Groupe de travail Réseau
Request for Comments : 5040
Catégorie : Sur la voie de la normalisation
Traduction Claude Brière de L'Isle

R. Recio, IBM Corporation
B. Metzler, IBM Corporation
P. Culley, Hewlett-Packard Company
J. Hilland, Hewlett-Packard Company
D. Garcia
octobre 2007

Spécification d'un protocole d'accès direct à une mémoire distante

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.

Résumé

Le présent document définit un protocole d'accès direct à une mémoire distante (RDMAP, *Remote Direct Memory Access Protocol*) qui fonctionne sur le protocole de placement direct des données (DDP, *Direct Data Placement*). RDMAP fournit directement des services d'écriture et de lecture aux applications et permet le transfert direct des données dans les antémémoires de protocole de couche supérieure (ULP, *Upper Layer Protocol*) sans copies intermédiaires des données. Il permet aussi une mise en œuvre qui saute le noyau.

Table des Matières

1. Introduction....2
1.1. Buts architecturaux....2
1.2 Vue d'ensemble du protocole....3
1.3 Mise en couches de RDMAP....4
2. Glossaire....5
2.1 Général....5
2.2 LLP....6
2.3. Placement direct des données (DDP, Direct Data Placement)....6
2.4 Accès direct à la mémoire distante (RDMA, Remote Direct Memory Access)....7
3. Attributs de couche supérieure et de transport....9
3.1 Exigences et hypothèses de transport....9
3.2 Interactions de RDMAP avec la couche supérieure....9
4. Format d'en-tête....10
4.1 Champ contrôle RDMAP et STag Invalidate....11
4.2 Définition des messages RDMA....12
4.3 En-tête RDMA Write....12
4.4 En-tête Demande RDMA Read....12
4.5 En-tête Réponse RDMA Read....13
4.6 En-tête Send et Send avec événement sollicité....13
4.7 En-tête Send avec Invalidate et Send avec SE et Invalidate....14
4.8 En-tête Terminate....14
5. Transfert des données....16
5.1 Message RDMA Write....16
5.2 Opération RDMA Read....17
5.3 Type de message Send....18
5.4 Message Terminé....19
5.5 Rangement et achèvement....19
6. Gestion de flux RDMAP....21
6.1 Initialisation du flux....21
6.2 Suppression de flux....22
7. Gestion d'erreur RDMAP....23
7.1. Nettoyage d'erreurs RDMAP....23
7.2 Erreurs détectées chez l'homologue distant sur les messages RDMA entrants....24
8. Considérations sur la sécurité....24

8.1 Résumé des exigences de sécurité spécifiques de RDMAP....24
8.2 Services de sécurité pour RDMAP....26
9. Considérations relatives à l'IANA....27
10. Références....27
10.1 Références normatives....27
10.2 Références pour information....28
Appendice A. Formats de segment DDP pour les messages RDMA....28
A.1 Segment DDP pour RDMA Write....28
A.2. Segment DDP pour demande RDMA Read....29
A.3 Segment DDP pour réponse RDMA Read....29
A.4 Segment DDP pour Send et Send avec événement sollicité....30
A.5. Segment DDP pour Send avec Invalidate et Send avec SE et Invalidate....30
A.6 Segment DDP pour Terminate....31
Appendice B. Tableau d'ordre et d'achèvement....31
Appendice C. Contributeurs....32
Adresse des auteurs....33
Déclaration complète de droits de reproduction....33

1. Introduction

Aujourd'hui, les communications sur TCP/IP exigent normalement des opérations de copie, qui ajoutent de la latence et consomment des ressources significatives de CPU et de mémoire. Le protocole d'accès direct à la mémoire distante (RDMAP, *Remote Direct Memory Access Protocol*) permet de supprimer les opérations de copie de données et permet la réduction des latences en faisant lire ou écrire les données par une application locale sur la mémoire d'un ordinateur distant avec une pression minimale sur la bande passante du bus de mémoire et les frais généraux de traitement de CPU, tout en préservant la sémantique de protection de la mémoire.

RDMAP est mis en couche par dessus le placement direct des données (DDP, *Direct Data Placement*) et utilise les deux modèles de mémoire tampon disponibles dans DDP. La terminologie relative à DDP est exposée au paragraphe 2.3. Comme RDMAP s'appuie sur DDP, le lecteur devrait se familiariser avec la [RFC5041].

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].

1.1. Buts architecturaux

RDMAP a été conçu avec les buts architecturaux de haut niveau suivants :

* Fournir une opération de transfert des données qui permette à un homologue local de transférer jusqu'à 2^32 - 1 octets directement dans une mémoire tampon annoncée précédemment (c'est-à-dire, une mémoire tampon étiquetée) située sur un homologue distant sans exiger une opération de copie. Ceci est appelé l'opération de transfert des données RDMA Write.

* Fournir une opération de transfert des données qui permette à un homologue local de restituer jusqu'à 2^32 - 1 octets directement d'une mémoire tampon annoncée précédemment (c'est-à-dire, une mémoire tampon étiquetée) située sur un homologue distant sans exiger d'opération de copie. Ceci est appelé l'opération de transfert des données RDMA Read.

* Fournir une opération de transfert des données qui permette à un homologue local d'envoyer jusqu'à 2^32 - 1 octets directement dans une mémoire tampon située chez un homologue distant qui n'a pas été explicitement annoncé. Ceci est appelé l'opération de transfert des données Send (Send avec Invalidate, Send avec événement sollicité, et Send avec événement sollicité et Invalidate).

* Permettre à l'ULP local d'utiliser le type d'opération Send (incluant Send, Send avec Invalidate, Send avec événement sollicité, et Send avec événement sollicité et Invalidate) pour signaler à l'ULP distant l'achèvement de tous les messages précédents initiés par l'ULP local.

* Faire que toutes les opérations sur un seul flux RDMAP soient transmises de façon fiable dans l'ordre de leur soumission.

* Fournir une capacité RDMAP indépendante pour chaque flux quand le LLP prend en charge plusieurs flux de données au sein d'une connexion LLP.

1.2 Vue d'ensemble du protocole

RDMAP fournit sept opérations de transfert des données. Sauf pour l'opération RDMA Read, chaque opération génère exactement un message RDMA. Voici un bref survol des opérations et messages RDMA :

1. Send : une opération Send utilise un message Send pour transférer des données de la source de données dans une mémoire tampon qui n'a pas été explicitement annoncée par le collecteur de données. Le message Send utilise le modèle de mémoire tampon DDP non étiquetée pour transférer le message d'ULP dans la mémoire tampon non étiquetée du collecteur de données.

2. Send avec Invalidate : une opération Send avec Invalidate utilise un message Send avec Invalidate pour transférer des données de la source de données dans une mémoire tampon qui n'a pas été explicitement annoncée par le collecteur de données. Le message Send avec Invalidate inclut toutes les fonctionnalités du message Send avec un ajout : un champ STag est inclus dans le message Send avec Invalidate. Après que le message a été placé et livré au collecteur de données, la mémoire tampon de l'homologue distant identifiée par la STag ne peut plus être accessible à distance jusqu'à ce que l'ULP de l'homologue distant rétablisse l'accès et annonce la mémoire tampon.

3. Send avec événement sollicité (Send avec SE) : une opération Send avec événement sollicité utilise un message Send avec événement sollicité pour transférer des données depuis la source de données dans une mémoire tampon non étiquetée au collecteur de données. Le message Send avec événement sollicité est similaire au message Send, avec un ajout : quand le message Send avec événement sollicité a été placé et livré, un événement peut être généré chez le receveur, si celui-ci est configuré à générer un tel événement.

4. Send avec événement sollicité et Invalidate (Send avec SE et Invalidate) - une opération Send avec événement sollicité et Invalidate utilise un message Send avec événement sollicité et Invalidate pour transférer des données de la source de données dans une mémoire tampon qui n'a pas été explicitement annoncée par le collecteur de données. Le message Send avec événement sollicité et Invalidate est similaire au message Send avec Invalidate, avec un ajout : quand le message Send avec événement sollicité et Invalidate a été placé et livré, un événement peut être généré chez le receveur, si celui-ci est configuré à générer un tel événement.

5. Accès direct en écriture à une mémoire distante (RDMA Write, *Remote Direct Memory Access Write*) : une opération RDMA Write utilise un message RDMA Write pour transférer des données de la source de données à une mémoire tampon annoncée précédemment au collecteur de données. L'ULP chez l'homologue distant, qui dans ce cas est le collecteur de données, permet l'accès à la mémoire tampon étiquetée du collecteur de données et annonce la taille de la mémoire tampon (longueur) sa situation (décalage étiqueté *(Tagged Offset)*) et l'étiquette de pilotage (STag, *Steering Tag*) à la source de données par un mécanisme spécifique de l'ULP. L'ULP chez l'homologue local, qui dans ce cas est la source de données, initie l'opération RDMA Write. Le message RDMA Write utilise le modèle de mémoire tampon étiquetée DDP pour transférer le message d'ULP dans la mémoire tampon étiquetée du collecteur de données. Noter que la STag associée à la mémoire tampon étiquetée reste valide jusqu'à ce que l'ULP chez l'homologue distant, ou l'ULP chez l'homologue local, l'invalide avec un Send avec Invalidate ou un Send avec événement sollicité et Invalidate.

6. Accès direct en lecture à une mémoire distante (RDMA Read, *Remote Direct Memory Access Read*) : l'opération RDMA Read transfère les données à une mémoire tampon étiquetée chez l'homologue local, qui dans ce cas est le collecteur de données, depuis une mémoire tampon étiquetée chez l'homologue distant, qui dans ce cas est la source de données. L'ULP à la source de données permet l'accès à la mémoire tampon étiquetée de la source de données et annonce la taille de la mémoire tampon (longueur) sa situation (décalage étiqueté *(Tagged Offset)*) et l'étiquette de pilotage (STag, *Steering Tag*) au collecteur de données par un mécanisme spécifique de l'ULP. L'ULP au collecteur de données permet l'accès à la mémoire tampon étiquetée du collecteur de données et initie l'opération RDMA Read. Celle-ci consiste en un seul message de demande RDMA Read et un seul message RDMA Read de réponse, et ce dernier peut être segmenté en plusieurs segments DDP. Le message de demande RDMA Read utilise le modèle DDP de mémoire tampon non étiquetée pour livrer la STag, en commençant par le décalage étiqueté *(Tagged Offset)* et la longueur pour les deux mémoires tampon étiquetées de la source de données et du collecteur de données à la file d'attente de demandes RDMA Read de l'homologue distant. Le message de réponse RDMA Read utilise le modèle DDP de mémoire tampon étiquetée pour livrer la mémoire tampon étiquetée de la source de données au collecteur de données, sans aucune implication de l'ULP de la source de données.
   Noter que la STag de source de données associée à la mémoire tampon étiquetée reste valide jusqu'à ce que l'ULP à la

source de données, ou l'ULP au collecteur de données, l'invalide avec un Send avec Invalidate ou un Send avec événement sollicité et Invalidate. La STag de collecteur de données associée à la mémoire tampon étiquetée reste valide jusqu'à ce que l'ULP au collecteur de données l'invalide.

7. Terminate : une opération Terminate utilise un message Terminé pour transférer à l'homologue distant les informations associées à une erreur qui s'est produite chez l'homologue local. Le message Terminé utilise le modèle DDP de mémoire tampon non étiquetée pour transférer le message dans la mémoire tampon non étiquetée du collecteur de données.

1.3 Mise en couches de RDMAP

RDMAP dépend de DDP, sous réserve des exigences définies au paragraphe 3.1, "Exigences et hypothèses de transport". La Figure 1, "Mise en couche de RDMAP", décrit les relations entre les protocoles de couche supérieure (les ULP), RDMAP, le protocole DDP, la couche de tramage, et le transport. Pour les définitions de protocole LLP de chaque LLP, voir les [RFC0793] [RFC4960] et [RFC5044].

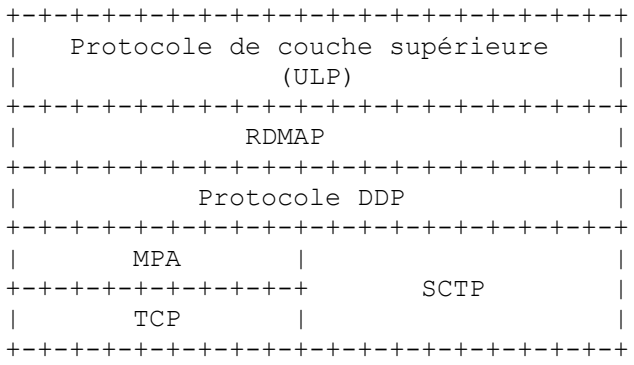

Figure 1 : Mise en couches de RDMAP

Si RDMAP est mis en couche sur DDP/MPA/TCP, les en-têtes respectifs et la charge utile d'ULP sont arrangés comme suit (noter pour être clair que les champs d'en-tête MPA et de CRC sont inclus mais les marqueurs MPA ne sont pas montrés).

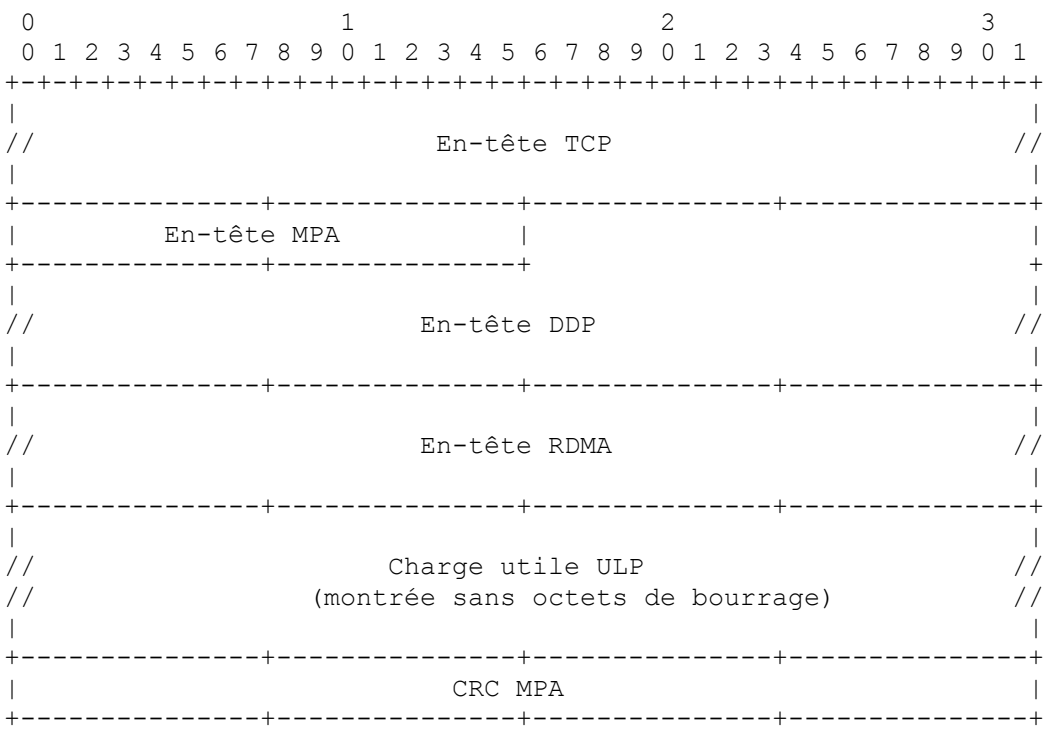

Figure 2 : Exemple d'alignement d'en-têtes MPA, DDP, et RDMAP sur TCP

2. Glossaire

2.1 Général

Annoncer (annoncé, annonce, annonces) : acte d'informer un homologue distant qu'une mémoire tampon locale RDMA lui est disponible. Un nœud rend disponible une mémoire tampon RDMA pour l'accès entrant RDMA Read ou RDMA Write en informant son homologue RDMA/DDP des identifiants de la mémoire tampon étiquetée (STag, adresse de base, et longueur de mémoire tampon). Ces informations d'annonce de mémoire tampon étiquetée ne sont pas définies par RDMA/DDP et sont laissées à l'ULP. Une méthode normale serait que l'homologue local incorpore l'étiquette de pilotage, l'adresse de base et la longueur de mémoire tampon étiquetée dans un message Send destiné à l'homologue distant.

Achèvement (achevé, achever, achève) : voir "Achèvement RDMA" au paragraphe 2.4.

Collecteur de données : l'homologue qui reçoit une charge utile de données. Noter que le collecteur de données peut être obligé d'envoyer et de recevoir des messages RDMA/DDP pour transférer une charge utile de données.

Source de données - l'homologue qui envoie une charge utile de données. Noter que la source de données peut être obligée d'envoyer et recevoir des messages RDMA/DDP pour transférer une charge utile de données.

Livraison des données (livrer, livrées) : livraison est défini comme le processus d'informer l'ULP ou le consommateur qu'un message particulier est disponible à l'utilisation. Ceci est spécifiquement différent de "placement", qui peut généralement se produire dans n'importe quel ordre, tandis que l'ordre de "livraison" est strictement défini. Voir "placement de données" au paragraphe 2.3.

Livraison (livré, livrer) : voir livraison des données ci-dessus.

Tissu : collection de liaisons, commutateurs, et routeurs qui connectent un ensemble de nœuds avec les mises en œuvre de protocole RDMA/DDP.

Barrer (barré, barrières) : bloquer l'exécution de l'opération RDMA en cours jusqu'à ce que les opérations RDMA précédentes soient achevées.

iWARP : suite de protocoles filaires composée de RDMAP, DDP, et MPA. La suite de protocoles iWARP peut être mise en couches par dessus TCP, SCTP, ou d'autres protocoles de transport.

Homologue local : mise en œuvre du protocole RDMA/DDP sur l'extrémité locale de la connexion. Utilisé pour se référer à l'entité locale quand on décrit un échange de protocole ou une autre interaction entre deux nœuds.

Nœud : appareil informatique rattaché à une ou plusieurs liaisons d'un tissu (réseau). Un nœud dans ce contexte ne se réfère pas à une application ou instanciation de protocole spécifique fonctionnant sur l'ordinateur. Un nœud peut consister en un ou plusieurs RNIC installés dans un ordinateur hôte.

Placement (placé, place) : voir "placement de données" au paragraphe 2.3.

Homologue distant - mise en œuvre du protocole RDMA/DDP sur l'extrémité opposée de la connexion. Utilisé pour se référer à l'entité distante quand on décrit des échanges de protocole ou d'autres interactions entre deux nœuds.

RNIC (*RDMA Network Interface Controller*) contrôleur d'interface de réseau RDMA : dans ce contexte, ce serait un adaptateur d'entrée/sortie de réseau ou un contrôleur incorporé avec la fonction iWARP et verbes.

Interface RNIC (RI, *RNIC Interface*) : présentation du RNIC au consommateur de verbes comme mis en œuvre par la combinaison du RNIC et du pilote de RNIC.

Terminaison (Terminé, Terminer, Termine) : voir "Terminaison RDMAP interruptive" au paragraphe 2.4.

Protocole de couche supérieure (ULP, *Upper Layer Protocol*) : couche de protocole au-dessus de celle actuellement référencée. L'ULP pour RDMA/DDP est supposée être un OS, une application, une couche d'adaptation, ou un appareil propriétaire. Les documents RDMA/DDP ne spécifient pas d'ULP -- ils fournissent un ensemble sémantique qui permet à un ULP d'être conçu pour utiliser RDMA/DDP.

Charge utile d'ULP : données d'ULP qui sont contenues dans un seul segment ou paquet de protocole (par exemple, un segment DDP).

Verbes : description abstraite de la fonction d'une interface de RNIC. L'OS peut exposer certaines de ces fonctions ou toutes via une ou plusieurs API aux applications. L'OS va aussi utiliser certaines des fonctions pour gérer l'interface de RNIC.

2.2 LLP

LLP *(Lower Layer Protocol)* : protocole de couche inférieure. Couche de protocole en dessous de la couche de protocole actuellement référencée. Par exemple, pour DDP, le LLP est SCTP, MPA, ou d'autres protocoles de transport. Pour RDMA, le LLP est DDP.

Connexion de LLP : correspond à une connexion de niveau transport de LLP entre les couches LLP de l'homologue sur deux nœuds.

Flux de LLP : correspond à un seul flux de niveau transport de LLP entre les couches LLP de l'homologue sur deux nœuds. Un ou plusieurs flux de LLP peuvent se transposer en une seule connexion de niveau transport de LLP. Pour les protocoles de transport qui peuvent prendre en charge plusieurs flux par connexion (par exemple, SCTP) un flux de LLP correspond à un flux de niveau transport.

MULPDU (Maximum ULPDU) : taille maximum courante de l'enregistrement qui est acceptable pour que DDP le passe au LLP pour transmission.

ULPDU *(Upper Layer Protocol Data Unit)* unité de données de protocole de couche supérieure : enregistrement de données défini par la couche au-dessus de MPA.

2.3. Placement direct des données (DDP, *Direct Data Placement*)

Placement de données (placement, placé, placer) : pour DDP, ce terme est spécifiquement utilisé pour indiquer le processus d'écriture dans une mémoire tampon de données par une mise en œuvre de DDP. Les segments DDP portent des informations de placement, qui peuvent être utilisées par la mise en œuvre DDP receveuse pour effectuer le placement de données de la charge utile du segment DDP ULP. Voir "livraison des données".

Suppression de DDP interruptive : action de clôture d'un flux DDP sans tenter d'achever les messages DDP en cours et en instance.

Suppression de DDP en douceur : acte de clôture d'un flux DDP tel que tous les messages DDP en cours et en instance puissent s'achever avec succès.

Champ de contrôle de DDP : champ fixe de 16 bits dans l'en-tête DDP. Le champ de contrôle de DDP contient un champ de 8 bits dont le contenu est réservé à l'usage de l'ULP.

En-tête DDP : en-tête présent dans tous les segments DDP. L'en-tête DDP contient les champs de contrôle et de placement qui sont utilisés pour définir la localisation du placement final de la charge utile d'ULP portée dans un segment DDP.

Message DDP : unité d'échange de données définie par l'ULP, qui est subdivisée en un ou plusieurs segments DDP. Cette segmentation peut survenir pour diverses raisons, incluant la segmentation pour respecter la taille maximum de segment du protocole de transport sous-jacent.

Segment DDP : plus petite unité de transfert de données pour le protocole DDP. Il inclut un en-tête DDP et une charge utile d'ULP (si elle est présente). Un segment DDP devrait être dimensionné pour tenir dans la MULPDU du protocole de transport sous-jacent.

Flux DDP : séquence de messages DDP dont l'ordre est défini par le LLP. Pour SCTP, un flux DDP se transpose directement en un flux SCTP. Pour MPA, un flux DDP se transpose directement en une connexion TCP, et un seul flux DDP est pris en charge. Noter que DDP ne donne pas de garantie d'ordre entre les flux DDP.

Placement direct de données : mécanisme par lequel les données d'ULP contenues dans les segments DDP peuvent être

placées directement dans leur destination finale en mémoire sans traitement de la part de l'ULP. Cela peut se produire même quand les segments DDP arrivent dans le désordre. La prise en charge du placement dans le désordre peut exiger que le collecteur de données mette en œuvre LLP et DDP comme un bloc fonctionnel.

Protocole de placement direct de données (DDP) : c'est aussi un protocole réseau qui prend en charge le placement direct de données en associant des informations de placement explicite en mémoire tampon aux unités de charge utile de LLP.

Décalage de mémoire (MO, *Memory Offset*) : pour le modèle de mémoire tampon non étiquetée de DDP, spécifie le décalage, en octets, depuis le début d'un message DDP.

Numéro de séquence de message (MSN, *Message Sequence Number*) : pour le modèle de mémoire tampon non étiquetée de DDP, spécifie un numéro de séquence qui augmente avec chaque message DDP.

Numéro de file d'attente (QN, *Queue Number*) : pour le modèle de mémoire tampon non étiquetée de DDP, identifie une file d'attente de collecteur de données de destination pour un segment DDP.

Étiquette de pilotage (STag, *Steering Tag*) : identifiant d'une mémoire tampon étiquetée sur un nœud, dont la validité est définie dans une spécification de protocole.

Mémoire tampon étiquetée : mémoire tampon qui est explicitement annoncée à l'homologue distant par un échange d'une STag, d'un décalage étiqueté, et d'une longueur.

Modèle de mémoire tampon étiquetée : modèle de transfert de données DDP utilisé pour transférer des mémoires tampon étiquetées de l'homologue local à l'homologue distant.

Message DDP étiqueté : message DDP qui cible une mémoire tampon étiquetée.

Décalage étiqueté (TO, *Tagged Offset*) - décalage au sein d'une mémoire tampon étiquetée sur un nœud.

Mémoire tampon non étiquetée - mémoire tampon qui n'est pas explicitement annoncée à l'homologue distant. Les mémoires tampon non étiquetées prennent en charge un des deux mécanismes de transfert de données disponibles appelé le modèle de mémoire tampon non étiquetée. Une mémoire tampon non étiquetée est utilisée pour envoyer des messages de contrôle asynchrones à l'homologue distant pour les demandes RDMA Read, Send, et Terminate. Les mémoires tampon non étiquetées traitent les messages DDP non étiquetés.

Modèle de mémoire tampon non étiquetée : modèle de transfert de données DDP utilisé pour transférer des mémoires tampon non étiquetées de l'homologue local à l'homologue distant.

Message DDP non étiqueté : message DDP qui cible une mémoire tampon non étiquetée.

2.4 Accès direct à la mémoire distante (RDMA, *Remote Direct Memory Access*)

File d'attente d'achèvement (CQ, *Completion Queue*) : composant logique de l'interface de RNIC qui représente conceptuellement comment un RNIC notifie à l'ULP l'achèvement de la transmission des données, ou l'achèvement de la réception des données ; voir la [RFC5042].

Événement : indication fournie par la couche RDMAP à l'ULP pour indiquer un achèvement ou autre condition qui exige une attention immédiate.

STag invalidante : mécanisme utilisé pour empêcher l'homologue distant de réutiliser une STAG annoncée explicitement précédente, jusqu'à ce que l'homologue local la rende disponible par une annonce explicite suivante. La STag ne peut pas être accédée à distance jusqu'à ce qu'elle soit explicitement annoncée à nouveau.

Achèvement RDMA (achevé, achève, achever) : pour RDMA, l'achèvement est défini comme le processus d'information de l'ULP qu'une opération RDMA particulière a effectué toutes les fonctions spécifiées pour les opérations RDMA, incluant le placement et la livraison. La sémantique d'achèvement de chaque opération RDMA est définie de façon distincte.

Message RDMA : mécanisme de transfert de données utilisé pour accomplir une opération RDMA.

Opération RDMA : séquence de messages RDMA, incluant des messages de contrôle, pour transférer des données d'une source de données à un collecteur de données. Les opérations RDMA suivantes sont définies : RDMA Write, RDMA Read, Send, Send avec Invalidate, Send avec événement sollicité, Send avec événement sollicité et Invalidate, et Terminate.

Protocole RDMA (RDMAP) : protocole du réseau qui prend en charge les opérations RDMA pour transférer les données d'ULP entre un homologue local et l'homologue distant.

Terminaison RDMAP interruptive (Terminaison, Terminé, Terminer, Termine) : acte de clôture d'un flux RDMAP sans tenter d'achever les opérations RDMA en cours et en instance.

Terminaison RDMAP en douceur : acte de clôture d'un flux RDMAP de façon à ce que toutes les opérations RDMA en cours et en instance puissent s'achever avec succès.

RDMA Read : opération RDMA utilisée par le collecteur de données pour transférer le contenu d'une mémoire tampon d'une source RDMA de l'homologue distant à l'homologue local. Une opération RDMA Read consiste en un seul message de demande RDMA Read et un seul messsage de réponse RDMA Read.

Demande RDMA Read : message RDMA utilisé par le collecteur de données pour demander à la source de données de transférer le contenu d'une mémoire tampon RDMA. Le message de demande RDMA Read décrit les deux mémoires tampon RDMA de la source de données et du collecteur de données.

File d'attente de demandes RDMA Read : file d'attente utilisée pour traiter les demandes RDMA Read. La file d'attente de demandes RDMA Read a un numéro de file d'attente DDP de 1.

Réponse RDMA Read : message RDMA utilisé par la source de données pour transférer le contenu d'une mémoire tampon RDMA au collecteur de données, en réponse à une demande RDMA Read. Le message de réponse RDMA Read décrit seulement la mémoire tampon RDMA du collecteur de données.

Flux RDMAP : association entre une paire de mises en œuvre de RDMAP, éventuellement sur des nœuds différents, qui transfère les données d'ULP en utilisant les opérations RDMA. Il peut y avoir plusieurs flux RDMAP sur un seul nœud. Un flux RDMAP se transpose directement en un seul flux DDP.

RDMA Write : opération RDMA qui transfère le contenu d'une mémoire tampon RDMA de source de l'homologue local à une mémoire tampon RDMA de destination chez l'homologue distant en utilisant RDMA. Le message RDMA Write décrit seulement la mémoire tampon RDMA du collecteur de données.

Accès direct à la mémoire distante (RDMA, *Remote Direct Memory Access*) : méthode d'accès à la mémoire sur un système distant dans lequel le système local spécifie la localisation distante des données à transférer. Employer un RNIC dans le système distant permet que l'accès ait lieu sans interrompre le traitement de la ou des CPU sur le système.

Send :- opération RDMA qui transfère le contenu d'une mémoire tampon d'ULP de l'homologue local à une mémoire tampon non étiquetée chez l'homologue distant.

Type de message Send : message Send, message Send avec Invalidate, message Send avec événement sollicité, ou message Send avec événement sollicité et Invalidate.

Type d'opération Send - opération Send, Send avec Invalidate, Send avec événement sollicité, ou Send avec événement sollicité et Invalidate.

Événement sollicité (SE, *Solicited Event*) : facilité par laquelle l'envoyeur d'une opération RDMA peut causer la génération d'un événement chez le receveur, si le receveur est configuré à générer un tel événement, quand un message Send avec événement sollicité ou Send avec événement sollicité et Invalidate est reçu. Noter que l'ULP de l'homologue local peut utiliser le mécanisme d'événement sollicité pour s'assurer que les messages désignés comme importants pour l'ULP soient traités d'une manière expéditive par l'ULP de l'homologue distant. L'ULP chez l'homologue local peut indiquer qu'un certain type de message Send est important en utilisant le message Send avec événement sollicité ou le message Send avec événement sollicité et Invalidate. L'ULP à l'homologue distant peut choisir d'être seulement notifié quand des messages valides Send avec événement sollicité et/ou des messages Send avec événement sollicité et Invalidate arrivent et traiter les autres messages Send ou Send avec Invalidate entrants valides comme bon lui semble.

Terminé : message RDMA utilisé par un nœud pour passer une indication d'erreur au nœud homologue sur un flux

RDMAP. Cette opération est pour le seul usage de RDMAP.

Mémoire tampon d'ULP : mémoire tampon appartenant à une couche au dessus de la couche RDMAP et annoncée à la couche RDMAP soit comme une mémoire tampon étiquetée, soit comme une mémoire tampon d'ULP non étiquetée.

Message d'ULP : données d'ULP qui sont passées à une couche de protocole spécifique pour transmission. Les limites de données sont préservées lorsque elles sont transmises par iWARP.

3. Attributs de couche supérieure et de transport

3.1 Exigences et hypothèses de transport

RDMAP DOIT être mis en couche par dessus le protocole de placement direct de données [RFC5041].

RDMAP exige la prise en charge DDP suivante :
* RDMAP utilise trois files d'attente pour les mémoires tampon non étiquetées :
  - numéro de file d'attente 0 (utilisé par RDMAP pour les opérations Send, Send avec Invalidate, Send avec événement sollicité, et Send avec événement sollicité et Invalidate).
  - numéro de file d'attente 1 (utilisé par RDMAP pour les opérations RDMA Read).
  - numéro de file d'attente 2 (utilisé par RDMAP pour les opérations Terminate).
* DDP transpose un seul message RDMA en un seul message DDP.
* DDP utilise la STag et le décalage étiqueté fournis par le RDMAP pour les messages de mémoire tampon étiquetée (c'est-à-dire, les réponses à RDMA Write et RDMA Read).
* Quand la couche DDP livre un message DDP non étiqueté à la couche RDMAP, DDP fournit la longueur du message DDP. Cela assure que RDMAP n'a pas à porter un champ de longueur dans son en-tête.
* Quand la couche RDMAP fournit un message RDMA à la couche DDP, DDP doit insérer la valeur du champ RsvdULP fournie par la couche RDMAP dans le message DDP associé.
* Quand la couche DDP livre un message DDP à la couche RDMAP, DDP fournit le champ RsvdULP.
* Le champ RsvdULP doit être de 1 octet pour les messages DDP étiquetés et de 5 octets pour les messages DDP non étiquetés.
* DDP propage à RDMAP toutes les erreurs d'opération ou de protection (utilisées par RDMAP Terminate) et, quand c'est approprié, les champs d'en-tête DDP du segment DDP qui a rencontré l'erreur.
* Si une opération RDMA est interrompue par DDP ou une couche inférieure, le contenu des mémoires tampon du collecteur de données associé à l'opération sont considérées comme étant indéterminé.
* DDP, en conjonction avec les couches inférieures, fournit une livraison fiable, en ordre.

3.2 Interactions de RDMAP avec la couche supérieure

RDMAP fournit à l' ULP l'accès aux opérations RDMA suivantes comme défini dans la présente spécification :
* Send
* Send avec événement sollicité
* Send avec Invalidate
* Send avec événement sollicité et Invalidate
* RDMA Write
* RDMA Read

Pour les types d'opération Send, les interactions suivantes ont lieu entre la couche RDMAP et l'ULP :
* À la source de données :
  * L'ULP passe ce qui suit à la couche RDMAP :
    * longueur du message d'ULP
    * message d'ULP
    * une indication du type d'opération Send, où les types valides sont : Send, Send avec événement sollicité, Send avec Invalidate, ou Send avec événement sollicité et Invalidate.
    * une STag invalidante, si le type d'opération Send était Send avec Invalidate ou Send avec événement sollicité et Invalidate.
  * Quand le type d'opération Send s'achève, il en résulte une indication de l'achèvement.
* Au collecteur de données :
  * Si le type d'opération Send s'achève avex succès, la couche RDMAP passe les informations suivantes à la couche ULP :

    * longueur du message d'ULP
    * message d'ULP
    * un événement, si le collecteur de données est configuré à générer un événement.
    * Une STag invalidée, si le type d'opération Send était Send avec Invalidate ou Send avec événement sollicité et Invalidate.
  * Si le type d'opération Send s'achève avec une erreur, la couche RDMAP du collecteur de données va passer les informations d'erreur correspondantes à l'ULP du collecteur de données et envoyer un message Terminé à la couche RDMAP de la source de données. La couche RDMAP de la source de données va alors passer le message Terminé à l'ULP.

Pour les opérations RDMA Write, les interactions suivantes ont lieu entre la couche RDMAP et l'ULP :
* À la source de données :
  * L'ULP passe ce qui suit à la couche RDMAP :
    * longueur du message ULP
    * message ULP
    * STag de collecteur de données
    * décalage étiqueté de collecteur de données
  * Quand l'opération RDMA Write s'achève, il en résulte une indication d'achèvement.
* Au collecteur de données :
  * Si l'opération RDMA Write s'achève avec succès, la couche RDMAP ne livre pas le RDMA Write à l'ULP. Elle place le message d'ULP transféré par le message RDMA Write dans la mémoire tampon d'ULP.
  * Si le RDMA Write s'achève par une erreur, la couche RDMAP du collecteur de données va passer les informations d'erreur correspondantes à l'ULP du collecteur de données et envoyer un message Terminé à la couche RDMAP de la source de données. La couche RDMAP de la source de données va alors passer le message Terminé à l'ULP.

Pour les opérations RDMA Read, les interactions suivantes ont lieu entre la couche RDMAP et l'ULP :
* Au collecteur de données :
  * L'ULP passe ce qui suit à la couche RDMAP :
    * longueur du message d'ULP
    * STag de la source de données
    * STag du collecteur de données
    * décalage étiqueté de la source de données
    * décalage étiqueté du collecteur de données
  * Quand l'opération RDMA Read s'achève, il en résulte une indication de l'achèvement.
* À la source de données :
  * Si aucune erreur ne s'est produite lors du traitement de la demande RDMA Read, la source de données ne va passer aucune information à l'ULP.
  * Si une erreur s'est produite lors du traitement de la demande RDMA Read, la couche RDMAP de la source de données va passer les informations d'erreur correspondantes à l'ULP de la source de données et envoyer un message Terminé à la couche RDMAP du collecteur de données. La couche RDMAP du collecteur de données va alors passer le message Terminé à l'ULP.

Pour les STag rendues disponibles à la couche RDMAP, les interactions suivantes ont lieu entre la couche RDMAP et l'ULP :
* Si l'ULP active une STag, l'ULP passe les éléments suivants à la couche RDMAP :
  * STag ;
  * gamme des décalages étiquetés qui sont associés à une STag donnée ;
  * droits d'accès distants (lecture, écriture, ou lecture et écriture) associés à la STag valide ; et
  * association entre une certaine STag et un flux RDMAP donné.
* Si l'ULP désactive une STag, l'ULP passe la STag à la couche RDMAP.

Si une erreur survient à la couche RDMAP, la couche RDMAP peut repasser les informations d'erreur (par exemple, le contenu d'un message Terminé) à l'ULP.

4. Format d'en-tête

Les informations de contrôle des messages RDMA sont incluses dans les champs d'en-tête définis par le protocole DDP, avec les exceptions suivantes :

* Le premier octet réservé pour l'usage de l'ULP sur tous les messages DDP dans le protocole DDP (c'est-à-dire, le champ RsvdULP) est utilisé par RDMAP pour porter le Opcode *(code de fonctionnement)* du message RDMA et la version RDMAP. Cet octet est appelé le champ de contrôle RDMAP dans la présente spécification. Pour Send avec Invalidate et Send avec événement sollicité et Invalidate, RDMAP utilise les octets deux à cinq, fournis par DDP sur les messages DDP non étiquetés, pour porter la STag qui va être invalidée.

* La longueur du message RDMA est passée par la couche RDMAP à la couche DDP sur tous les transferts sortants.

* Pour les messages de demande RDMA Read, la taille du message RDMA Read est incluse dans l'en-tête de la demande RDMA Read.

* La longueur du message RDMA est passée à la couche RDMAP par la couche DDP sur les transferts entrants de mémoire tampon non étiquetée.

* Deux messages RDMA portent des en-têtes RDMAP supplémentaires. La demande RDMA Read porte les descriptions de mémoire tampon de collecteur de données et de source de données, incluant la longueur de la mémoire tampon. Le message Terminé porte des informations supplémentaires associées à l'erreur qui a causé le Terminé.

4.1 Champ contrôle RDMAP et STag Invalidate

La version de RDMAP définie dans la présente spécification utilise tous les 8 bits du champ de contrôle RDMAP. Le premier octet réservé pour l'usage de l'ULP dans le protocole DDP DOIT être utilisé par RDMAP pour porter le champ Contrôle RDMAP. L'ordre des bits dans le premier octet DOIT être comme défini dans la Figure 3, "Champs Contrôle DDP, Contrôle RDMAP, et STag invalidante". Pour Send avec Invalidate et Send avec événement sollicité et Invalidate, les octets du second au cinquième du champ DDP RsvdULP DOIVENT être utilisés par RDMAP pour porter la STag invalidante. La Figure 3 décrit le format des champs Contrôle DDP et Contrôle RDMAP. (Noter que dans la Figure 3, l'en-tête DDP est décalé de 16 bits pour s'accommoder de l'en-tête MPA défini dans la [RFC5044]. L'en-tête MPA est seulement présent si DDP est mis en couche par dessus MPA.)

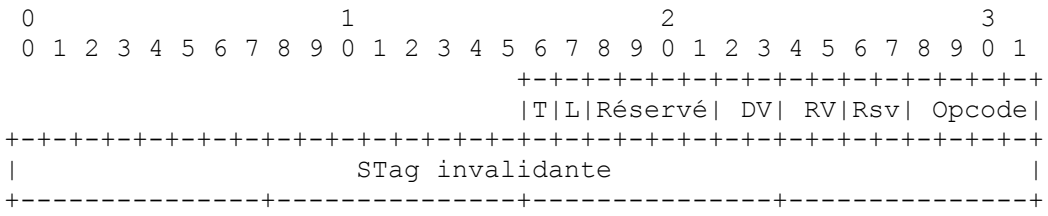

Figure 3 : Champs Contrôle DDP, Contrôle RDMAP, et STag invalidante

Tous les messages RDMA passés de la couche RDMAP à la couche DDP DOIVENT définir la valeur du fanion Étiqueté dans l'en-tête DDP. La Figure 4, "Usage par RDMA des champs DDP", DOIT être utilisée pour définir la valeur du fanion Étiqueté qui est passée à la couche DDP pour chaque message RDMA.

La Figure 4 définit la valeur du champ RDMA Opcode qui DOIT être utilisé pour chaque message RDMA.

La Figure 4 définit quand les champs STag, Numéro de file d'attente, et Décalage étiqueté DOIVENT être fournis pour chaque message RDMA.

Pour la présente version de RDMAP, tous les messages RDMA DOIVENT avoir :

* aux bits 24-25, champ Version RDMA : 01b pour un RNIC qui se conforme à la présente spécification de protocole RDMA ; 00b pour un RNIC qui se conforme à la spécification de protocole RDMA du Consortium RDMA. Les deux numéros de version sont valides. L'interopérabilité dépend de la négociation de la version de protocole MPA (par exemple, marqueur MPA et CRC MPA).

* aux bits 26-27, réservé : DOIVENT être réglés à zéro par l'envoyeur, et ignorés par le receveur.

* aux bits 28-31, champ OpCode : voir la Figure 4.

* aux bits 32-63, STag invalidante. Cependant, ce champ est seulement valide pour les messages Send avec Invalidate et

Send avec événement sollicité et Invalidate (voir la Figure 4).

Pour Send, Send avec événement sollicité, demande RDMA Read, et Terminate, le champ STag Invalidante DOIT être réglé à zéro à l'émission et ignoré à réception.

OpCode du message RDMA	Type de message	Fanion Étiqueté	STag et TO	Numéro de file d'attente	STag invalidante	Longueur de message communiquée entre DDP et RDMAP
0000b	RDMA Write	1	Valide	N/A	N/A	oui
0001b	Demande RDMA Read	0	N/A	1	N/A	oui
0010b	Réponse RDMA Read	1	Valide	N/A	N/A	oui
0011b	Send	0	N/A	0	N/A	oui
0100b	Send avec Invalidate	0	N/A	0	Valide	oui
0101b	Send avec SE	0	N/A	0	N/A	oui
0110b	Send avec SE et Invalidate	0	N/A	0	Valide	oui
0111b	Terminate	0	N/A	2	N/A	oui
1000b à 1111b	Réservé	Non spécifié				

Figure 4 : Usage par RDMA des champs DDP

Note : N/A signifie "Non applicable".

4.2 Définition des messages RDMA

La figure suivante définit quels en-têtes RDMA DOIVENT être utilisés sur chaque message RDMA et quels messages RDMA sont autorisés à porter une charge utile d'ULP :

OpCode de message RDMA	Type de message	En-tête RDMA utilisé	Message d'ULP permis dans le message RDMA
0000b	RDMA Write	aucun	oui
0001b	Demande RDMA Read	En-tête de demande RDMA Read	non
0010b	Réponse RDMA Read	aucun	oui
0011b	Send	aucun	oui
0100b	Send avec Invalidate	aucun	oui
0101b	Send avec SE	aucun	oui
0110b	Send avec SE et Invalidate	aucun	oui
0111b	Terminé	En-tête Terminate	non
1000b à 1111b	Réservé	Non spécifié	

Figure 5 : Définitions de message RDMA

4.3 En-tête RDMA Write

Le message RDMA Write ne comporte pas d'en-tête RDMAP. La couche RDMAP passe à la couche DDP un champ de contrôle RDMAP. Le message RDMA Write est pleinement décrit par les en-têtes DDP des segments DDP associés au message.

Voir à l'Appendice A une description du format de segment DDP associé au message RDMA Write.

4.4 En-tête Demande RDMA Read

Le message de demande RDMA Read porte un en-tête Demande RDMA Read qui décrit les mémoires tampon de collecteur de données et de source de données utilisées par l'opération RDMA Read. L'en-tête Demande RDMA Read suit immédiatement l'en-tête DDP. La couche RDMAP passe à la couche DDP un champ Contrôle RDMAP. La Figure qui suit décrit l'en-tête Demande RDMA Read qui DOIT être utilisé pour tous les messages de demande RDMA Read :

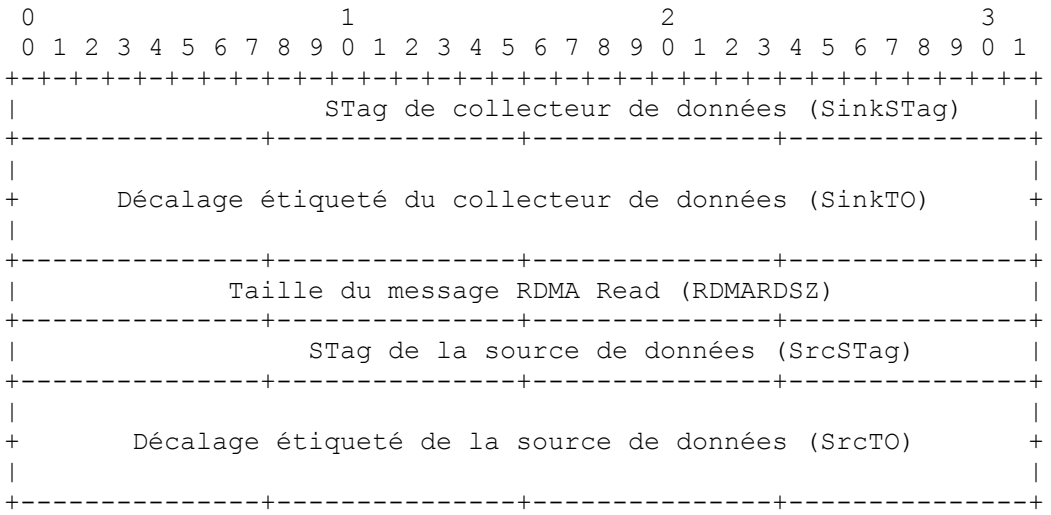

Figure 6 : Format d'en-tête Demande RDMA Read

STag de collecteur de données : 32 bits. L'étiquette de pilotage du collecteur de données identifie la mémoire tampon étiquetée du collecteur de données. Ce champ DOIT être copié, sans interprétation, de la demande RDMA Read dans la réponse RDMA Read correspondante ; ce champ permet au collecteur de données de placer les données retournées. La STag est associée au flux RDMAP par un mécanisme qui sort du domaine d'application de la spécification de RDMAP.

Décalage étiqueté de collecteur de données : 64 bits. Le décalage étiqueté de collecteur de données spécifie le décalage de départ, en octets, par rapport à la base de la mémoire tampon étiquetée du collecteur de données, où les données sont à écrire par la source de données. Ce champ est copié de la demande RDMA Read dans la réponse RDMA Read correspondante et permet au collecteur de données de placer les données retournées. Le décalage étiqueté de collecteur de données PEUT commencer à un décalage arbitraire. Les champs STag de collecteur de données et Décalage étiqueté de collecteur de données décrivent la mémoire tampon dans laquelle les données de RDMA Read sont écrites.
Note : la couche DDP protège contre un retour à zéro du décalage étiqueté du collecteur de données.

Taille du message RDMA Read : 32 bits. C'est la quantité de données, en octets, lues de la source de données. Un seul message de demande RDMA Read peut restituer de 0 à 2^32-1 octets de données de la source de données.

STag de la source de données : 32 bits. L'étiquette de pilotage de la source de données identifie la mémoire tampon étiquetée de la source de données. La STag est associée au flux RDMAP par un mécanisme qui sort du domaine d'application de la spécification RDMAP.

Décalage étiqueté de la source de données : 64 bits. Le décalage étiqueté spécifie le décalage de départ, en octets, qui est à lire de la mémoire tampon étiquetée de la source de données. Le décalage étiqueté de la source de données PEUT commencer à un décalage arbitraire. Les champs STag de source de données et Décalage étiqueté de source de données décrivent la mémoire tampon d'où les données de RDMA Read sont lues.

Voir au paragraphe 7.2, "Erreurs détectées à l'homologue distant sur les messages RDMA entrants" une description de la vérification d'erreur requise lors du traitement d'une demande RDMA Read à la source de données.

4.5 En-tête Réponse RDMA Read

Le message de réponse RDMA Read n'inclut pas d'en-tête RDMAP. La couche RDMAP passe à la couche DDP un champ Contrôle RDMAP. Le message de réponse RDMA Read est pleinement décrit par les en-têtes DDP des segments DDP associés au message.

Voir à l'Appendice A la description du format de segment DDP associé aux messages de réponse RDMA Read.

4.6 En-tête Send et Send avec événement sollicité

Les messages Send et Send avec événement sollicité n'incluent pas d'en-tête RDMAP. La couche RDMAP passe à la couche DDP un champ Contrôle RDMAP. Les messages Send et Send avec événement sollicité sont pleinement décrits par

les en-têtes DDP des segments DDP associés aux messages.

Voir à l'Appendice A la description du format de segment DDP associé aux messages Send et Send avec événement sollicité.

4.7 En-tête Send avec Invalidate et Send avec SE et Invalidate

Les messages Send avec Invalidate et Send avec événement sollicité et Invalidate n'incluent pas d'en-tête RDMAP. La couche RDMAP passe à la couche DDP un champ Contrôle RDMAP et le champ STag Invalidante (voir au paragraphe 4.1 le champ Contrôle RDMAP et STag Invalidante). Les messages Send avec Invalidate et Send avec événement sollicité et Invalidate sont pleinement décrits par les en-têtes DDP des segments DDP associés aux messages.

Voir à l'Appendice A la description du format de segment DDP associé aux messages Send et Send avec événement sollicité et Invalidate.

4.8 En-tête Terminate

Le message Terminé porte un en-tête Terminate qui contient des informations supplémentaires associées à la cause du Terminé. L'en-tête Terminate suit immédiatement l'en-tête DDP. La couche RDMAP passe à la couche DDP un champ Contrôle RDMAP. La figure suivante décrit un en-tête Terminate qui DOIT être utilisé pour le message Terminé :

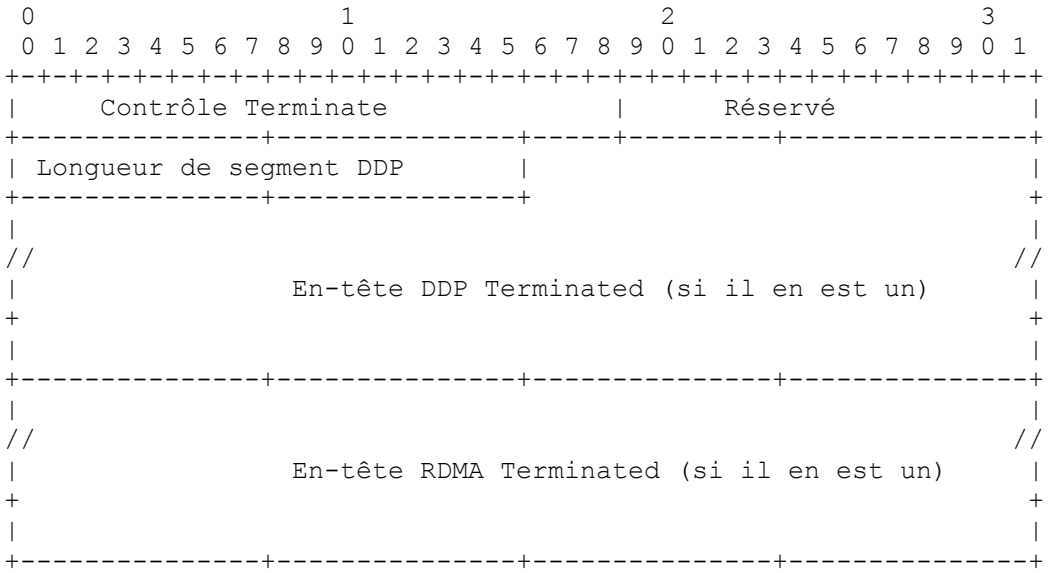

Figure 7 : Format d'en-tête Terminate

Contrôle Terminate : 19 bits. Le champ Contrôle Terminate DOIT avoir le format défini à la Figure 8.

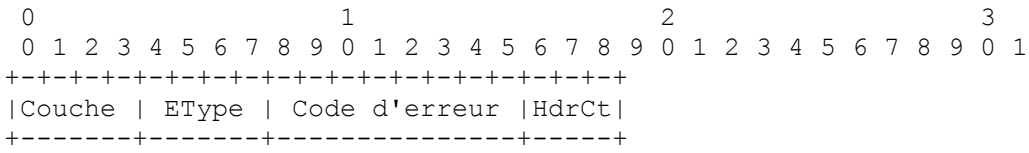

Figure 8 : Champ Contrôle Terminate

* La Figure 9, "Valeurs du champ Contrôle Terminate", définit les valeurs valides qui DOIVENT être utilisées pour ce champ.

* Couche : 4 bits. Identifie la couche qui a rencontré l'erreur.

* EType (Type d'erreur RDMA) : 4 bits. Identifie le type d'erreur qui a causé le Terminé. Quand l'erreur est détectée à la couche RDMAP, la couche RDMAP insère le type d'erreur dans ce champ. Quand l'erreur est détectée à une couche de

LLP, la couche LLP crée le type d'erreur et la couche DDP le passe à la couche RDMAP, et la couche RDMAP l'insère dans ce champ.

* Code d'erreur : 8 bits. Ce champ identifie l'erreur spécifique qui a causé le Terminé. Quand l'erreur est détectée à la couche RDMAP, la couche RDMAP crée le code d'erreur. Quand l'erreur est détectée à une couche de LLP, la couche LLP crée le code d'erreur, la couche DDP le passe à la couche RDMAP, et la couche RDMAP l'insère dans ce champ.

* HdrCt : 3 bits. Bits de contrôle d'en-tête :
  M : bit 16. Longueur de segment DDP valide. Voir à la Figure 10 quand ce bit DEVRAIT être établi.
  D : bit 17. En-tête DDP inclus. Voir à la Figure 10 quand ce bit DEVRAIT être établi.
  R : bit 18. En-tête RDMAP inclus. Voir à la Figure 10 quand ce bit DEVRAIT être établi.

Couche	Nom de couche	Type d'erreur	Nom de type d'erreur	Code d'erreur	Nom de code d'erreur
0000b	RDMA	0000b	Erreur locale catastrophique	aucun	aucun - ce type d'erreur n'a pas de code d'erreur. Toute valeur est acceptable dans ce champ.
		0001b	Erreur protection distante	00X	STag invalide
				01X	Violation de base ou limites
				02X	Violation des droits d'accès
				03X	STag non associée avec flux RDMAP
				04X	Retour à zéro de TO
				09X	STag ne peut pas être invalidée
				FFX	Erreur non spécifiée
		0010b	Erreur protection distante	05X	Version RDMAP invalide
				06X	OpCode inattendu
				07X	Erreur catastrophique, localisée au flux RDMAP
				08X	Erreur catastrophique, global
				09X	STag ne peut pas être Invalidée
				FFX	Erreur non spécifiée
0001b	DDP	Voir dans la spécification DDP [RFC5041] la description des valeurs et des noms.			
0010b	LLP (par exemple, MPA)	Pour MPA, voir dans la spécification MPA [RFC5044] la description des valeurs et noms.			

Figure 9 : Valeurs du champ Terminate Control

Réservé : 13 bits. Ce champ DOIT être réglé à zéro à l'émission, et ignoré à réception.

Longueur de segment DDP : 16 bits. Longueur passée par la couche DDP quand l'erreur a été détectée. Elle DOIT être valide si le bit M est établi. Elle DOIT être présente quand le bit D est établi.

En-tête DDP Terminated : 112 bits pour les messages étiquetés et 144 bits pour les messages non étiquetés. C'est l'en-tête DDP du message entrant qui est associé au Terminé. L'en-tête DDP n'est pas présent si le type d'erreur Terminate est une erreur locale catastrophique. Il DOIT être présent si le bit D est établi.

En-tête RDMA Terminated : 224 bits. L'en-tête RDMA Terminated n'est renvoyé que si le Terminé est associé à un message de demande RDMA Read. Il DOIT être présent si le bit R est établi. Si le Terminé se produit avant que le premier octet de la demande RDMA Read soit traité, l'en-tête original de la demande RDMA Read est renvoyé. Si le Terminé se produit après le traitement du premier octet de la demande RDMA Read, l'en-tête de demande RDMA Read est mis à jour pour refléter la situation actuelle de l'opération RDMA Read qui est en cours :

* STag de collecteur de données = la STag de collecteur de données envoyée à l'origine dans la demande RDMA Read.
* Décalage étiqueté du collecteur de données = décalage actuel dans la mémoire tampon étiquetée du collecteur de données. Par exemple, si la demande RDMA Read s'est terminée après l'envoi de 2048 octets, alors le décalage étiqueté du collecteur de données = le décalage étiqueté original du collecteur de données + 2048.
* Taille du message de données = nombre d'octets restants à transférer.
* STag de source de données = STag de la source de données dans la demande RDMA Read.
* Décalage étiqueté de la source de données = décalage actuel de la mémoire tampon étiquetée de la source de données. Par exemple, si la demande RDMA Read s'est terminée après l'envoi de 2048 octets, alors le décalage étiqueté de la source de données = décalage étiqueté original de la source de données + 2048.

Note : si un certain LLP ne définit pas de code de terminaison pour le message RDMAP Terminé à utiliser, aucun ne va être utilisé pour ce LLP.

La Figure 10, "Transposition de type d'erreur à message RDMA", transpose le nom de couche et les types d'erreur en chaque type de message RDMA :

Nom de couche	Nom de type d'erreur	Terminate inclut En-tête DDP et Longueur de Segment DDP	Terminate inclut en-tête RDMA	Quel type de message RDMA peut causer l'erreur
	Erreur locale catastrophique	Non	Non	Tous
RDMA	Erreur de protection distante	Oui, si possible	Oui	Seuls Demande RDMA Read, Send avec Invalidate, et Send avec SE et Invalidate
	Erreur d'opération distante	Oui, si possible	Non	Tous
DDP	Voir la [RFC5041]	Oui	Non	Tous
LLP	Voir la spéc. de LLP (par exemple, MPA)	Non	Non	Tous

Figure 10 : Transposition de type d'erreur en message RDMA

5. Transfert des données

5.1 Message RDMA Write

Un message RDMA Write est utilisé par la source de données pour transférer des données à une mémoire tampon étiquetée précédemment annoncée au collecteur de données. Le message RDMA Write a la sémantique suivante :

* Un message RDMA Write DOIT faire référence à une mémoire tampon étiquetée. C'est-à-dire, la couche RDMAP de la source de données DOIT demander que la couche DDP marque le message comme étiqueté.

* Un message RDMA Write valide NE DOIT PAS être livré à l'ULP du collecteur de données (c'est-à-dire, il est placé par la couche DDP).

* Chez l'homologue distant, quand un message RDMA Write invalide est livré à la couche RDMAP de l'homologue distant, une erreur est nettoyée (voir au paragraphe 7.1, "Nettoyage d'erreur RDMAP").

* Le décalage étiqueté d'une mémoire tampon étiquetée PEUT commencer à une valeur non à zéro.

* Un message RDMA Write PEUT cibler tout ou partie d'une mémoire tampon annoncée précédemment.

* RDMAP ne définit pas comment la ou les mémoires tampon sont utilisées par un RDMA Write sortant ou comment elles sont adressées. Par exemple, une mise en œuvre de RDMA peut choisir de permettre qu'une liste rassemblée de blocs de données non contigus soit la source d'un RDMA Write. Dans ce cas, les blocs de données vont être combinés par la source de données et envoyés comme un seul message RDMA Write au collecteur de données.

* La couche RDMAP de la source de données DOIT produire des messages RDMA Write à la couche DDP dans l'ordre de leur soumission par l'ULP.

* À la source de données, un message Send suivant (Send avec Invalidate, Send avec événement sollicité, ou Send avec événement sollicité et Invalidate) PEUT être utilisé pour signaler la livraison de messages RDMA Write précécents au collecteur de données, si l'ULP choisit de signaler la livraison de cette façon.

* Si l'homologue local souhaite écrire sur plusieurs mémoires tampon étiquetées sur l'homologue distant, l'homologue local DOIT utiliser plusieurs messages RDMA Write. C'est-à-dire, un seul message RDMA Write peut seulement écrire sur une mémoire tampon étiquetée distante.

* La source de données PEUT produire un message RDMA Write de longueur zéro.

5.2 Opération RDMA Read

L'opération RDMA Read DOIT consister en un seul message de demande RDMA Read et un seul message de réponse RDMA Read.

5.2.1 Message Demande RDMA Read

Une demande RDMA Read est utilisée par le collecteur de données pour transférer des données d'une mémoire tampon étiquetée précédemment annoncée à la source de données sur une mémoire tampon étiquetée au collecteur de données. Le message de demande RDMA Read a la sémantique suivante :

* Un message de demande RDMA Read DOIT faire référence à une mémoire tampon non étiquetée. C'est-à-dire, la couche RDMAP de l'homologue local DOIT demander que le DDP marque le message comme non étiqueté.

* Une message de demande RDMA Read DOIT consommer une mémoire tampon non étiquetée.

* La couche RDMAP de l'homologue distant DOIT traiter un message de demande RDMA Read. Un message de demande RDMA Read valide NE DOIT PAS être livré à l'ULP du collecteur de données (c'est-à-dire, il est traité par la couche RDMAP).

* À l'homologue distant, quand un message de demande RDMA Read invalide est livré à la couche RDMAP de l'homologue distant, une erreur est nettoyée (voir au paragraphe 7.1, "Nettoyage d'erreur RDMAP").

* Un message de demande RDMA Read DOIT faire référence à la file d'attente de demandes RDMA Read. C'est-à-dire, la couche RDMAP de l'homologue local DOIT demander que la couche DDP règle le champ Numéro de file d'attente à un.

* L'homologue local DOIT passer à la couche DDP les messages de demande RDMA Read dans l'ordre de leur soumission par l'ULP.

* L'homologue distant DOIT traiter les messages de demande RDMA Read dans l'ordre de leur envoi.

* Si l'homologue local souhaite lire à partir de plusieurs mémoires tampon étiquetées sur l'homologue distant, l'homologue local DOIT utiliser plusieurs messages de demande RDMA Read. C'est-à-dire, un seul message de demande RDMA Read DOIT seulement lire à partir d'une mémoire tampon étiquetée distante.

* Un message de demande RDMA Read PEUT cibler tout ou partie d'une mémoire tampon annoncée précédemment.

* Si la source de données reçoit un message de demande RDMA Read valide, elle DOIT répondre avec un message de réponse RDMA Read valide.

* Le collecteur de données PEUT produire un message de demande RDMA Read de longueur zéro en réglant le champ Taille de message RDMA Read à zéro dans l'en-tête Demande RDMA Read.

* Si la source de données reçoit un message RDMA Read d'une taille non zéro, la source de données RDMAP DOIT valider la STag de la source de données et le décalage étiqueté de source de données contenus dans l'en-tête de demande RDMA Read.

* Si la source de données reçoit un en-tête de demande RDMA Read avec la taille de message RDMA Read réglée à zéro, la source de données RDMAP :
  * NE DOIT PAS valider la STag de source de données et le décalage étiqueté de source de données contenus dans l'en-tête de demande RDMA Read, et
  * DOIT répondre avec un message de réponse RDMA Read de longueur zéro.

5.2.2 Message de réponse RDMA Read

Le message de réponse RDMA Read utilise le modèle de mémoire tampon étiquetée DDP pour livrer le contenu d'une mémoire tampon étiquetée de source de données précédemment demandée au collecteur de données, sans aucune implication de l'ULP à l'homologue distant. Le message de réponse RDMA Read a la sémantique suivante :

* Le message de réponse RDMA Read pour le message de demande RDMA Read associé voyage dans la direction opposée.

* Un message de réponse RDMA Read DOIT faire référence à une mémoire tampon étiquetée. C'est-à-dire, la couche RDMAP de la source de données DOIT demander que le DDP marque le message comme étiqueté.

* La source de données DOIT s'assurer qu'un nombre suffisant de mémoires tampon non étiquetées est disponible sur la file d'attente de demandes RDMA Read (file d'attente avec le numéro de file d'attente DDP 1) pour prendre en charge le nombre maximum de demandes RDMA Read négociées par l'ULP.

* La couche RDMAP DOIT livrer le message de réponse RDMA Read à l'ULP.

* À l'homologue distant, quand un message de réponse RDMA Read invalide est livré à la couche RDMAP de l'homologue distant, une erreur est nettoyée (voir au paragraphe 7.1, "Nettoyage d'erreur RDMAP").

* Le décalage étiqueté d'une mémoire tampon étiquetée PEUT commencer à une valeur non zéro.

* La couche RDMAP de la source de données DOIT passer les messages de réponse RDMA Read à la couche DDP, dans l'ordre de réception des messages de demande RDMA Read par la couche RDMAP, à la source de données.

* Le collecteur de données PEUT valider que la STag, le décalage étiqueté, et la longueur du message de réponse RDMA Read sont les mêmes que la STag, le décalage étiqueté, et la longueur inclus dans le message de demande RDMA Read correspondant.

* Un seul message de réponse RDMA Read DOIT écrire sur une mémoire tampon étiquetée distante. Si le collecteur de données souhaite lire plusieurs mémoires tampon étiquetées, le collecteur de données peut utiliser plusieurs messages de demande RDMA Read.

5.3 Type de message Send

Le type de message Send utilise le modèle de mémoire tampon non étiquetée DDP pour transférer les données de la source de données dans une mémoire tampon non étiquetée au collecteur de données.

* Un type de message Send DOIT faire référence à une mémoire tampon non étiquetée. C'est-à-dire, la couche RDMAP de l'homologue local DOIT demander que la couche DDP marque le message comme non étiqueté.

* Un type de message Send DOIT consommer une mémoire tampon non étiquetée.

    * Le message d'ULP envoyé en utilisant un type de message Send PEUT être inférieur ou égal à la taille de la mémoire tampon non étiquetée consommée. La couche RDMAP communique à l'ULP la taille des données écrites dans la mémoire tampon non étiquetée.

    * Si le message d'ULP envoyé via le type de message Send est supérieur à la mémoire tampon non étiquetée du collecteur de données, c'est une erreur (voir au paragraphe 9.1, "Nettoyage d'erreur RDMAP").

* À l'homologue distant, le type des messages Send DOIT être livré à l'ULP de l'homologue distant dans l'ordre de leur envoi.

* Après la livraison du message Send avec événement sollicité ou Send avec événement sollicité et Invalidate à l'ULP, le RDMAP PEUT générer un événement, si le collecteur de données est configuré à générer un tel événement.

* À l'homologue distant, quand un type de message Send invalide est livré à la couche RDMAP de l'homologue distant, une erreur est nettoyée (voir au paragraphe 7.1, "Nettoyage d'erreur RDMAP").

* Le RDMAP ne spécifie pas la structure de la ou des mémoires tampon utilisées par un RDMA Write sortant ni comment la ou les mémoires tampon sont adressées. Par exemple, une mise en œuvre de RDMA peut choisir de permettre qu'une liste rassemblant des blocs de données non contigus soit la source du type de message Send. Dans ce cas, les blocs de données vont être combinés par la source de données et envoyés comme un seul type de message Send au collecteur de données.

* Pour un type de message Send, la couche RDMAP de l'homologue local DOIT demander que la couche DDP règle le champ Numéro de file d'attente à zéro.

* L'homologue local DOIT produire les types de messages Send dans l'ordre de leur soumission par l'ULP.

* La source de données PEUT passer un type de message Send de longueur zéro. Un type de message Send de longueur zéro DOIT consommer une mémoire tampon non étiquetée au collecteur de données. Un message Send avec Invalidate ou Send avec événement sollicité et Invalidate DOIT faire référence à une STag. C'est-à-dire, la couche RDMAP de l'homologue local DOIT passer le champ Contrôle RDMA et la STag qui vont être invalidés à la couche DDP.

* Quand les messages Send avec Invalidate et Send avec événement sollicité et Invalidate sont livrés à la couche RDMAP de l'homologue distant, la couche RDMAP DOIT :
  * vérifier la STag associée au flux RDMAP et
  * invalider la STag si elle est associée au flux RDMAP; ou produire un message Terminé avec le code d'erreur de terminaison "La STag ne peut pas être invalidée", si la STag n'est pas associée au flux RDMAP.

5.4 Message Terminé

Le message Terminé utilise le modèle de mémoire tampon non étiquetée DDP pour transférer des informations relatives à l'erreur de la source de données dans une mémoire tampon non étiquetée au collecteur de données et ensuite cesser toute communication sur le flux DDP sous-jacent. Le message Terminé a la sémantique suivante :

* Un message Terminé DOIT faire référence à une mémoire tampon non étiquetée. C'est-à-dire, la couche RDMAP de l'homologue local DOIT demander que la couche DDP marque le message comme non étiqueté.

* Un message Terminé fait référence à la file d'attente Terminate. C'est-à-dire, la couche RDMAP de l'homologue local DOIT demander que la couche DDP règle le champ Numéro de file d'attente à deux.

* Un message Terminé DOIT consommer une mémoire tampon non étiquetée.

* Sur un seul flux RDMAP, la couche RDMAP DOIT garantir le placement d'un seul message Terminé.

* Un message Terminé DOIT être livré à la couche RDMAP de l'homologue distant. La couche RDMAP DOIT livrer le message Terminé à l'ULP.

* À l'homologue distant, quand un message Terminé invalide est livré à la couche RDMAP de l'homologue distant, une erreur est nettoyée (voir au paragraphe 7.1 "Nettoyage d'erreur RDMAP").

* La couche RDMAP aohève en erreur toutes les opérations d'ULP qui n'ont pas été fournies à la couche DDP.

* Après l'envoi d'un message Terminé sur un flux RDMAP, l'homologue local NE DOIT PAS envoyer d'autre message sur ce flux RDMAP spécifique.

* Après la réception d'un message Terminé sur un flux RDMAP, l'homologue distant PEUT arrêter d'envoyer des messages sur ce flux RDMAP spécifique.

5.5 Rangement et achèvement

Il est important de comprendre la différence entre placement et ordre de livraison car RDMAP donne aux deux notions une sémantique assez différente.

Noter que de nombreux protocoles courants, utilisés dans l'Internet et ailleurs, supposent que les données sont à la fois placées et livrées dans l'ordre. Tirer parti de ce fait permet aux applications de prendre divers raccourcis. Pour RDMAP, beaucoup de ces raccourcis ne sont plus d'utilisation sûre, et pourraient causer des échecs d'application.

Les règles suivantes s'appliquent aux mises en œuvre de RDMAP. Noter que dans ces règles, Send inclut Send, Send avec Invalidate, Send avec événement sollicité, et Send avec événement sollicité et Invalidate :

1. RDMAP n'assure pas d'ordre parmi les messages sur des flux RDMAP differents.

2. RDMAP n'assure pas d'ordre entre des opérations générées des deux extrémités d'un flux RDMAP.

3. Les messages RDMA qui utilisent des mémoires tampon étiquetées et non étiquetées PEUVENT être placés dans n'importe quel ordre. Si une application utilise des mémoires tampon qui se chevauchent (pointent sur des messages differents ou des portions d'un seul message dans la même mémoire tampon) il est alors possible que la dernière écriture entrante sur la mémoire tampon du collecteur de données ne soit pas les dernières données sortantes envoyées de la source de données.

4. Pour une opération Send, le contenu d'une mémoire tampon non étiquetée au collecteur de données PEUT être indéterminé jusqu'à ce que le Send soit livré à l'ULP au collecteur de données.

5. Pour une opération RDMA Write, le contenu de la mémoire tampon étiquetée au collecteur de données PEUT être indéterminé jusqu'à ce qu'un Send suivant soit livré à l'ULP chez le collecteur de données.

6. Pour une opération RDMA Read, le contenu de la mémoire tampon étiquetée au collecteur de données PEUT être indéterminé jusqu'à ce qu'un message de réponse RDMA Read ait été livré à l'homologue local.

Les déclarations 4, 5, et 6 impliquent qu'on ne "jette pas un coup d'œil" aux données pour voir si c'est fait. Il est possible que certaines données arrivent avant des données logiquement antérieures, et le "coup d'œil" peut causer une défaillance d'application imprévisible.

7. Si l'ULP ou l'application modifie le contenu des mémoires tampon étiquetées ou non étiquetées, qui sont en cours de modification par une opération RDMA tandis que RDMAP traite l'opération RDMA, l'état des mémoires tampon est indéterminé.

8. Si l'ULP ou l'application modifie le contenu des mémoires tampon étiqueées ou non étiquetées, qui sont lues par une opération RDMA tandis que RDMAP traite l'opération RDMA, le résultat de la lecture est indéterminé.

9. L'achèvement d'une opération RDMA Write ou Send chez l'homologue local ne garantit pas que le message d'ULP a déjà atteint la mémoire tampon d'ULP de l'homologue distant ou a été examinée par l'ULP distant.

10. Les messages Send DOIVENT être livrés à l'ULP chez l'homologue distant après qu'ils sont livrés à RDMAP par DDP et dans l'ordre de leur livraison à RDMAP.

    Noter que les règles d'ordre de DDP assurent que ce sera le même ordre que celui de leur soumission chez l'homologue local et que tous les RDMA Write antérieurs ont été soumis pour un placement ordonné chez l'homologue distant. Cela signifie que quand l'ULP voit la livraison du Send, les mémoires tampon ciblées par tout RDMA Write et Send précédent sont disponibles pour un accès local ou distant selon ce qui est autorisé. Si l'ULP fait chevaucher ses mémoires tampon pour différentes opérations, les données provenant du RDMA Write ou Send peuvent être écrasées par des opérations RDMA suivantes avant que l'ULP reçoive et traite la livraison.

11. Les messages de réponse RDMA Read DOIVENT être livrés à l'ULP chez l'homologue distant après qu'elles sont livrées à RDMAP par DDP et dans l'ordre de leur livraison à RDMAP. Les règles d'ordre de DDP assurent que ce sera le même ordre que celui de leur soumission à l'homologue local. Cela signifie que quand l'ULP voit la livraison de la réponse RDMA Read, les mémoires tampon ciblées par la réponse RDMA Read sont disponibles pour un accès local ou distant selon ce qui est autorisé. Si l'ULP fait chevaucher ses mémoires tampon pour des opérations différentes, les données provenant de la réponse RDMA Read peuvent être écrasées par des opérations RDMA suivantes avant que l'ULP reçoive et traite la livraison.

12. Les messages de demande RDMA Read, incluant les demandes RDMA Read de longueur zéro, NE DOIVENT PAS commencer le traitement chez l'homologue distant avant d'avoir été livrés à RDMAP par DDP.

    Note : l'ULP est assuré que les données écrites peuvent être lues. Par exemple,
    a) si une demande RDMA Read est produite par l'homologue local,
    b) si la demande cible la même mémoire tampon d'ULP qu'un RDMA Send ou Write précédent (dans la même direction que la demande RDMA Read) et
    c) si il n'y a pas d'autre source de mise à jour de la mémoire tampon d'ULP, alors l'homologue distant va renvoyer les données écrites par le RDMA Send ou Write. C'est-à-dire, pour cet exemple, que la mémoire tampon d'ULP annoncée pour être utilisée sur une série de messages RDMA, est seulement valide sur le flux RDMAP pour lequel elle est annoncée, et n'est pas mise à jour en local pendant que la série de messages RDMAP est effectuée. Pour cet

exemple, la règle d'ordre (12) assure que les accès suivants, en local ou à distance, à la mémoire tampon d'ULP contiennent les données écrites par le RDMA Send ou Write. Les messages de réponse RDMA Read PEUVENT être générés chez l'homologue distant après que les messages RDMA Write ou Send suivants ont été placés ou livrés. Donc, quand une application fait une demande RDMA Read suivie par un RDMA Write (ou Send) à la même mémoire tampon, elle peut obtenir les données du dernier RDMA Write (ou Send) dans le message de réponse RDMA Read, même si les opérations se sont achevées dans l'ordre chez l'homologue local. Si ce comportement n'est pas désiré, l'ULP d'homologue local doit barrer le dernier RDMA Write (ou Send) en retenant le message RDMA Write jusqu'à ce que toutes les réponses RDMA Read en instance aient été livrées.

13. La couche RDMAP DOIT soumettre les messages RDMA à la couche DDP dans l'ordre où les opérations RDMA sont soumises à la couche RDMAP par l'ULP.

14. Un message RDMA Send ou Write NE DOIT PAS être considéré achevé chez l'homologue local (source de données) jusqu'à ce qu'il ait été achevé avec succès à la couche DDP.

15. Les opérations RDMA DOIVENT être achevées chez l'homologue local dans l'ordre de leur soumission par l'ULP.

16. Au collecteur de données, un message Send entrant DOIT être livré à l'ULP seulement après que le message DDP a été livré à la couche RDMAP par la couche DDP.

17. Le traitement du message de réponse RDMA Read chez l'homologue distant (lire la mémoire tampon étiquetée spécifiée) DOIT être commencé seulement après que le message de demande RDMA Read a été livré par la couche DDP (donc, tous les messages RDMA précédents ont été correctement soumis pour un placement ordonné).

18. Les messages Send PEUVENT être achevés chez l'homologue distant (collecteur de données) avant que des messages de demande RDMA Read entrants précédents aient achevé leur traitement de réponse.

19. Une opération RDMA Read NE DOIT PAS être achevée chez l'homologue local jusqu'à ce que la couche DDP livre le message de réponse RDMA Read entrant associé.

20. Si plus d'un message de demande RDMA Read entrant est pris en charge par les deux homologues, les messages de réponse RDMA Read DOIVENT être soumis à la couche DDP chez l'homologue distant dans l'ordre de livraison des messages de demande RDMA Read par DDP, mais la lecture réelle du contenu de la mémoire tampon PEUT avoir lieu dans n'importe quel ordre chez l'homologue distant.

Cela simplifiele traitement d'achèvement de l'homologue local pour les RDMA Read en ce qu'une réponse RDMA Read livrée DOIT être suffisante pour achever l'opération RDMA Read.

6. Gestion de flux RDMAP

La gestion de flux RDMAP consiste en l'initialisation et la terminaison du flux RDMAP.

6.1 Initialisation du flux

L'initialisation du flux RDMAP survient après que le flux de LLP a été créé (par exemple, pour DDP/MPA sur TCP, le premier segment TCP après l'échange SYN, SYN/ACK). L'ULP est responsable de la transition du flux de LLP en mode à capacité RDMA. Le passage au mode RDMA se produit normalement parfois après l'établissement du flux de LLP. Une fois dans le mode à capacité RDMA, une mise en œuvre DOIT envoyer seulement des messages RDMA à travers le flux de transport jusqu'à ce que le flux RDMAP soit supprimé.

Pour chaque direction d'un flux RDMAP :

* Pour un flux RDMAP donné, le nombre de demandes RDMA Read en instance est limité par direction de flux RDMAP.

* Il est de la responsabilité de l'ULP de fixer le nombre maximum de demandes RDMA Read entrantes en instance par direction de flux RDMAP.

* La couche RDMAP DOIT fournir le nombre maximum de demandes RDMA Read entrantes en instance par direction

de flux RDMAP qui a été négocié entre l'ULP et la couche RDMAP de l'homologue local. Le mécanisme de négociation sort du domaine d'application de la présente spécification.

* Il est de la responsabilité de l'ULP de fixer le nombre maximum de demandes RDMA Read sortantes en instance par direction de flux RDMAP.

* La couche RDMAP DOIT fournir le nombre maximum de demandes RDMA Read sortantes en instance par direction de flux RDMAP qui a été négocié entre l'ULP et la couche RDMAP de l'homologue local. Le mécanisme de négociation sort du domaine d'application de la présente spécification.

* L'ULP de l'homologue local est chargé de négocier avec l'ULP de l'homologue distant le nombre maximum de demandes RDMA Read en instance pour la direction du flux RDMAP. Il est recommandé que l'ULP règle le nombre maximum de demandes RDMA Read entrantes en instance à être égal au nombre maximum de demandes RDMA Read en instance sortantes pour une direction de flux RDMAP donnée.

* Pour les demandes RDMA Read sortantes, la couche RDMAP NE DOIT PAS excéder le nombre maximum de demandes RDMA Read sortantes en instance qui ont été négociées entre l'ULP et la couche RDMAP de l'homologue local.

* Pour les demandes RDMA Read entrantes, la couche RDMAP NE DOIT PAS excéder le nombre maximum de demandes RDMA Read entrantes en instance qui a été négocié entre l'ULP et la couche RDMAP de l'homologue local.

6.2 Suppression de flux

Il y a trois méthodes pour terminer un flux RDMAP : terminaison d'ULP en douceur, terminaison RDMAP interruptive, et terminaison de LLP interruptive.

L'ULP est chargé d'effectuer la terminaison d'ULP en douceur. Après une terminaison d'ULP en douceur, l'un ou l'autre côté du flux peut initier la terminaison de LLP en douceur, en utilisant le mécanisme de terminaison en douceur fourni par le LLP.

La terminaison RDMAP interruptive permet à RDMAP de produire un message Terminé décrivant la raison de la terminaison du flux RDMAP. Le paragraphe suivant (6.2.1, "Terminaison RDMAP interruptive") décrit en détails la terminaison RDMAP interruptive.

La terminaison de LLP interruptive résulte d'une erreur de LLP et cause la suppression du flux RDMAP en cours, sans message RDMAP Terminé. Bien que cette dernière méthode soit indésirable, elle est possible, et l'ULP devrait la prendre en compte.

6.2.1 Terminaison RDMAP interruptive

RDMAP définit une opération Terminate qui DEVRAIT être invoquée quand une erreur RDMAP est rencontrée ou quand une erreur de LLP est nettoyée à la couche RDMAP par le LLP.

Il n'est pas toujours possible d'envoyer le message Terminé. Par exemple, certaines erreurs de LLP peuvent survenir qui causent la suppression du flux de LLP a) avant que RDMAP connaisse l'erreur, b) avant que RDMAP soit capable d'envoyer le message Terminé, ou c) après que RDMAP a envoyé le message Terminé au LLP, mais qu'il n'a pas encore été transmis par le LLP.

Noter qu'une terminaison RDMAP interruptive peut entraîner la perte de données. En général, quand un message Terminé est reçu, il est impossible de dire avec certitude quels messages RDMA non acquittés ont été achevés avec succès chez l'homologue distant. Donc, l'état de tous les messages RDMA en instance est indéterminé, et les messages DEVRAIENT être considérés comme achevés avec erreur.

Quand un homologue envoie ou reçoit un message Terminé, il PEUT immédiatement supprimer le flux de LLP. L'homologue DEVRAIT effectuer une suppression de LLP en douceur pour s'assurer que le message Terminé est bien livré.

Voir au paragraphe 4.8, "En-tête Terminate", la description du message Terminé et de son contenu. Voir au paragraphe 5.4,

"Message Terminé", la description de la sémantique du message Terminé.

7. Gestion d'erreur RDMAP

Le protocole RDMAP n'a pas d'opération de récupération d'erreur de couche RDMAP ou DDP incorporée. Si tout marche bien, les garanties de LLP vont assurer que les messages sont arrivés à destination.

Si des erreurs sont détectées à la couche RDMAP ou DDP, les flux RDMAP, DDP, et LLP sont terminés de façon interruptive (voir au paragraphe 4.8, "En-tête Terminate").

En général, de mauvaises mises en œuvre ou une programmation inappropriée d'ULP causent les erreurs détectées aux couches RDMAP et DDP. Dans ce cas, retourner un message de diagnostic d'erreur de terminaison et clore le flux RDMAP est bien plus simple que de tenter de maintenir le flux RDMAP, en particulier quand la cause de l'erreur n'est pas connue.

Si un LLP ne prend pas en charge la suppression d'un flux indépendamment des autres flux, et si une erreur RDMAP résulte en la terminaison d'un flux spécifique, alors le LLP DOIT étiqueter le flux comme erroné et NE DOIT PAS permettre d'autres transferts de données sur ce flux après que RDMAP a demandé que le flux soit supprimé.

Pour une connexion de LLP spécifique, quand tous les flux sont soit supprimés en douceur soit marqués comme flux erronés, la connexion de LLP DOIT être supprimée.

Comme les erreurs sont détectées chez l'homologue distant (éventuellement longtemps) après que les messages RDMA sont passés au DDP et LLP chez l'homologue local et après l'achèvement des opérations RDMA convoyées par les messages, l'envoyeur ne peut pas facilement déterminer lesquels de ses messages ont été reçus. (Les RDMA Read sont une exception à cette règle.)

Pour une liste des erreurs retournées à l'homologue distant par suite d'une terminaison interruptive, voir au paragraphe 4.8, "En-tête Terminate".

7.1. Nettoyage d'erreurs RDMAP

Si une erreur survient chez l'homologue local, la couche RDMAP DOIT tenter d'informer l'ULP local que l'erreur s'est produite.

L'homologue local DOIT envoyer un message Terminé pour chacun des cas suivants :

1. Pour les erreurs détectées lors de la création de demandes RDMA Write, Send, Send avec Invalidate, Send avec événement sollicité, Send avec événement sollicité et Invalidate, ou RDMA Read, ou pour d'autres raisons non directement associées à un message entrant, le message Terminé et un code d'erreur sont envoyés à la place de la demande. Dans ce cas, les champs Type d'erreur et Code d'erreur sont inclus dans le message Terminé, mais les champs d'en-tête DDP et RDMA Terminated sont réglés à zéro.

2. Pour les erreurs détectées sur un message entrant RDMA Write, Send, Send avec Invalidate, Send avec événement sollicité, Send avec événement sollicité et Invalidate, ou de réponse Read (après que le message a été livré par DDP) le message Terminé est envoyé à la première opportunité, de préférence dans le prochain message RDMA sortant. Dans ce cas, les champs Type d'erreur, Code d'erreur, Longueur de PDU d'ULP, et En-tête DDP Terminated sont inclus dans le message Terminé, mais le champ En-tête RDMA Terminated est réglé à zéro.

3. Pour les erreurs détectées sur un message de demande RDMA Read entrant (après que le message a été livré par DDP) le message Terminé est envoyé à la première opportunité, de préférence dans le prochain message RDMA sortant. Dans ce cas, les champs Type d'erreur, Code d'erreur, Longueur de PDU d'ULP, En-tête DDP Terminated, et En-tête RDMA Terminated sont inclus dans le message Terminé.

4. Si plus d'une erreur est détectée sur les messages RDMA entrants, avant que le message Terminé puisse être envoyé, alors le premier message RDMA (et son segment DDP associé) qui a rencontré une erreur DOIT être capturé par le message Terminé, en accord avec les règles 2 et 3 ci-dessus.

7.2 Erreurs détectées chez l'homologue distant sur les messages RDMA entrants

Sur les messages entrants RDMA Write, réponse RDMA Read, Send, Send avec Invalidate, Send avec événement sollicité, Send avec événement sollicité et Invalidate, et Terminate, on doit valider ce qui suit :
1. La couche DDP DOIT valider tous les champs de segment DDP.
2. Le OpCode RDMA DOIT être valide.
3. La version RDMA DOIT être valide.

De plus, sur les messages entrants Send avec Invalidate et Send avec événement sollicité et Invalidate, on doit aussi valider ce qui suit :
4. La STag Invalidate DOIT être valide.
5. La STag DOIT être associée à ce flux RDMAP.

Sur les messages de demande RDMA entrants, on doit valider ce qui suit :
1. La couche DDP DOIT valider tous les champs Segment DDP non étiqueté.
2. Le OpCode RDMA DOIT être valide.
3. La version RDMA DOIT être valide.
4. Pour les messages de demande RDMA Read de longueur non zéro :
   a. La STag de source de données DOIT être valide.
   b. La STag de source de données DOIT être associée à ce flux RDMAP.
   c. Le décalage étiqueté de la source de données DOIT tomber dans la gamme des décalages légaux associée à la STag de la source de données.
   d. La sommme du décalage étiqueté de la source de données et de la taille du message RDMA Read DOIT tomber dans la gamme des décalages légaux associée à la STag de la source de données.
   e. La somme du décalage étiqueté de la source de données et de la taille du message RDMA Read NE DOIT PAS causer le retour à zéro du décalage étiqueté de la source de données.

8. Considérations sur la sécurité

La présente Section fait références aux ressources qui discutent des considérations et implications de sécurité spécifiques du protocole de l'utilisation de RDMAP avec les services de sécurité existants. Une analyse détaillée des questions de sécurité autour de la mise en œuvre et l'utilisation de RDMAP se trouve dans la [RFC5042].

La [RFC5042] introduit le modèle de référence de RDMA et expose comment les ressources de ce modèle sont vulnérables aux attaques et les types d'attaques auxquelles ces vulnérabilités sont sujettes. Elle précise aussi les niveaux de confiance disponibles dans ce modèle d'homologue à homologue et comment cela définit la nature du partage de ressources.

Les exigences de IPsec pour RDDP se fondent sur la version de IPsec spécifiée dans la [RFC2401] et dans les RFC qui s'y rapportent, avec le profil de la [RFC3723], en dépit de l'existence d'une version plus récente de IPsec spécifiée dans la [RFC4301] et les RFC qui s'y rapportent [RFC4303], [RFC4306], [RFC4835]. Une des premières applications importantes des protocoles RDDP est leur utilisation avec iSCSI [RFC5046] ; les exigences IPsec de RDDP suivent celles de IPsec afin de faciliter cet usage en permettant l'utilisation d'un profil commun de IPsec avec iSCSI et les protocoles RDDP. À l'avenir, la RFC 3723 pourra être mise à jour avec une version plus récente de IPsec, et les exigences de sécurité de IPsec d'une telle mise à jour devraient s'appliquer uniformément à iSCSI et aux protocoles RDDP.

8.1 Résumé des exigences de sécurité spécifiques de RDMAP

La [RFC5042] définit les exigences de sécurité pour la mise en œuvre des composants du modèle de référence RDMA, à savoir le contrôleur d'interface réseau (NIC, *Network Interface Controller*) à capacité RDMA (RNIC) et le gestionnaire de ressource privilégié. Une mise en œuvre de RDMAP qui se conforme à la présente spécification DOIT se conformer à ces exigences.

8.1.1. Exigences pour RDMAP (RNIC)

RDMAP fournit plusieurs contre-mesures pour tous les types d'attaques mentionnés dans la [RFC5042]. Dans ce qui suit, la présente spécification fait la liste de toutes les exigences de sécurité qui DOIVENT être mises en œuvre par le RNIC. Une discussion plus détaillée des exigences de sécurité de RNIC se trouve à la Section 5 de la [RFC5042].

1. Un RNIC DOIT s'assurer qu'un flux spécifique dans un domaine de protection spécifique ne peut pas accéder à la STag dans un domaine de protection différent.

2. Un RNIC DOIT s'assurer que si une STag est limitée en portée à un seul flux, aucun autre flux ne peut utilisr la STag.

3. Un RNIC DOIT s'assurer qu'un homologue distant n'est pas capable d'accéder à une mémoire en dehors de la mémoire tampon spécifiée quand la STag a été activée pour l'accès à distance.

4. Un RNIC DOIT fournir un mécanisme pour que l'ULP établisse et révoque l'association d'une mémoire tampon d'ULP à une STag et une gamme de TO.

5. Un RNIC DOIT fournir un mécanisme pour que l'ULP établisse et révoque l'accès en lecture, écriture, ou lecture et écriture à la mémoire tampon d'ULP référencée par une STag.

6. Un RNIC DOIT s'assurer que l'interface réseau ne peut plus modifier une mémoire tampon annoncée après que l'ULP a révoqué les droits d'accès distants pour une STag.

7. Un RNIC DOIT s'assurer qu'un homologue distant n'est pas capable d'invalider une STag activée pour l'accès à distance, si la STag est partagée sur plusieurs flux.

8. Un RNIC DOIT choisir la valeur des STag d'une façon difficile à prédire. Il est RECOMMANDÉ de les remplir dispersées sur toute la gamme disponible.

9. Un RNIC NE DOIT PAS permettre le partage d'une file d'attente d'achèvement (CQ) à travers les ULP qui ne partagent pas une confiance mutuelle partielle.

10. Un RNIC DOIT s'assurer que si une CQ déborde, tous les flux qui n'utilisent pas la CQ DOIVENT rester non affectés.

11. Une mise en œuvre de RNIC DEVRAIT fournir un mécanisme pour contrôler le nombre de demandes RDMA Read en instance.

12. Un RNIC NE DOIT PAS activer un programme à charger sur le RNIC directement d'un homologue local ou homologue distant qui n'est pas de confiance, sauf si l'homologue est correctement authentifié*, et si la mise à jour est faite via un protocole sûr, tel que IPsec.

* par un mécanisme qui sort du domaine d'application de la présente spécification. Le mécanisme englobe probablement d'authentifier que l'ULP distant a le droit d'effectuer la mise à jour.

8.1.2 Exigences pour le gestionnaire de ressource privilégiée

Avec RDMAP, toutes les réservations de ressources locales sont initiées à partir des ULP locaux. Pour protéger contre des attaques locales incluant une distribution de ressource inéquitable et l'obtention d'un accès non autorisé aux ressources de RNIC, un gestionnaire de ressource privilégié (PRM, *Priviledged Resource Manager*) doit être mis en œuvre, qui gère toute l'allocation de ressources locale. Noter que le PRM ne doit pas être fourni comme un composant indépendant, et sa fonction peut aussi être mise en œuvre au titre de l'ULP privilégié ou au titre du RNIC lui-même.

Une mise en œuvre de PRM doit satisfaire les exigences de sécurité suivantes (une discussion plus détaillée des exigences de sécurité de PRM se trouve à la Section 5 de la [RFC5042]) :

1. Toutes les interactions d'ULP non privilégié avec le moteur RNIC qui pourraient affecter d'autres ULP DOIVENT être faites en utilisant le gestionnaire de ressources comme mandataire.

2. Toutes les demandes d'allocation de ressources d'ULP pour des ressources dispersées DOIVENT aussi être faites en utilisant un gestionnaire de ressource privilégié.

3. Le gestionnaire de ressource privilégié NE DOIT PAS supposer que des ULP différents partagent une confiance mutuelle partielle sauf si il y a un mécanisme pour assurer que les ULP partagent bien une confiance mutuelle partielle.

4. Si des ULP non privilégiés sont pris en charge, le gestionnaire de ressource privilégié DOIT vérifier que l'ULP non privilégié a le droit d'accéder à une mémoire tampon de données spécifique avant de permettre une STag pour laquelle

l'ULP a des droits d'accès à être associé à uue mémoire tampon de données spécifique.

5. Le gestionnaire de ressource privilégié DOIT contrôler l'allocation des entrées de CQ.

6. Le gestionnaire de ressource privilégié DEVRAIT empêcher un homologue local d'allouer plus que sa part équitable de ressources.

7. La consommation de ressource de file d'attente de demandes RDMA Read DOIT être contrôlée par le gestionnaire de ressource privilégié de telle façon que les flux RDMAP/DDP qui ne partagent pas de confiance mutuelle partielle ne partagent pas les ressources de file d'attente de demandes RDMA Read.

8. Si un RNIC fournit la capacité de partager des mémoires tampon de réception à travers plusieurs flux, la combinaison du RNIC et du gestionnaire de ressource privilégié DOIT être capable de détecter si l'homologue distant tente de consommer plus que sa part équitable de ressources afin que l'homologue local puisse appliquer des contre-mesures pour détecter et prévenir l'attaque.

8.2 Services de sécurité pour RDMAP

RDMAP utilise des services réseau fondés sur IP pour contrôler, lire, et écrire sur des mémoires tampon de données sur le réseau. Donc, tous les paquets de contrôle et de données échangés sont vulnérables à des attaques d'usurpation d'identité, d'altération, et de divulgation d'informations.

Les flux RDMAP qui sont soumis à des attaques d'usurpation d'identité ou de capture de flux peuvent être authentifiés, avoir une protection de leur intégrité et être protégés contre les attaques en répétition. De plus, la protection de la confidentialité peut être utilisée pour protéger contre l'espionnage.

8.2.1 Services de sécurité disponibles

La suite de protocoles IPsec [RFC2401] définit de fortes contre-mesures pour protéger un flux IP de ces attaques. Plusieurs niveaux de protection peuvent garantir la confidentialité d'une session, l'authentification de la source par paquet, l'intégrité par paquet, et le séquençage correct des paquets.

La sécurité de RDMAP peut aussi bénéficier des services de sécurité SSL ou TLS fournis pour les ULP fondés sur TCP [RFC4346]. Utilisés en dessous de RDMAP, ces services de sécurité fournissent aussi l'authentification du flux, l'intégrité des données, et la confidentialité. Comme exposé dans la [RFC5042], des limitations à la longueur maximum de paquet porté sur le réseau et au traitement potentiellement inefficace des paquets déclassés au collecteur de données rendent SSL et TLS moins appropriés que IPsec pour RDMAP.

Si SSL est mis en couche par dessus RDMAP, SSL ne protège pas les en-têtes RDMAP. Donc, une attaque par interposition peut toujours se produire en modifiant l'en-tête RDMAP pour placer incorrectement les données dans la mauvaise mémoire tampon, corrompant donc effectivement le flux de données.

En restant indépendant des protocoles de sécurité d'ULP et de LLP, RDMAP va bénéficier des améliorations continues sur ces couches. Les utilisateurs ont de la souplesse pour s'adapter à leurs exigences spécifiques de sécurité et la capacité de s'adapter aux défis de sécurité futurs. Cela étant, la vulnérabilité de RDMAP aux interférences actives de tiers n'est pas supérieure à celle de tout autre protocole fonctionnant sur un LLP comme TCP ou SCTP.

8.2.2 Exigences des services IPsec pour RDMAP

Parce que IPsec est conçu pour sécuriser des flux arbitraires de paquets IP, incluant des flux où des paquets sont perdus, RDMAP peut fonctionner par dessus IPsec sans aucun changement. Les paquets IPsec sont traités (par exemple, vérification de l'intégrité, et éventuellement déchiffrés) dans l'ordre de leur réception, et un collecteur de données RDMAP va traiter les messages RDMA déchiffrés contenus dans ces paquets de la même manière que les messages RDMA contenus dans des paquets IP non sécurisés.

Le groupe de travail "Mémorisation IP" a défini les exigences normatives de IPsec pour la mémorisation IP [RFC3723]. Des portions de cette spécification sont applicables à RDMAP. En particulier, une mise en œuvre conforme des services IPsec pour RDMAP DOIT satisfaire les exigences mentionnées au paragraphe 2.3 de la [RFC3723]. Sans reprendre en détails la discussion de la [RFC3723], cela inclut les exigences suivantes :

1. La mise en œuvre DOIT prendre en charge IPsec ESP [RFC2406], ainsi que les mécanismes de protection contre la répétition de IPsec. Quand ESP est utilisé, l'authentification par paquet de l'origine des données, la protection de l'intégrité, et contre la répétition DOIVENT être utilisées.

2. Elle DOIT prendre en charge ESP en mode tunnel et PEUT mettre en œuvre ESP en mode transport.

3. Elle DOIT prendre en charge IKE [RFC2409] pour l'authentification de l'homologue, la négociation des associations de sécurité, et la gestion de clés, en utilisant le DOI IPsec [RFC2407].

4. Elle NE DOIT PAS interpréter la réception d'un message de suppression IKE phase 2 comme une raison pour supprimer le flux RDMAP. Comme le matériel d'accélération IPsec peut seulement être capable de traiter un nombre limité de SA IKE phase 2 actives, les SA inactives peuvent être supprimées dynamiquement, et une nouvelle SA être remise à nouveau en place, si l'activité reprend.

5. Elle DOIT prendre en charge l'authentification de l'homologue en utilisant une clé pré-partagée, et PEUT prendre en charge l'authentification de l'homologue fondée sur le certificat en utilisant des signatures numériques. L'authentification de l'homologue en utilisant des méthodes de chiffrement de clé publique [RFC2409] NE DEVRAIT PAS être utilisée.

6. Elle DOIT prendre en charge IKE en mode principal et DEVRAIT prendre en charge le mode aggressif. IKE en mode principal avec authentification par clé pré-partagée NE DEVRAIT PAS être utilisé quand l'un des homologues utilise une adresse IP allouée de façon dynamique.

7. Quand des signatures numériques sont utilisées pour réaliser l'authentification, IKE en mode principal ou IKE en mode aggressif PEUT être utilisé. Dans ces cas, un négociateur IKE DEVRAIT utiliser une ou des charges utiles de demande de certificat IKE pour spécifier la ou les autorités de certification qui sont de confiance en accord avec sa politique locale. Les négociateurs IKE DEVRAIENT vérifier la liste de révocation de certificat (CRL, *Certificate Revocation List*) pertinente avant d'accepter un certificat de PKI à utiliser dans les procédures d'authentification de IKE.

8. L'accès à des informations secrètes mémorisées localement (clé pré-partagée ou privée pour la signature numérique) doit être convenablement contrôlé, car la compromission des informations secrètes annulle les propriétés de sécurité des protocoles IKE/IPsec.

9. Elle DOIT suivre les lignes directrices du paragraphe 2.3.4 de la [RFC3723] sur le réglage des paramètres de IKE afin d'obtenir un haut niveau d'interopérabilité sans exiger une configuration extensive.

De plus, la mise en œuvre et le déploiement des services IPsec pour RDDP devraient suivre les considérations sur la sécurité mentionnées à la Section 5 de la [RFC3723].

9. Considérations relatives à l'IANA

Le présent document ne demande pas d'action directe de la part de l'IANA. La considération qui suit est un simple commentaire.

Si RDMAP était activé à priori pour un ULP en se connectant sur un accès bien connu, cet accès serait enregistré auprès de l'IANA pour RDMAP. L'enregistrement de l'accès bien connu sera de la responsabilité de la spécification de l'ULP.

10. Références

10.1 Références normatives

[RFC0793] J. Postel (éd.), "Protocole de commande de transmission – Spécification du protocole du programme Internet DARPA", STD 7, septembre 1981. (*Remplacée par* RFC9293)

[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)

[RFC2406] S. Kent et R. Atkinson, "Encapsulation de charge utile de sécurité IP (ESP)", novembre 1998. *(Ob., voir* RFC4303)

[RFC2407] D. Piper, "Le domaine d'interprétation de sécurité IP de l'Internet pour ISAKMP", novembre 1998. *(Obs., voir* 4306)

[RFC2409] D. Harkins et D. Carrel, "L'échange de clés Internet (IKE)", novembre 1998. *(Obsolète, voir la* RFC4306)

[RFC3723] B. Aboba et autres, "Protocoles de sécurisation de mémorisation de blocs sur IP", avril 2004. *(P.S.)*

[RFC4960] R. Stewart, éd., "Protocole de transmission de commandes de flux (SCTP)", septembre 2007. (*Remplace* RFC2960, RFC3309 ; *P.S. ; Remplacée par* RFC9260)

[RFC5041] H. Shah et autres, "Placement direct des données sur transports fiables", octobre 2007. *(P.S. ; MàJ par* RFC7146*)*

[RFC5042] J. Pinkerton, E. Deleganes, "Sécurité du protocole de placement direct des données (DDP) / protocole d'accès direct à une mémoire distante (RDMAP)", octobre 2007. *(P.S. ; MàJ par* RFC7146*)*

[RFC5044] P. Culley et autres, "Tramage verrouillé sur la PDU de marqueur pour la spécification de TCP", octobre 2007. *(P.S. ; MàJ par* RFC6581*,* RFC7146*)*

[RFC5046] M. Ko et autres, "Extensions pour l'accès direct à une mémoire distante (RDMA) à l'interface système de petit ordinateur à l'Internet (iSCSI)", octobre 2007. *(P.S.) (Remplacée par* RFC7145)

10.2 Références pour information

[RFC4301] S. Kent et K. Seo, "Architecture de sécurité pour le protocole Internet", décembre 2005. *(P.S.) (Remplace la* RFC2401)

[RFC4303] S. Kent, "Encapsulation de charge utile de sécurité dans IP (ESP)", décembre 2005. (*Remplace* RFC2406) *(P.S.)*

[RFC4306] C. Kaufman, "Protocole d'échange de clés sur Internet (IKEv2)", décembre 2005. *(Obsolète, voir la* RFC5996)

[RFC4346] T. Dierks et E. Rescorla, "Protocole de sécurité de la couche Transport (TLS) version 1.1", avril 2006. *(Remplace* RFC2246 *; Remplacée par* RFC5246 *; MàJ par* RFC4366, 4680, 4681, 5746, 6176, 7465, 7507, 7919)

[RFC4835] V. Manral, "Exigences pour la mise en œuvre d'algorithme de chiffrement pour l'encapsulation de charge utile de sécurité (ESP) et l'en-tête d'authentification (AH)", avril 2007. (*Remplace* RFC4305) *(P.S.)*

Appendice A. Formats de segment DDP pour les messages RDMA

Cet Appendice est seulement pour information et NE FAIT PAS partie de la norme. Il décrit simplement le format de segment DDP pour les divers messages RDMA.

A.1 Segment DDP pour RDMA Write

La figure suivante décrit un segment DDP de RDMA Write :

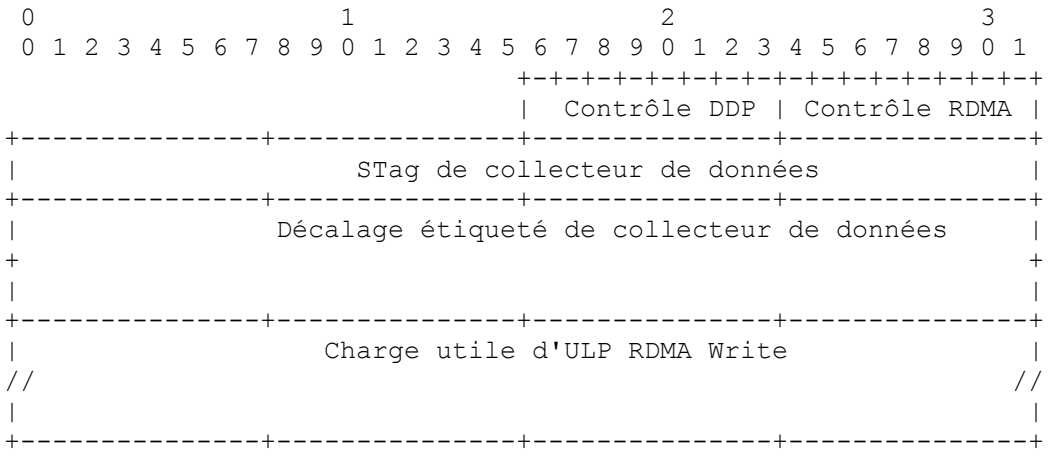

Figure 11 : Format de segment DDP de RDMA Write

A.2. Segment DDP pour demande RDMA Read

La figure suivante décrit un segment DDP de demande RDMA Read :

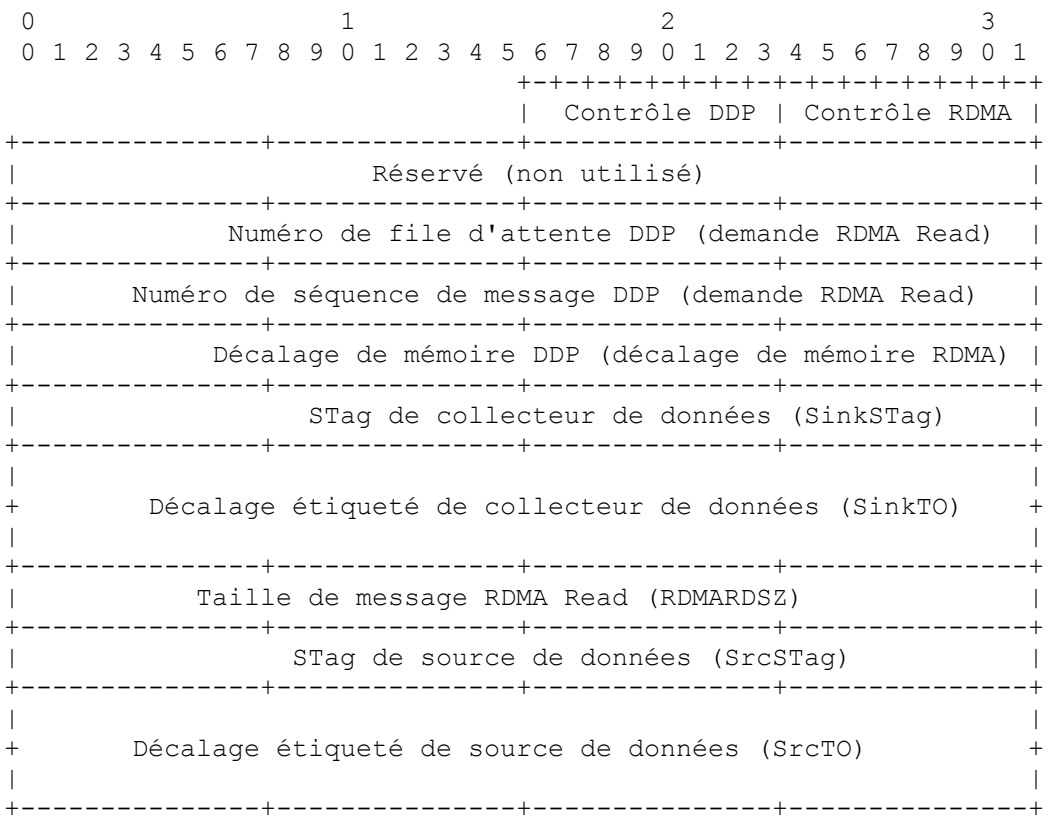

Figure 12 : Format de segment DDP de demande RDMA Read

A.3 Segment DDP pour réponse RDMA Read

La figure suivante décrit un segment DDP de réponse RDMA Read :

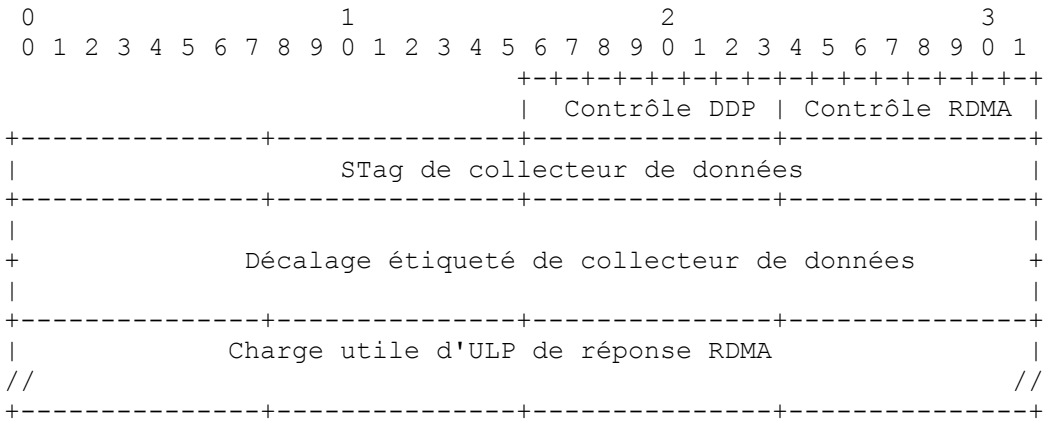

Figure 13 : Format de segment DDP de réponse RDMA Read

A.4 Segment DDP pour Send et Send avec événement sollicité

La figure suivante décrit un segment DDP de demande Send et Send avec événement sollicité :

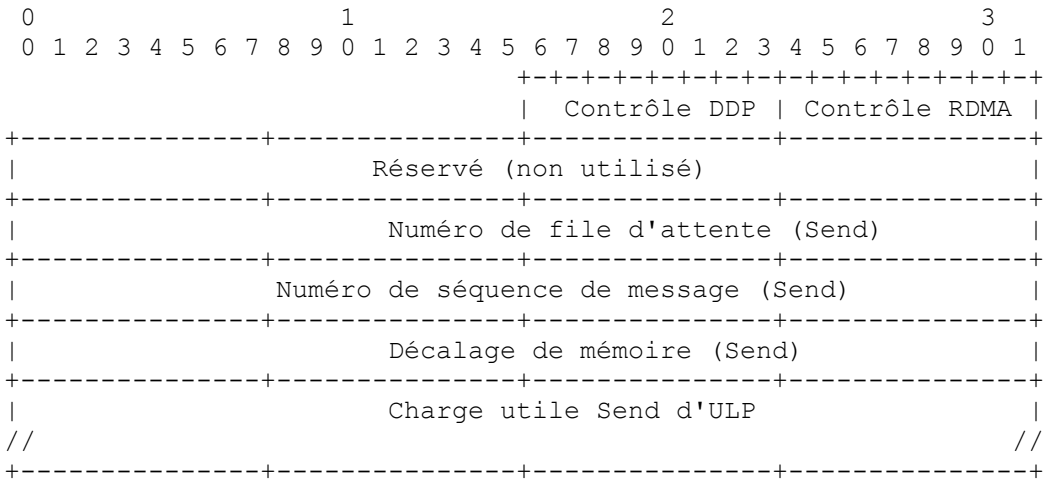

Figure 14 : Format de segment DDP de Send et Send avec événement sollicité

A.5. Segment DDP pour Send avec Invalidate et Send avec SE et Invalidate

La figure suivante décrit un segment DDP de demande Send avec Invalidate et Send avec SE et Invalidate :

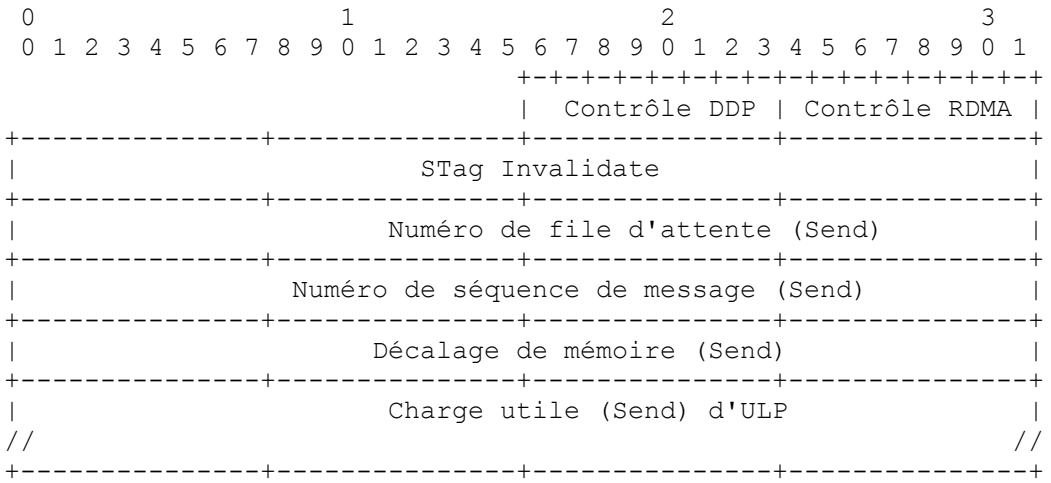

Figure 15 : Format de segment DDP de Send avec Invalidate et Send avec SE et Invalidate

A.6 Segment DDP pour Terminate

La figure suivante décrit un segment DDP de Terminate :

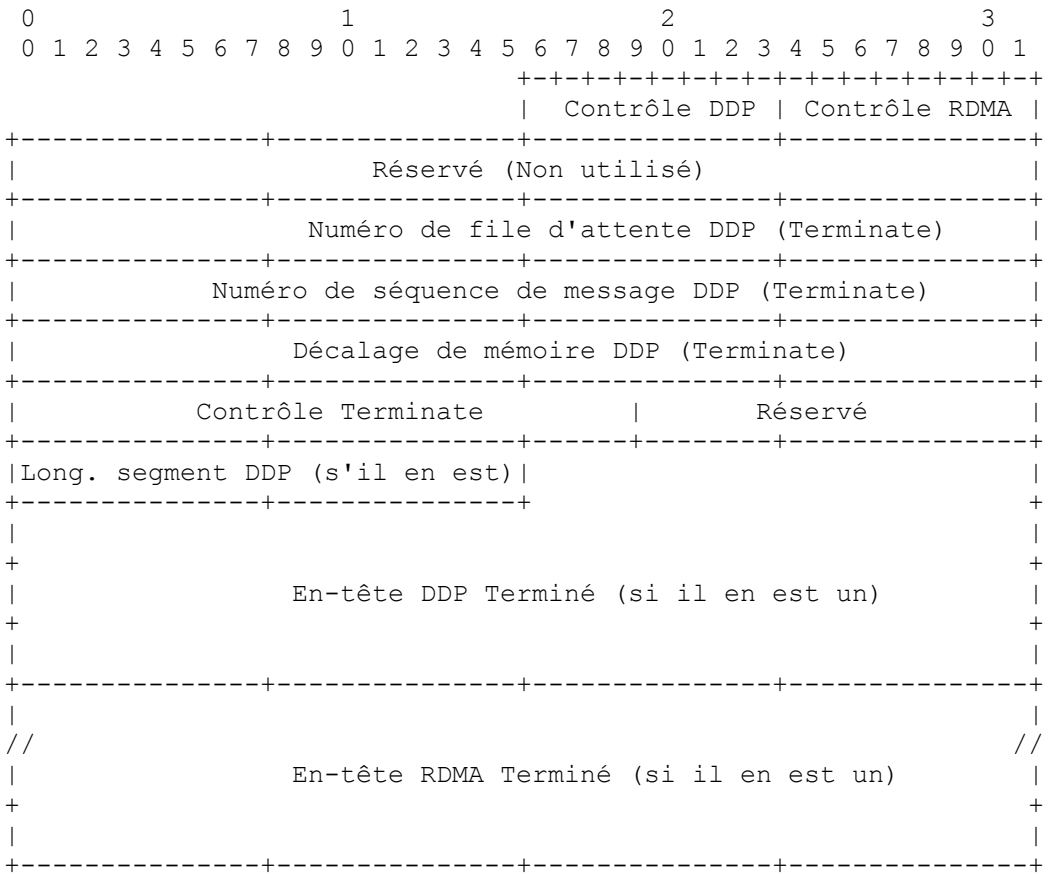

Figure 16 : Format de segment DDP Terminé

Appendice B. Tableau d'ordre et d'achèvement

Le tableau suivant résume les relations d'ordre qui sont définies au paragraphe 5.5, "Ordre et achèvement", du point de vue de l'homologue local qui produit les deux opérations. Noter que dans le tableau qui suit, Send inclut Send, Send avec Invalidate, Send avec événement sollicité, et Send avec événement sollicité et Invalidate.

Première opération	Opération ultérieure	Placement garanti à l'homologue distant	Placement garanti à l'homologue local	Ordre garanti à l'homologue distant
Send	Send	Pas de placement garanti. Si la garantie est nécessaire, voir note 1.	Non applicable	Achevé dans l'ordre.
Send	RDMA Write	Pas de placement garanti. Si la garantie est nécessaire, voir note 1.	Non applicable	Non applicable
Send	RDMA Read	Pas de placement garanti entre charge utile Send et en-tête de demande RDMA Read	La charge utile de réponse RDMA Read ne sera pas placée à l'homologue local jusqu'à ce que la charge utile Send soit placée chez l'homologue distant.	Le message de réponse RDMA Read ne va pas être généré tant que Send n'est pas achevé
RDMA Write	Send	Pas de placement garanti. Si la garantie est nécessaire, voir note 1.	Non applicable	Non applicable
RDMA Write	RDMA Write	Pas de placement garanti. Si la garantie est nécessaire, voir note 1.	Non applicable	Non applicable
RDMA Write	RDMA Read	Pas de placement garanti entre charge utile RDMA Write et en-	La charge utile de réponse RDMA Read ne sera pas placée	Non applicable

		tête de demande RDMA Read	chez l'homologue local jusqu'à ce que la charge utile RDMA Write soit placée chez l'homologue distant	
RDMA Read	Send	Pas de placement garanti entre en-tête de demande RDMA Read et charge utile Send	La charge utile Send peut être placée chez l'homologue distant avant que la réponse RDMA Read soit générée. Si la garantie est nécessaire, voir la note 2.	Non applicable
RDMA Read	RDMA Write	Pas de placement garanti entre en-tête de demande RDMA Read et charge utile RDMA Write	La charge utile RDMA Write peut être placed chez l'homologue distant avant que la réponse RDMA Read soit générée. Si la garantie est nécessaire, voir la note 2.	Non applicable
RDMA Read	RDMA Read	Pas de placement garanti des deux charges utile de réponse RDMA Read. De plus, il n'est pas garanti que les mémoires tampon étiquetées référencées dans le RDMA Read soient lues dans l'ordre.	Pas de placement garanti des deux charges utiles de réponse RDMA Read	La seconde réponse RDMA Read ne va pas être générée avant que la première réponse RDMA Read soit générée.

Figure 17 : Ordre des opérations

Note 1 : Si la garantie est nécessaire, un ULP peut insérer une opération RDMA Read et attendre qu'elle soit achevée pour agir comme une barrière.

Note 2 : Si la garantie est nécessaire, un ULP peut attendre l'achèvement de l'opération RDMA Read avant d'effectuer le Send.

Appendice C. Contributeurs

Dwight Barron, Hewlett-Packard Company ; mél : dwight.barron@hp.com
Caitlin Bestler, Broadcom Corporation ; mél : caitlinb@broadcom.com
John Carrier, Cray, Inc. ; mél : carrier@cray.com
Ted Compton, EMC Corporation ; mél : compton_ted@emc.com
Uri Elzur, Broadcom Corporation ; mél : Uri@Broadcom.com
Hari Ghadia, Gen10 Technology, Inc. ; mél : hghadia@gen10technology.com
Howard C. Herbert, Intel Corporation ; mél : howard.c.herbert@intel.com
Mike Ko, IBM ; mél : mako@us.ibm.com
Mike Krause, Hewlett-Packard Company ; mél : krause@cup.hp.com
Dave Minturn, Intel Corporation ; mél : dave.b.minturn@intel.com
Mike Penna, Broadcom Corporation ; mél : MPenna@Broadcom.com
Jim Pinkerton, Microsoft, Inc. ; mél : jpink@microsoft.com
Hemal Shah, Broadcom Corporation ; mél : hemal@broadcom.com
Allyn Romanow, Cisco Systems ; mél : allyn@cisco.com
Tom Talpey, Network Appliance ; mél : thomas.talpey@netapp.com
Patricia Thaler, Broadcom Corporation ; mél : pthaler@broadcom.com
Jim Wendt, Hewlett-Packard Company ; mél : jim_wendt@hp.com
Madeline Vega, IBM ; mél : mvega1@us.ibm.com
Claudia Salzberg, IBM ; mél : salzberg@us.ibm.com

Adresse des auteurs

Renato J. Recio
IBM Corp.

Bernard Metzler
IBM Research GmbH

Paul R. Culley
Hewlett-Packard Company

11501 Burnett Road
Austin, TX 78758 USA
téléphone : 512-838-3685
mél : recio@us.ibm.com

Zurich Research Laboratory
Saeumerstrasse 4
CH-8803 Rueschlikon, Switzerland
téléphone : +41 44 724 8605
mél : bmt@zurich.ibm.com

20555 SH 249
Houston, TX 77070-2698 USA
téléphone : 281-514-5543
mél : paul.culley@hp.com

Jeff Hilland
Hewlett-Packard Company
20555 SH 249
Houston, TX 77070-2698 USA
téléphone : 281-514-9489
mél : jeff.hilland@hp.com

Dave Garcia
24100 Hutchinson Rd.
Los Gatos, CA 95033 USA
téléphone : +1 (831) 247-4464
mél : Dave.Garcia@StanfordAlumni.org

Déclaration complète de droits de reproduction