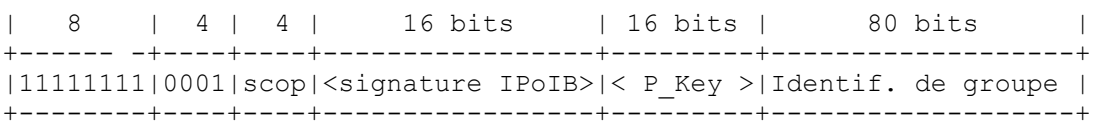

Figure 1

Comme un MGID alloué pour transporter des datagrammes en diffusion groupée IP est considéré seulement comme une adresse transitoire de couche de liaison de diffusion groupée [RFC4392], tous les MGID IB alloués pour les besoins de IPoIB DOIVENT établir le fanion T à 1 [IBTA].

Une signature spéciale est incorporée pour identifier le MGID pour la seule utilisation de IPoIB. Pour IPv4 sur IB, la signature DOIT être "0x401B". Pour IPv6 sur IB, la signature DOIT être "0x601B".

L'adresse de diffusion groupée IP est utilisée avec une P_Key de liaison IPoIB donnée pour former le MGID du groupe de diffusion groupée IB. Pour IPv6 les 80 bits inférieurs de l'identifiant de groupe sont utilisés directement dans les 80 bits inférieurs du MGID. Pour IPv4, l'identifiant de groupe est seulement de 28 bits, et est placé directement dans les 28 bits inférieurs du MGID. Les bits restants de l'identifiant de groupe dans le MGID sont remplis de 0.

Par exemple, sur une liaison IPoIB qui est entièrement contenue dans un seul sous réseau IB avec une P_Key de 0x8000, les MGID pour le groupe de diffusion groupée tous les routeurs avec l'identifiant de groupe [AARCH], [IGMP3] sont :

FF12:401B:8000::2, pour IPv4 en format compressé, et
FF12:601B:8000::2, pour IPv6 en format compressé.

Un cas particulier existe pour l'adresse de diffusion limitée IPv4 "255.255.255.255" [RFC1122]. L'adresse DOIT être transposée en le "GID de diffusion", qui est défini comme suit :

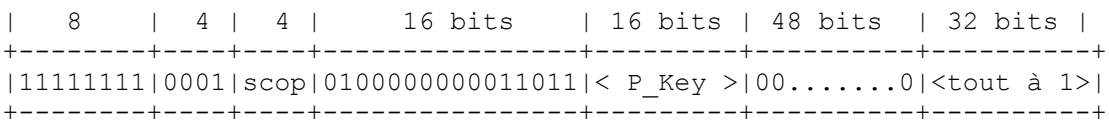

Figure 2

Tous les MGID utilisés dans le sous réseau IPoIB DOIVENT utiliser les mêmes bits de "scop" que dans le GID de diffusion correspondant.

4.1 Paramètres de GID de diffusion

Le GID de diffusion est établi avec les attributs suivants :

1. P_Key : une P_Key "Full Membership" *(membre à part entière)* (le bit de poids fort est réglé à 1) DOIT être utilisée afin que tous les membres puissent communiquer les uns avec les autres.

2. Q_Key : il est RECOMMANDÉ qu'une Q_Key contrôlée soit utilisée avec le bit de poids fort établi. C'est pour empêcher un logiciel non privilégié de fabriquer et envoyer des datagrammes IP bogués.

3. MTU IB : la valeur allouée au GID de diffusion ne doit pas être supérieure à toute MTU de liaison physique s'étendant sur le sous réseau IPoIB.

Les attributs suivants sont exigés dans les transmissions de diffusion groupée et aussi dans les transmissions en envoi

individuel si une liaison IPoIB couvre plus d'un seul sous réseau IB.

4. Autres paramètres : le choix des valeurs de TClass, FlowLabel, et HopLimit dépend de la mise en œuvre. Mais il doit prendre en compte la topologie des sous réseaux IB constituant la liaison IPoIB afin de permettre une communication réussie entre deux nœuds quelconques dans la même liaison IPoIB.

Un SL a aussi besoin d'être alloué au GID de diffusion. Ce SL est utilisé dans toutes les communications en diffusion groupée dans le sous réseau.

Les bits de portée du GID de diffusion doivent être établis sur la base de si la liaison IPoIB est confinée dans un sous réseau IB ou si la liaison IPoIB s'étend sur plusieurs sous réseaux IB. Une portée de sous réseau local par défaut (c'est-à-dire, 0x2) est RECOMMANDÉE. Un nœud peut déterminer les bits de portée à utiliser en cherchant de façon interactive une GID de diffusion de portée encore plus grande en commençant par la portée locale. Ou, une mise en œuvre peut inclure les bits de portée comme paramètre de configuration.

5. Établissement d'une liaison IPoIB

Le GID de diffusion, comme défini dans la section précédente, DOIT être établi pour qu'un sous réseau IPoIB soit formé. Chaque interface IPoIB DOIT se joindre comme "FullMember" au groupe de diffusion groupée IB défini par le GID de diffusion. Ce groupe de diffusion groupée va à partir de là être mentionné comme le groupe de diffusion. L'opération de jonction retourne la MTU, la Q_Key, et les autres paramètres associés au groupe de diffusion. Le nœud associe alors les paramètres reçus par suite de l'opération de jonction à son interface IPoIB. Le groupe de diffusion sert aussi à fournir un service de diffusion de couche de liaison pour des protocoles comme ARP, de diffusions dirigées sur le réseau, dirigées sur les sous réseaux, et dirigées sur tous les sous réseaux dans les réseaux IPv4 sur IB.

L'opération de jonction n'est réussie que si le gestionnaire de sous réseau (SM, *Subnet Manager*) détermine que le nœud qui se joint peut supporter la MTU enregistrée avec le groupe de diffusion [RFC4392] assurant la prise en charge d'une MTU de liaison commune. Le SM s'assure aussi que tous les nœuds qui se joignent au GID de diffusion ont des chemins les uns avec les autres et peuvent donc envoyer et recevoir des paquets en envoi individuel. Il s'assure de plus que tous les nœuds forment bien une arborescence de diffusion groupée qui permet que les paquets envoyés de tout membre soient reproduits à chaque autre membre. Donc, la liaison IPoIB est formée par les nœuds IPoIB qui se joignent au groupe de diffusion. Il n'y a pas de démarcation physique de la liaison IPoIB autre que celle déterminée par l'adhésion au groupe de diffusion.

La P_Key est un paramètre de configuration qui doit être connu avant que le GID de diffusion puisse être formé. Pour qu'un nœud se joigne à une partition, un de ses accès doit être alloué à la P_Key pertinente par le SM [RFC4392].

La méthode de création du groupe de diffusion et l'allocation/choix de ses paramètres relèvent de la mise en œuvre et/ou de l'administrateur du sous réseau IPoIB. Le groupe de diffusion peut être créé par le premier nœud IPoIB à être initialisé, ou il peut être créé administrativement avant l'établissement du sous réseau IPoIB. Il est RECOMMANDÉ que la création et suppression du groupe de diffusion soit sous contrôle administratif.

La gestion de la diffusion groupée InfiniBand, qui inclut la création, la jonction, et sortie des groupes de diffusion groupée IB par les nœuds IB, est décrite dans la [RFC4392].

6. Format de trame

Tous les datagrammes IP et ARP transportés sur InfiniBand sont préfixés d'un en-tête d'encapsulation de quatre octets comme illustré ci-dessous.

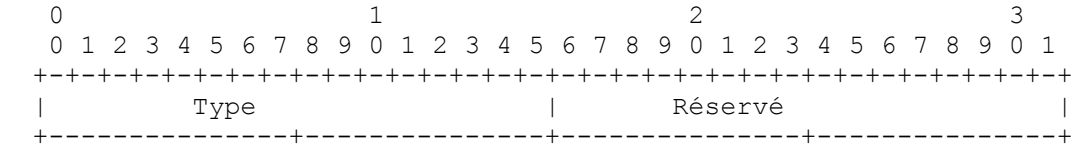

Figure 3

Le champ "Réservé" DOIT être réglé à zéro à l'envoi et ignoré à réception sauf spécification contraire dans un futur document.

Le champ "Type" DEVRA indiquer le protocole encapsulé conformément au tableau suivant.

Type	Protocole
0x800	IPv4
0x806	ARP
0x8035	RARP
0x86DD	IPv6

Tableau 1

Ces valeurs sont prises des numéros de "ETHER TYPE" alloués par l'IANA [IANA]. D'autres protocoles de réseau, identifiés par des valeurs différentes de "ETHER TYPE", peuvent utiliser le format d'encapsulation défini ici, mais une telle utilisation sort du domaine d'application de de document.

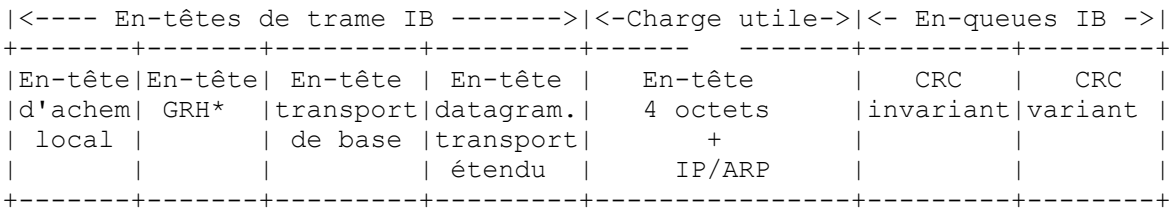

Figure 4

La Figure 4 décrit la trame IB encapsulant un datagramme IP/ARP. La spécification InfiniBand exige l'utilisation d'un en-tête d'acheminement mondial (GRH, *Global Routing Header*) [RFC4392] pour la diffusion groupée ou quand un paquet InfiniBand traverse d'un sous réseau IB à un autre à travers un routeur IB. Son utilisation est facultative pour la transmission en envoi individuel entre des nœuds au sein d'un sous réseau IB. La mise en œuvre IPoIB DOIT être capable de traiter les paquets reçus avec ou sans l'utilisation de GRH.

7. Unité maximum de transmission

MTU IB : les composants IB, c'est-à-dire, liaisons IB, commutateurs, adaptateurs de canaux (CA, *Channel Adapter*), et routeurs IB, peuvent prendre en charge des charges utiles maximum de 256, 512, 1024, 2048, ou 4096 octets. La charge utile maximum IB supportée par les composants IB dans tout chemin IB est la MTU IB pour le chemin.

MTU de liaison IPoIB : c'est la valeur de MTU associée au groupe de diffusion. La MTU de liaison IPoIB peut être réglée à toute valeur jusqu'à la plus petite MTU IB supportée par les composants IB constituant la liaison IPoIB.

Afin de réduire les problèmes de fragmentation et de découverte de MTU de chemin, le présent document exige que toutes les mises en œuvre de IPoIB prennent en charge une MTU de 2044 octets, c'est-à-dire, une MTU de liaison IPoIB de 2048 octets moins les frais généraux d'encapsulation de quatre octets. Des MTU plus grandes et plus petites PEUVENT être prises en charge sous réserve d'autres exigences de MTU existanres [RFC2460], mais la configuration par défaut doit prendre en charge une MTU de 2044 octets.

8. Autoconfiguration IPv6 sans état

L'architecture IB associe un identifiant EUI-64, appelé un identifiant unique au monde (GUID, *Globally Unique Identifier*) [RFC4392], [IBTA] à chaque accès. L'identifiant local (LID, *Local Identifier*) n'est unique qu'au sein d'un sous réseau IB.

L'identifiant d'interface peut être choisi à partir du :

1) GUID conforme à EUI-64 alloué par le fabricant ;

2) si le sous réseau IPoIB est entièrement contenu dans un sous réseau IB, tout LID unique de 16 bits de l'accès associé à l'interface IPoIB. Les valeurs de LID d'un accès peuvent changer après un réamorçage/cycle d'alimentation du nœud IB. Donc, si on désire une valeur persistante, il serait prudent de ne pas utiliser de LID pour former l'identifiant d'interface. Par ailleurs, le LID donne un identifiant qui peut être utilisé pour créer une adresse IPv6 plus anonyme car le LID n'est pas unique au monde et est sujet à changer avec le temps.

Il est RECOMMANDÉ que l'adresse de liaison locale soit construite à partir de l'identifiant EUI-64 de l'accès comme indiqué ci-dessous.

La [RFC3513] exige que l'identifiant d'interface soit créé dans le format "EUI-64 modifié" quand il est déduit d'un identifiant EUI-64. [IBTA] n'est pas clair sur la question de savoir si le GUID devrait utiliser le format IEEE EUI-64 ou le format "EUI-64 modifié". Donc, quand elle crée un identifiant d'interface à partir du GUID, une mise en œuvre DOIT faire ce qui suit :

=> Déterminer si le GUID est un identifiant EUI-64 modifié (le bit "u" est basculé) comme défini dans la [RFC3513]

=> Si le GUID est un identifiant EUI-64 modifié, alors le bit "u" NE DOIT PAS être basculé lors de la création de l'identifiant d'interface.

=> Si le GUID est un identifiant EUI-64 non modifié, alors le bit "u" DOIT être basculé en accord avec la [RFC3513]

8.1 Adresse IPv6 de liaison locale

L'adresse IPv6 de liaison locale pour une interface IPoIB est formée comme décrit dans la [RFC3513] en utilisant l'identifiant d'interface décrit au paragraphe précédent.

9. Transposition d'adresse – envoi individuel

La résolution d'adresse dans les sous réseaux IPv4 est accomplie par le protocole de résolution d'adresse (ARP, *Address Resolution Protocol*) [RFC0826]. Elle est accomplie dans les sous réseaux IPv6 en utilisant le protocle de découverte de voisin [RFC2461].

9.1 Informations de liaison

Un paquet InfiniBand sur le mode UD inclut plusieurs en-têtes comme l'en-tête de chemin local (LRH, *local route header*), l'en-tête de chemin mondial (GRH, *global route header*), l'en-tête de transport de base (BTH, *base transport header*), l'en-tête de transport de datagramme étendu (DETH, *datagram extended transport header*) comme décrit à la Figure 4 et spécifié dans l'architecture InfiniBand [IBTA]. Tous ces en-têtes constituent la couche de liaison d'une liaison IPoIB.

Les paramètres nécessaires dans ces en-têtes IBA constituent les informations de couche de liaison qui doivent être déterminées avant qu'un paquet IP puisse être transmis sur la liaison IPoIB.

Les paramètres qui doivent être déterminé sont les suivants :

a) LID : le LID est toujours nécessaire. Un paquet comporte toujours le LRH qui est ciblé sur le LID du nœud distant, ou sur le LID d'un routeur IB pour mettre le nœud distant dans un autre sous réseau IB.

b) Identifiant mondial (GID, *Global Identifier*) : le GID n'est pas nécessaire quand on échange des informations au sein d'un sous réseau IB bien qu'il puisse être inclus dans tout paquet. Il est d'une nécessité absolue lors de la transmission à travers le sous réseau IB car les routeurs IB utilisent le GID pour transmettre correctement les paquets. Les GID de source et de destination sont des champs inclus dans le GRH. Le GID, si il est formé en utilisant le GUID, peut être utilisé pour identifier sans ambiguïté un point d'extrémité.

c) Numéro de paire de file d'attente (QPN, *Queue Pair Number*) : chaque communication UD en envoi individuel est toujours dirigée sur une paire de file d'attente (QP, *queue pair*) particulière chez l'homologue.

d) Q_Key : une Q_Key est associée à chaque QPN de datagramme non fiable. Les paquets reçus doivent contenir une Q_Key qui correspond à la Q_Key de la PQ pour être acceptés.

e) P_Key : une communication réussie entre deux nœuds IB utilisant le mode UD ne peut se produire que si les deux nœuds ont des P_Key compatibles. Ceci est appelé être dans la même partition [IBTA].

f) SL : chaque paquet IBA contient une valeur de SL. Un chemin dans IBA est défini comme étant le triplet (LID de source, LID de destination, SL). Le SL à son tour est transposé en un chemin virtuel (VL, *virtual lane*) à chaque CA, commutateur qui envoie/transmet le paquet [RFC4392]. Plusieurs SL peuvent être utilisés entre deux points d'extrémité pour assurer l'équilibrage de charge. Les SL peuvent être utilisés pour fournir une infrastructure de qualité de service (QS) ou peuvent être utilisés pour éviter des impasses dans le tissu IBA.

Un autre élément auxiliaire d'information, non inclus dans les en-têtes IBA, est le suivant :

g) Débit de chemin : IBA définit plusieurs vitesses de liaison. Un émetteur de vitesse supérieure peut inonder les commutateurs et les CA. Pour éviter un tel encombrement, chaque source qui transmet à des vitesses supérieures à 1x doit déterminer le "débit de chemin" avant que les données puissent être transmises [IBTA].

9.1.1 Adresse de couche liaison /Adresse de matériel

Bien que la liste des informations exigées pour une transmission réussie d'un paquet IPoIB soit grande, toutes les informations n'ont pas besoin d'être déterminées durant le processus de résolution d'adresse IP.

L'adresse de couche liaison IPoIB de 20 octets utilisée dans l'option d'adresse de couche liaison de source/cible dans IPv6 et l'adresse de matériel dans IPv4/ARP ont le même format.

Ce format est décrit ci-dessous :

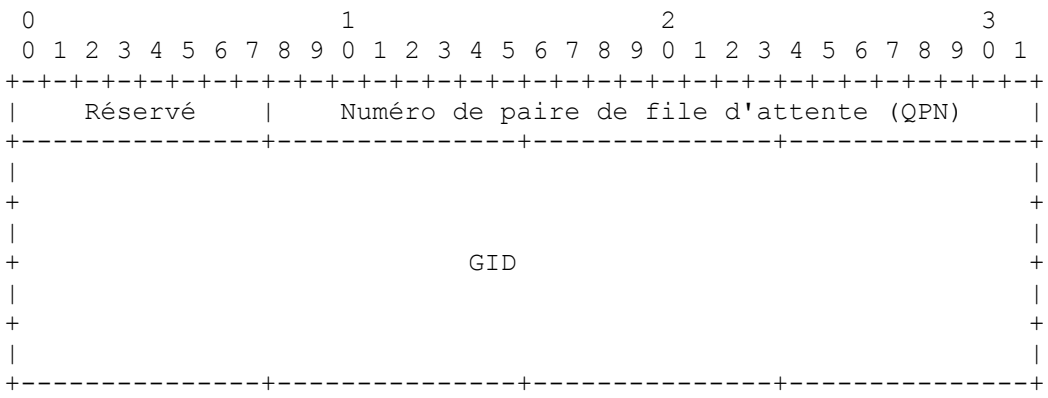

Figure 5

a) Fanions réservés : ces 8 bits sont réservé pour une utilisation future. Ces bits DOIVENT être réglés à zéro à l'envoi et ignorés à réception sauf spécification différente dans un futur document.

b) QPN : chaque communication en envoi individuel dans l'architecture IB est dirigée sur une QP spécifique [IBTA]. Ce numéro de QP est inclus dans la description de la liaison. Toutes les communications IP sur l'interface IPoIB pertinente DOIVENT être dirigées sur ce QPN. Dans le cas de sous réseaux IPv4, les paquets de réponse ARP sont aussi dirigés sur le même QPN.
Le choix du QPN pour les communications IP/ARP relève de la mise en œuvre.

c) GID : c'est un des GID de l'accès associé à l'interface IPoIB [IBTA]. IB associe plusieurs GID à un accès. Il est RECOMMANDÉ que le GID formé par la combinaison du préfixe de sous réseau IB et le "GUID d'accès" de l'accès [IBTA] soit inclus dans l'adresse de couche de liaison/de matériel.

9.1.2 Informations auxiliaires de liaison

Le reste des paramètres est déterminé comme suit :

a) LID : la méthode pour déterminer le LID de l'homologue n'est pas définie dans le présent document. Il appartient à la mise en œuvre d'utiliser toute méthode approuvée par l'IBA pour déterminer le LID de destination. Une de ces méthodes est d'utiliser le GID déterminé durant la résolution d'adresse, de restituer le LID associé à partir de

l'infrastructure d'acheminement IB ou de l'administrateur de sous réseau (SA, *Subnet Administrator*). Il est de la responsabilité de l'administrateur de s'assurer que le ou les sous réseaux IB ont une connexité d'envoi individuel entre les nœuds IPoIB. Le GID échangé entre deux points d'extrémité dans un message en diffusion groupée (ARP/ND) ne garantit pas l'existence d'un chemin d'envoi individuel entre les deux. Il peut y avoir plusieurs LID, et donc plusieurs chemins, entre les points d'extrémité. Les critères pour le choix des LID sortent du domaine d'application de ce document.

b) Q_Key : la Q_Key reçue en se joignant au groupe de diffusion DOIT être utilisée pour toute communication IPoIB sur cette liaison IPoIB particulière.

c) P_Key : la P_Key à utiliser dans le sous réseau IP n'est pas découverte mais est un paramètre de configuration.

d) SL : la méthode de détermination du SL n'est pas définie dans ce document. Le SL est déterminé par toute méthode approuvée dans l'IBA.

e) Débit de chemin : la mise en œuvre doit utiliser les méthodes IB pour déterminer le débit de chemin comme exigé.

9.2 Résolution d'adresse dans les sous réseaux IPv4

L'en-tête de paquet ARP est définie dans la [RFC0826]. Le type de matériel est réglé à 32 (décimal) comme spécifié par l'IANA [IANA]. Le reste des champs est utilisé conformément à la [RFC0826].

16 bits : type de matériel
16 bits : protocole
8 bits : longueur de l'adresse de matériel
8 bits : longueur de l'adresse de protocole
16 bits : opération ARP

Les champs restant dans le paquet contiennent les adresses de matériel et de protocole de l'envoyeur et de la cible.

[adresse de matériel de l'envoyeur]
[adresse de protocole de l'envoyeur]
[adresse de matériel de la cible]
[adresse de protocole de la cible]

L'adresse de matériel incluse dans le paquet ARP va être comme spécifiée au paragraphe 9.1.1 et décrite à la Figure 5.

La longueur de l'adresse de matériel utilisée dans l'en-tête d'un paquet ARP est donc 20.

9.3 Résolution d'adresse dans les sous réseaux IPv6

L'option Adresse de couche de liaison de source/cible est utilisée dans les messages Sollicitation de routeur, Annonce de routeur, Redirection, Sollicitation de voisin, et Annonce de voisin quand de tels messages sont transmis sur des réseaux InfiniBand.

L'option Adresse de source/cible est spécifiée comme suit :

Type :
1 : Adresse de couche de liaison de source
2 : Adresse de couche de liaison de cible

Longueur : 3

Adresse de couche de liaison : elle est comme spécifiée au paragraphe 9.1.1 et décrite à la Figure 5.
La [RFC2461] spécifie la longueur de l'option source/cible en nombre de 8 octets comme indiqué par une longueur de '3' ci-dessus. Comme l'adresse de couche liaison IPoIB est seulement longue de 20 octets, deux octets de zéro DOIVENT être ajoutés devant pour remplir la longueur totale d'option de 24 octets.

9.4 Note d'avertissement sur la mise en antémémoire de QPN

L'adresse de couche liaison pour IPoIB inclut le QPN, qui peut ne pas être constant après des réamorçages ou des réinitialisations d'interface réseau. Les entrées de QPN en antémémoires, comme des entrées statiques d'ARP ou dans des serveurs de protocole de résolution inverse d'adresse (RARP, *Reverse Address Resolution Protocol*) ne vont fonctionner que si la ou les mises en œuvre qui utilisent ces options s'assurent que le QPN associé à une interface est invariant à travers les réamorçages/réinitialisations de réseau.

Il est RECOMMANDÉ que les mises en œuvre revalident périodiquement les antémémoires ARP à cause de la volatilité sus-mentionnée induite par le QPN des adresses de couche de liaison IPoIB.

10. Envoi et réception des paquets IP en diffusion groupée

La diffusion groupée dans InfiniBand diffère d'un certain nombre de façons de la diffusion groupée dans Ethernet. Cela ajoute un peu de complexité à une mise en œuvre de IPoIB quand elle prend en charge la diffusion groupée IPoIB.

A) Un groupe de diffusion groupée IB doit être explicitement créé à travers le SA avant qu'il puisse être utilisé.
Cela implique qu'afin d'envoyer un paquet destiné à une adresse de diffusion groupée IP, la mise en œuvre IPoIB doit d'abord vérifier avec le SA sur la liaison de sortie qu'un "MCMemberRecord" correspond au MGID. Si il en existe un, l'identifiant local de diffusion groupée (MLID, *Multicast Local Identifier*) associé au groupe de diffusion groupée est utilisé comme identifiant local de destination (DLID, *Destination Local Identifier*) pour le paquet. Autrement, cela implique qu'aucun membre n'existe sur la liaison locale. Si la portée du groupe de diffusion groupée IP est au delà de la liaison locale, le paquet doit être envoyé aux routeurs sur la liaison en utilisant le groupe de diffusion groupée Tous les routeurs ou le groupe de diffusion. Ceci est pour permettre aux routeurs locaux de transmettre le paquet aux écoutants de diffusion groupée sur les réseaux distants. Le groupe de diffusion groupée Tous les routeurs est préféré au groupe de diffusion pour une meilleure efficacité. Si le groupe de diffusion groupée Tous les routeurs n'existe pas, l'envoyeur peut supposer qu'il n'y a pas de routeur sur la liaison locale ; donc, le paquet peut être éliminé en toute sécurité.

B) Un envoyeur de diffusion groupée doit se joindre au groupe de diffusion groupée cible avant de produire des messages sortants de diffusion groupée qui puissent être acheminés avec succès. La jonction "SendOnlyNonMember" *(en envoi seul pour les non membres)* est différente de la jonction régulière "FullMember" sous deux aspects. D'abord, les deux types de jonction permettent que les paquets de diffusion groupée soient acheminés à partir de l'accès local, mais seule la jonction "FullMember" cause l'acheminement des paquets de diffusion groupée à l'accès. Ensuite, l'accès d'envoyeur d'une jonction "SendOnlyNonMember" ne va pas être compté comme membre du groupe de diffusion groupée pour les besoins de création et suppression du groupe.

Le pseudo code suivant montre les étapes d'une mise en œuvre normale lors du traitement d'un paquet de diffusion groupée sortant.

```
si l'accès de sortie est déjà un "SendOnlyNonMember", ou a "FullMember"
    => envoyer le paquet

autrement, si le groupe de diffusion groupée cible existe
    => faire une jonction "SendOnlyNonMember"
    => envoyer le paquet

autrement, si portée > liaison locale ET si le groupe de diffusion groupée Tous les routeurs existe
    => envoyer le paquet à tous les routeurs

autrement
    => éliminer le paquet
```

Les mises en œuvre devraient mettre en antémémoire les informations sur l'existence d'un groupe de diffusion groupée IB, son MLID et ses autres attributs. C'est pour éviter de coûteux appels de SA sur chaque paquet de diffusion groupée sortant. Les envoyeurs DOIVENT s'abonner aux signes de création et suppression de groupe de diffusion groupée afin de surveiller l'état de groupes de diffusion groupée IB spécifiques. Par exemple, les paquets de diffusion groupée dirigés sur le groupe de diffusion groupée Tous les routeurs à cause de l'absence d'écoutants sur le sous réseau local doivent être transmis au bon groupe de diffusion groupée si le groupe est créé ultérieurement. Cela arrive quand un écoutant se manifeste sur le sous réseau local.

Un nœud qui se joint à un groupe de diffusion groupée IP doit d'abord construire un MGID en accord avec la règle décrite à la Section 4. Une fois le MGID correct calculé, le nœud doit appeler le SA de la liaison sortante pour tenter une jonction "FullMember" au groupe de diffusion groupée IB correspondant au MGID. Si le groupe de diffusion groupée IB n'existe pas déjà, il doit d'abord en être créé un avec la MTU de liaison IPoIB. Le MGID DOIT utiliser les mêmes P_Key, Q_Key, SL, MTU, et HopLimit que celles utilisées dans le GID de diffusion. Le reste des attributs DEVRAIT suivre aussi les valeurs utilisées dans le GID de diffusion.

La demande de jonction va causer l'ajout de l'accès local au groupe de diffusion groupée. Elle permet aussi au SM de programmer les commutateurs et routeurs IB avec les nouvelles informations de diffusion groupée pour assurer la transmission correcte des paquets de diffusion groupée pour le groupe.

Quand un nœud quitte un groupe de diffusion groupée IP, il DEVRAIT faire une demande de sortie "FullMember" au SA. Cela donne au SM une opportunité de mettre à jour les informations de transmission pertinentes, de supprimer un groupe de diffusion groupée IB si la partie locale est le dernier FullMember à partir, et de libérer le MLID qui lui est alloué. L'algorithme spécifique dépend de la mise en œuvre et sort du domaine d'application du présent document.

Noter que pour une liaison IPoIB qui s'étend sur plus d'un sous réseau IB connecté par des routeurs IB, un soutien adéquat de la transmission de diffusion groupée au niveau IB est exigé pour que les paquets de diffusion groupée atteignent les écoutants sur un sous réseau IB distant. Le mécanisme spécifique pour cela sort du domaine d'application de IPoIB.

11. Acheminement de diffusion groupée IP

L'acheminement de diffusion groupée IP exige que chaque interface sur laquelle le routeur fonctionne soit configurée à écouter toutes les adresses de diffusion groupée de couche de liaison générées par IP [RFC1112], [RFC2710]. Pour une interface Ethernet, cela est souvent réalisé en activant le mode de diffusion groupée de promiscuité sur l'interface.

IBA ne fournit pas de prise en charge matérielle pour le mode de diffusion groupée de promiscuité. Heureusement, un mode de diffusion groupée de promiscuité peut être émulé dans le logiciel fonctionnant sur un routeur en suivant les étapes ci-dessous :

A) Obtenir une liste de tous les groupes de diffusion groupée IB actifs du SA local.

B) Faire une demande de jonction "NonMember" au SA pour chaque groupe qui a une signature dans son MGID correspondant à celle pour IPv4 ou IPv6.

C) Souscrire aux événements de création de groupe de diffusion groupée IB en utilisant un MGID générique afin que le routeur puisse faire une jonction "NonMember" à tous les groupes de diffusion groupée IB créés utltérieurement pour IPv4 ou IPv6.

La jonction "NonMember" a le même effet qu'une jonction "FullMember" sauf qu'elle ne sera pas comptée comme un membre du groupe de diffusion groupée pour les besoins de création ou suppression de groupe. C'est-à-dire, quand le dernier "FullMember" quitte un groupe de diffusion groupée, le groupe peut être supprimé en toute sécurité par le SA sans se soucier des routeurs "NonMember".

12. Nouveaux types de vulnérabilité dans la diffusion groupée IB

De nombreuses fonctions de diffusion groupée IB sont sujettes à des défaillances dues à un certain nombre de possibles contraintes de ressources. Cela inclut la création de groupes de diffusion groupée IB, les appels de jonction ("SendOnlyNonMember", "FullMember", et "NonMember") et le rattachement d'une QP à un groupe de diffusion groupée.

En général, l'occurrence de ces conditions d'échec est très dépendante de la mise en œuvre, et est estimée être rare. Usuellement, un échec d'opération de diffusion groupée au niveau IB peut être propagé en arrière au niveau IP, causant l'échec de l'opération originale et la notification à l'initiateur de l'opération. Mais certaines fonctions de diffusion groupée IB ne sont liées à aucune opération de premier plan, rendant leurs défaillances difficiles à détecter. Par exemple, si un routeur de diffusion groupée IB tente une jonction "NonMember" à un groupe de diffusion groupée nouvellement créé dans le sous réseau local, mais si la jonction échoue, la transmission de paquets pour ce groupe de diffusion groupée particulier

va probablement échouer en silence, c'est-à-dire, sans que les envoyeurs de diffusion groupée locaux y prêtent attention. Ce type de problème peut ajouter de la vulnérabilité aux opérations déjà non fiables de diffusion groupée IP.

Les mises en œuvre DEVRAIENT enregistrer des messages d'erreur sur toute défaillance d'une opération de diffusion groupée IB. Les administrateurs de réseau devraient être conscients de cette vulnérabilité, et préserver assez de ressources de diffusion groupée aux points où la diffusion groupée IP va être lourdement utilisée. Par exemple, les HCA qui ont des ressources amples de diffusion groupée devraient être utilisés à tous les routeurs de diffusion groupée IB.

13. Considérations sur la sécurité

Le présent document spécifie la transmission IP sur un réseau de diffusion groupée. Tout réseau de cette sorte est vulnérable à un envoyeur qui prétend avoir une autre identité et au trafic falsifié ou à l'espionnage. Il est de la responsabilité des couches supérieures ou des applications de mettre en œuvre des contre mesures convenables si cela pose problème.

La réussite de la transmission des paquets IP dépend de l'établissement correct de la liaison IPoIB, de la création du GID de diffusion, de la création de la QP et de son rattachement au GID de diffusion, et de la détermination correcte des divers paramètres de la liaison comme le LID, le niveau de service, et le débit de chemin. Ces opérations, dont beaucoup impliquent des interactions avec le SM/SA, DOIVENT être protégées par le système de fonctionnement sous-jacent. C'est pour empêcher un logiciel malveillant, non privilégié de capturer d'importantes ressources et configurations.

Des Q_Key contrôlées DEVRAIENT être utilisées dans toutes les transmissions. C'est pour empêcher un logiciel non privilégié de fabriquer des datagrammes IP.

14. Considérations relatives à l'IANA

Pour la prise en charge de ARP sur InfiniBand, une valeur est exigée pour le paramètre de résolution d'adresse "Number Hardware Type (hrd)". L'IANA a alloué le numéro "32" pour indiquer InfiniBand [IANA_ARP].

Les futures utilisations des bits réservés dans le format de trame (Figure 3) et l'adresse de couche liaison (Figure 5) DOIVENT être publiées comme RFC. Le présent document exige que les bits réservés soient réglés à zéro à l'envoi et ignorés à réception.

15. Remerciements

Les auteurs tiennent à remercier Bruce Beukema, David Brean, Dan Cassiday, Aditya Dube, Yaron Haviv, Michael Krause, Thomas Narten, Erik Nordmark, Greg Pfister, Jim Pinkerton, Renato Recio, Kevin Reilly, Kanoj Sarcar, Satya Sharma, Madhu Talluri, et David L. Stevens de leurs suggestions et de beaucoup de précisions sur la spécification IBA.

16. Références

16.1 Références normatives

[IANA] Internet Assigned Numbers Authority, URL http://www.iana.org

[IANA_ARP] URL http://www.iana.org/assignments/arp-parameters

[IBTA] InfiniBand Architecture Specification, URL http://www.infinibandta.org/specs

[RFC0826] D. Plummer, "Protocole de résolution d'adresses Ethernet : conversion des adresses de protocole réseau en adresses Ethernet à 48 bits pour transmission sur un matériel Ethernet", STD 37, novembre 1982.

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

[RFC2461] T. Narten, E. Nordmark, W. Simpson, "Découverte de voisins pour IP version 6 (IPv6)", décembre 1998. (*Obsolète, voir* RFC4861) *(D.S.)*

[RFC3513] R. Hinden et S. Deering, "Architecture d'adressage du protocole Internet version 6 (IPv6)", avril 2003. (*Obs. voir* RFC4291)

[RFC4392] V. Kashyap, "Architecture de IP sur InfiniBand (IPoIB)", avril 2006. *(Information)*

16.2 Références pour information

[RFC1112] S. Deering, "Extensions d'hôte pour diffusion groupée sur IP", STD 5, août 1989. *(Mise à jour par la* RFC2236*)*

[RFC1122] R. Braden, "Exigences pour les hôtes Internet – couches de communication", STD 3, octobre 1989. *(MàJ par* RFC6633, 8029*)*

[RFC2460] S. Deering et R. Hinden, "Spécification du protocole Internet, version 6 (IPv6)", décembre 1998. *(MàJ par* 5095, 6564 *; D.S ; Remplacée par* RFC8200, STD 86*)*

[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.)*

Adresse des auteurs

H.K. Jerry Chu
17 Network Circle, UMPK17-201
Menlo Park, CA 94025
USA
téléphone : +1 650 786 5146
mél : jerry.chu@sun.com

Vivek Kashyap
15350, SW Koll Parkway
Beaverton, OR 97006
USA
téléphone : +1 503 578 3422
mél : vivk@us.ibm.com

Déclaration complète de droits de reproduction

droits de propriété qui pourraient couvrir les technologies qui peuvent être nécessaires pour mettre en œuvre la présente norme. Prière d'adresser les informations à l'IETF à ietf-ipr@ietf.org.

Remerciement
Le financement de la fonction d'édition des RFC est fourni par l'activité de soutien administratif de l'IETF (IASA).