Groupe de travail Réseau
Request for Comments : 4755
Catégorie : Sur la voie de la normalisation

V. Kashyap, IBM
décembre 2006
Traduction Claude Brière de L'Isle

IP sur InfiniBand : mode connecté

Résumé
Le présent document spécifie la transmission de paquets IPv4/IPv6 et la résolution d'adresse sur les modes connectés de InfiniBand.

Table des matières

1. Introduction

La spécification InfiniBand [IB_ARCH] se trouve à www.infinibandta.org. Le document [RFC4392] fournit une brève vue d'ensemble de l'architecture InfiniBand avec des considérations sur la spécification de IP sur des réseaux InfiniBand.

L'architecture InfiniBand (IBA, *InfiniBand Architecture*) définit plusieurs modes de transport. Parmi eux la méthode de transport de datagrammes non fiable (UD, *unreliable datagram*) correspond le mieux aux besoins de IP. IP sur InfiniBand (IPoIB) sur UD est décrit dans la [RFC4391]. Le présent document décrit la transmission IP sur les modes connectés de IBA.

IBA définit deux modes connectés :
1. Connecté fiable (RC, *Reliable Connected*)
2. Connecté non fiable (UC, *Unreliable Connected*)

Comme il est évident par leur désignation, les deux modes diffèrent principalement par la fourniture de la fiabilité de la livraison des données à travers la connexion. Le présent document s'applique également aux deux modes connectés. IPoIB sur ces deux modes est appelé IPoIB-CM (mode connecté) dans ce document. Pour être clair, IPoIB sur le mode de datagramme non fiable, comme décrit dans la [RFC4391] est appelé IPoIB-UD.

IBA exige que tous les adaptateurs de canal d'hôte (HCA, *Host Channel Adapter*) prennent en charge les modes connectés fiable et non fiable [IB_ARCH]. Il est facultatif pour les adaptateurs de canal cible (TCA, *Target Channel Adapter*) de prendre en charge les modes connectés.

Les modes connectés offrent des MTU de liaison jusqu'à 2^31 octets de long. Donc, l'utilisation des modes connectés peut offrir des avantages significatifs en prenant en charge des MTU raisonnablement grandes. Les modes datagramme de l'architecture InfiniBand (IBA) sont limités à 4096 octets.

La fiabilité est aussi améliorée si la caractéristique sous-jacente de "migration automatique de chemin" prise en charge par les modes connectés est utilisée.

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. Mode IPoIB connecté

IPoIB sur mode connecté est une extension FACULTATIVE de IPoIB-UD. Chaque mise en œuvre IPoIB DOIT prendre en charge la [RFC4391] et PEUT prendre en charge les extensions décrites dans le présent document.

Donc, l'encapsulation IP, la MTU par défaut, le format d'adresse de couche liaison, et le mécanisme d’auto configuration sans état IPv6 s'appliquent à IPoIB-CM exactement comme décrit dans la [RFC4391].

2.1 Diffusion groupée

Les modes connectés de IBA définissent un réseau non de diffusion à plusieurs accès. Les modes connectés de IBA ne prennent pas en charge la diffusion groupée bien que chaque nœud puisse communiquer avec chaque autre nœud si désiré.

Cela exige que la diffusion groupée soit émulée d'une certaine façon par le réseau. Cependant, dans le cas d'un réseau InfiniBand, au lieu d'une émulation, une paire de file d'attente (QP, *queue pair*) de datagrammes non fiables (UD, *unreliable datagram*) peut être utilisée pour prendre en charge la diffusion groupée tandis que le mode QP connecté est utilisé pour le trafic en envoi individuel. Comme il est exigé de chaque mise en œuvre de IPoIB qu'elle prenne en charge le mode UD, chaque mise en œuvre qui prend en charge IPoIB-CM va être capable d'utiliser la QP IPoIB-UD pré-existante pour toutes les communications en diffusion/diffusion groupée. La transposition de diffusion groupée, la transmission, et la réception des paquets en diffusion groupée et l'acheminement de diffusion groupée DOIVENT utiliser la QP UD associée à l'interface IPoIB.

2.2 Présentation de la résolution d'adresse

Chaque interface IPoIB-CM DOIT être associée à deux jeux de QP :
1) une QP de datagrammes non fiables,
2) une ou plusieurs QP en mode connecté.

La [RFC4391] décrit la méthode de résolution d'adresse pour déterminer l'adresse de liaison de l'homologue. Cette réponse est reçue sur la QP UD associée à l'interface IPoIB.

2.3 Description de l'établissement de connexion

Une fois l'adresse de liaison du nœud distant connue, une connexion IB doit être établie entre les nœuds avant qu'une communication IP puisse se produire.

Pour faire une connexion, l'envoyeur doit connaître l'identifiant de service à utiliser dans la demande [IB_ARCH]. Il doit aussi fournir la paire de file d'attente en "mode connexion" au nœud distant. L'homologue répond avec sa paire de file d'attente. Chaque connexion IB est d'homologue à homologue et utilise une QP en mode connecté à chaque extrémité.

Bien que la résolution d'adresse se produise au niveau d'une adresse IP individuelle, la connexion entre les nœuds est à la couche IB. Donc, chaque résolution d'adresse individuelle n'implique pas une nouvelle connexion entre les homologues.

3. Résolution d'adresse

Les interrogations de résolution d'adresse sont envoyées sur l'identifiant de groupe de diffusion "broadcast-GID" (Broadcast-Group Identifier) sur la QP UD associée à l'interface IPoIB [RFC4391]. Une réponse en envoi individuel est reçue sur la QP UD.

3.1 Adresse de couche de liaison

L'encapsulation IPoIB [RFC4391] décrit l'adresse de couche de liaison comme suit :

<1 octet réservé>:QP: GID

Le présent document étend l'adresse de couche de liaison comme suit :

<Fanions>:QPN:GID

Fanions : c'est un champ d'un seul octet. Les bits indiquent les modes connectés supportés par l'interface.

Le bit 0 spécifie la prise en charge du mode connecté fiable (RC, *reliable connected*). Le bit 1 indique la prise en charge du mode connecté non fiable (UC, *unreliable connected*). Tous les autres bits de l'octet sont réservés et DOIVENT être réglés à 0 à l'émission et ignorés à réception. Le format des fanions est le suivant :

```
+--+--+--+--+--+--+--+--+
|RC|UC| 0| 0| 0| 0| 0| 0|
+--+--+--+--+--+--+--+--+
```

RC et UC PEUT être tous deux établis en même temps si l'interface prend en charge les deux modes. Comme le mode IPoIB-UD est toujours pris en charge, il n'y a pas de fanion pour indiquer la prise en charge de IPoIB-UD.

Si IPoIB-CM n'est pas pris en charge, c'est-à-dire, si la mise en œuvre prend seulement en charge IPoIB-UD, la mise en œuvre DOIT alors ignorer le champ Fanions à réception. Elle DOIT régler l'octet <Fanions> tout à zéro à l'émission comme spécifié dans la [RFC4391].

QPN *(queue-pair number)* : numéro de paire de file d'attente sur laquelle les réponses en envoi individuel de résolution d'adresse vont être reçues [RFC4391]. Une interface IPoIB a seulement une QP UD associée qu'elle prenne ou non en charge cette extension.

Le QPN sert aussi un autre objet : il est utilisé pour former l'identifiant de service qui sert à établir la connexion IB.

À réception de la demande de résolution d'adresse de diffusion/diffusion groupée, le receveur répond par sa propre adresse de liaison, incluant le QPN UD associé et les fanions appropriés.

La réponse du receveur est renvoyée en envoi individuel à l'envoyeur après que le receveur a, dans le cas de IPoIB-UD, résolu le GID en identifiant local (LID), et déterminé les autres paramètres requis [RFC4391]. Une fois la résolution d'adresse achevée, la connexion IB sous-jacente peut être établie sur les modes de connexion pris en charge. Il n'est pas EXIGÉ d'une mise en œuvre qu'elle établisse une connexion simplement parce que l'homologue indique cette capacité. La décision de faire une telle connexion relève de la mise en œuvre.

3.2 Établissement de connexion IB

Une fois la résolution d'adresse achevée, la connexion IB peut être établie par l'un ou l'autre des homologues. Pour établir une connexion, des datagrammes de gestion InfiniBand (MAD, *IB Management Datagram*) sont envoyés au gestionnaire de communication (CM, *communication manager*) de l'homologue. La demande de connexion contient toujours un identifiant de service pour que l'homologue associe la demande au service approprié. Si la demande est acceptée, l'homologue retourne le QPN de mode connecté pertinent dans la réponse MAD. Le format des messages de connexion au CM et le processus d'établissement de la connexion IB sont décrits dans [IB_ARCH]. La prise de contact globale est de la forme :

```
REQ ---->
   <---- REP [ou REJ(rejet)]
RTA ---->
[ou REJ(rejet)]
```

Les messages de CM incluent, entre autres paramètres, l'identifiant de service, le QPN de mode de connexion locale, et la taille de charge utile à utiliser sur la connexion.

Note : la connexion IB est établie en utilisant l'identifiant de service comme défini au paragraphe 3.5. Le nœud DOIT garder un enregistrement des connexions IB auxquelles il participe. Le nœud PEUT tenter une autre connexion avec l'homologue distant en utilisant le même identifiant de service qu'utilisé pour une connexion IB existante. De même, le receveur d'une telle connexion PEUT rejeter la demande avec une indication d'erreur convenable dans la réponse au CM. La décision d'accepter ou d'initier plusieurs connexions à partir de ou vers une interface IPoIB relève de la mise en œuvre.

Le nœud qui a initié la connexion a connaissance de l'adresse IP du nœud cible, comme décrit ci-dessus. Le nœud qui reçoit la demande de connexion IB, ne peut cependant pas déterminer l'adresse de liaison du nœud initiateur. Pour permettre cette détermination, chaque message CM échangé dans l'établissement de la connexion IB DOIT inclure le QPN IPoIB-UD de l'envoyeur dans le champ "Données privées" [IB_ARCH]. Le QPN IPoIB-UD DOIT aussi être inclus dans tous les messages "REJ" [IB_ARCH].

3.3 Connexions IB simultanées

Pour s'assurer que deux connexions IB ne sont pas établies entre les homologues du fait du croisement de demandes *(REQ)* les règles suivantes DOIVENT être suivies :

Le receveur forme l'adresse de couche de liaison du nœud distant en utilisant le QPN UD reçu dans le champ "Données privées" du message "REQ" et le GID de l'envoyeur inclus dans le message "REQ". L'adresse de couche de liaison est utilisée pour déterminer si il y a déjà une demande de connexion "REQ" en instance envoyée par l'interface locale à l'adresse de couche de liaison reçue donnée. Si une telle demande en cours est déterminée, alors les deux adresses (locale et distante) de couche de liaison sont comparées numériquement. Si l'adresse de couche de liaison locale est numériquement inférieure, la connexion est alors acceptée, et sinon rejetée. Le code d'erreur dans le MAD "REJ" est réglé à "Rejet du consommateur" [IB_ARCH].

Note : les adresses de couche de liaison formées pour comparaison mettent à zéro les fanions de mode de connexion spécifiés au paragraphe 3.1. La comparaison est effectuée de l'octet de poids fort à l'octet de moindre poids de l'adresse de couche de liaison.

Ce qui est ci-dessus tient même si le receveur prend en charge plusieurs connexions IB avec le même homologue. C'est pour s'assurer qu'une seule connexion de plus est établie quand les messages "REQ" passent.

3.4 Suppression de connexion IB IPoIB-CM

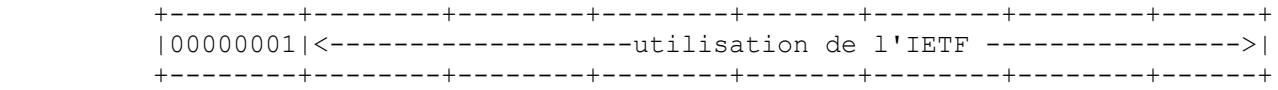
Les connexions IB créées par IPoIB-CM sont considérées faire partie d'une interface IPoIB. À ce titre, elles DEVRAIENT être supprimées quand les interfaces IPoIB auxquelles elles sont associées sont supprimées.

De plus, la connexion IB entre deux homologues PEUT être supprimée par l'un ou l'autre homologue chaque fois que l'entrée de résolution d'adresse expire. Une mise en œuvre est libre de mettre en œuvre d'autres politiques pour supprimer

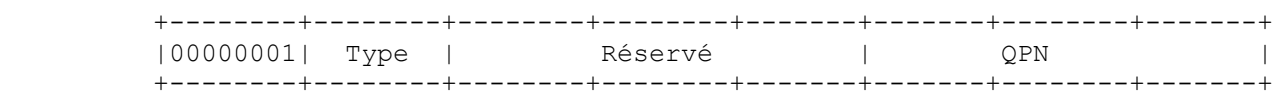
des connexions IB entre des homologues.

3.5 Identifiant de service

La spécification InfiniBand définit un bloc d'identifiants de service à l'usage de l'IETF. La spécification InfiniBand a laissé la définition et la gestion de ce bloc à l'IETF [IB_ARCH]. Le bloc de 64 bits est comme suit :



Les identifiants de service utilisés par IPoIB vont être du format suivant :



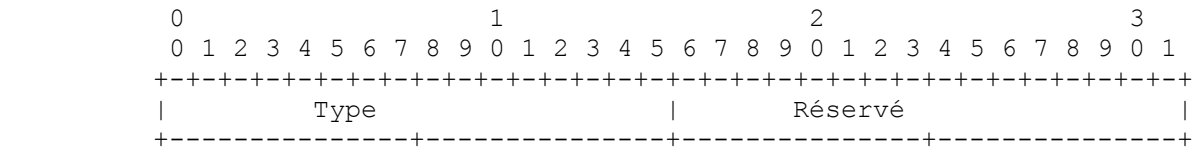
Le champ "Type" DOIT être réglé à 0.

Le champ "Réservé" DOIT être réglé tout à zéros.

Le QPN DOIT être le QP UD échangé durant la résolution d'adresse.

4. Format de trame

Tous les datagrammes IP transportés sur InfiniBand sont préfixés par un en-tête d'encapsulation de quatre octets comme décrit dans la [RFC4391].



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

Type	**Protocole**
0x800	IPv4
0x86DD	IPv6

Ces valeurs sont prises des numéros de "ETHER TYPE" allouées par l'IANA. D'autres protocoles 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 du présent document.

5. Unité de transmission maximum

L'établissement de la connexion IB pourrait être utilisé pour IPv4 et IPv6 ou il pourrait être utilisé seulement pour un des deux tandis qu'une connexion différente serait utilisée pour l'autre. La MTU de liaison DOIT être capable de supporter la MTU minimum requise par le protocole.

La MTU par défaut de l'interface IPoIB-CM est de 2044 octets, c'est-à-dire, 2048 octets de MTU de liaison IPoIB moins les 4 octets d'en-tête d'encapsulation.

Cependant, les modes connectés de InfiniBand permettent des tailles de message jusqu'à 2^31 octets. Donc, IPoIB-CM peut utiliser une MTU bien plus grande pour une communication en envoi individuel entre deux points d'extrémité quelconques. La charge utile maximum et/ou optimale qui peut être reçue ou envoyée sur une connexion InfiniBand dépend de la mise en œuvre, de l'adaptateur de canal IB, et des ressources configurées.

Une mise en œuvre PEUT utiliser le mécanisme suivant pour échanger la taille optimale de message à travers la connexion IB.

5.1 MTU par connexion

Chaque message d'établissement de connexion IB comporte un champ "Données privées" [IB_ARCH]. Le champ "Données privées" dans le message d'établissement de connexion (CM REQ) DOIT inclure la "MTU de réception". Cela indique la taille maximum de paquet que le demandeur peut accepter. Le demandeur DOIT aussi être capable d'accepter de plus petites tailles de MTU.

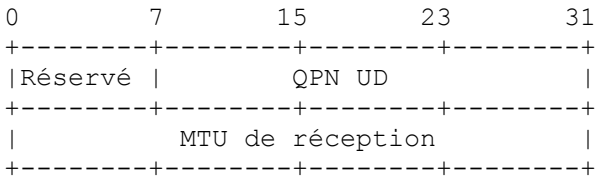
Il appartient à la mise en œuvre d'utiliser ce mécanisme pour régler la MTU par connexion IB. Pour calculer la MTU IPoIB résultante sur la connexion, la plus petite des valeurs des deux "MTU de réception" IB est utilisée par les deux homologues. L'interface IPoIB doit aussi tenir compte des quatre octets de l'en-tête d'encapsulation et donc la MTU IPoIB sur la connexion va être encore réduite de ce montant.

6. Format de données privées

Le champ "Données privées" dans chaque message CM pour l'établissement de connexion doit inclure les valeurs suivantes :

1. QPN UD de l'envoyeur
2. MTU de réception supportée par l'envoyeur

Le format du champ "Données privées" DOIT être comme suit :



La valeur Réservé DOIT être réglée à zéro à l'émission et ignorée à réception.

7. Considérations sur IPoIB-CM

Chaque interface IPoIB prend en charge IPoIB-UD. Elle peut de plus prendre en charge un des modes IPoIB-CM ou les deux. Donc, il peut y avoir plusieurs méthodes de communication entre deux homologues quelconques. Cela implique qu'une interface PEUT transmettre/recevoir un paquet sur tout mode RC, UC, ou UD selon les modes pris en charge entre elle et l'homologue. Il s'ensuit de plus que chaque mise en œuvre IPoIB conforme au présent document DOIT accepter toutes les transmissions IP en envoi individuel sur tous les modes IPoIB qu'elle supporte. Les paquets en diffusion groupée et en diffusion vont par leur nature même toujours être transmis et reçus sur la QP IPoIB-UD. De plus, toutes les réponses de résolution d'adresse (ARP ou découverte de voisin) DOIVENT toujours être encapsulées dans un paquet en mode UD.

7.1 Note de précaution sur IPoIB-RC

Le mode RC de InfiniBand garantit une livraison dans l'ordre des paquets. Chaque message transmis sur la connexion RC est coupé en paquets de la taille de la MTU physique par la connexion RC. Si un paquet est perdu, il est retransmis jusqu'à ce que le message complet soit échangé. Donc, il y a une possibilité qu'une couche supérieure de transport subisse une fin de temporisation, alors que la couche RC est encore dans le processus de transfert du message complet. TCP va voir la fin de temporisation comme un indicateur d'encombrement et entrer en démarrage lent affectant ainsi sévèrement le débit [RFC2581]. D'autres protocoles de couche supérieure pourraient insérer des retransmissions dans le tissu, ajoutant à l'encombrement déjà existant.

L’applicabilité de la fiabilité de InfiniBand est sur un tissu avec de faibles latences (pas de grandes zones). Donc, les valeurs de temporisateur RC devraient être courtes comparées aux valeurs de temps minimum de démarrage utilisées par

les transports de bout en bout supérieurs. De plus, parce que le mode RC n'a pas de transmission fiable fondée sur des mesures, son utilisation sur des tissus qui ont de longues latences ou des latences très dynamiques peut être un problème pour le trafic sensible à l'encombrement qui traverse ces tissus.

7.2 MTU par destination IPoIB-CM

Comme décrit ci-dessus, les interfaces sur le même sous réseau peuvent prendre en charge des MTU de liaison différentes sur la base de la valeur négociée ou à cause du type de liaison (mode UD ou connecté). Donc, une mise en œuvre pourrait choisir de définir une grande MTU IP, qui serait réduite sur la base de la MTU à la destination. La MTU pertinente peut être mémorisée dans un objet convenable par destination, comme une antémémoire de chemin ou de voisin. La MTU par destination est connue de l'interface IPoIB-CM comme décrit à la Section 5.

Les mises en œuvre pourraient choisir de ne pas prendre en charge des valeurs de MTU qui diffèrent et de toujours prendre en charge une MTU égale à la MTU IPoIB-UD déterminée à partir du GID de diffusion.

8. Considérations sur la sécurité

Un imposteur peut retourner un faux ensemble de fanions à une interface IPoIB. Cela peut causer des tentatives non nécessaires et des délais/interruptions dans la communication IPoIB. Il en est de même dans le cas de valeurs mauvaises ou parasites de QPN fournies durant la résolution d'adresse de diffusion/diffusion groupée.

9. Considérations relatives à l'IANA

Les futures utilisation des bits et octets réservés dans l'adresse de couche de liaison (paragraphe 3.1), d'identifiant de service (paragraphe 3.5), et de "Format de données privées" (Section 6) DOIVENT être publiées comme des RFC. Le présent document exige que les bits réservés soient réglés à zéro à l'émission.

10. Remerciements

L'auteur remercie le groupe de travail IPoIB de ses divers commentaires et suggestions. Un merci particulier à Bernie King-Smith et Dror Goldenberg pour leur relecture détaillée et leurs suggestions.

9. Références normatives

[IB_ARCH] "InfiniBand Architecture Specification, version 1.2". www.infinibandta.org

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

[RFC4391] J. Chu, V. Kashyap, "Transmission de IP sur InfiniBand (IPoIB)", avril 2006. *(P.S.)*

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

12. Références pour information

[RFC2581] M. Alman, V. Paxson et W. Stevens, "Contrôle d'encombrement avec TCP", avril 1999. *(Obsolète, voir* RFC5681)

Adresse de l'auteur

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

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