Groupe de travail Réseau
Request for Comments : 5041
Catégorie : Sur la voie de la normalisation
Traduction Claude Brière de L'Isle

H. Shah, Broadcom Corporation
J. Pinkerton, Microsoft Corporation
R. Recio, IBM Corporation
P. Culley, Hewlett-Packard Company
octobre 2007

Placement direct des données sur transports fiables

Résumé
Le protocole de placement direct de données donne des informations pour placer les données entrantes directement dans la mémoire tampon de réception d'une couche supérieure de protocole sans mémoire tampon intermédiaire. Cela supprime les excès d'utilisation de CPU et de mémoire associés au transfert de données à travers des mémoires tampon intermédiaires.

Table des Matières

1. Introduction

Note : les majuscules en tête de certains mots dans ce document indiquent qu'ils sont utilisés avec la signification spécifique donnée dans le glossaire (Section 2).

Le protocole de placement direct de données (DDP, *Direct Data Placement Protocol*) permet à un protocole de couche supérieure (ULP, *Upper Layer Protocol*) d'envoyer des données à un collecteur de données sans exiger du collecteur de données qu'il place les données dans une mémoire tampon intermédiaire - donc, quand les données arrivent au collecteur de données, l'interface réseau peut placer les données directement dans la mémoire tampon de l'ULP. Cela peut permettre au collecteur de données de consommer substantiellement moins de mémoire qu'un modèle de mémoire tampon parce que le collecteur de données n'est pas obligé de déplacer les données de la mémoire tampon intermédiaire à la destination finale. De plus, cela peut permettre au protocole réseau de consommer substantiellement moins de cycles de CPU que si la CPU était utilisée pour déplacer les données, et ceci peut supprimer la limitation de bande passante de seulement être capable de déplacer les données aussi vite que la CPU peut copier les données.

DDP préserve les limites d'enregistrement d'ULP (les messages) tout en fournissant divers mécanismes de transfert de données et d'achèvement à utiliser pour transférer les messages d'ULP.

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

DDP a été conçu avec les objectifs architecturaux de haut niveau suivants :

* Fournir un modèle de mémoire tampon qui permette à l'homologue local d'annoncer une mémoire tampon désignée (c'est-à-dire, une étiquette pour une mémoire tampon) à l'homologue distant, de telle sorte qu'à travers le réseau, l'homologue distant puisse placer les données dans la mémoire tampon aux endroits spécifiés par l'homologue distant. Ceci est appelé le modèle de mémoire tampon étiquetée.

* Fournir un second modèle de mémoire tampon de réception qui préserve les limites de message d'ULP provenant de l'homologue distant et garde l'anonymat des mémoires tampon de l'homologue local (c'est-à-dire, non étiquetées). Ceci est appelé le modèle de mémoire tampon non étiquetée.

* Fournir une sémantique de livraison fiable, dans l'ordre pour les deux modèles de mémoire tampon étiquetée et non étiquetée.

* Fournir une segmentation et un réassemblage des messages d'ULP.

* Permettre que la mémoire tampon d'ULP soit utilisée comme mémoire tampon de réassemblage, sans besoin de copie, même si des segments DDP entrants arrivent dans le désordre. Ceci exige que le protocole sépare le placement de données de charge utile d'ULP contenues dans un segment DDP entrant de la livraison des données de messages d'ULP achevés.

* Si le protocole de couche inférieure (LLP) prend en charge plusieurs flux de LLP au sein d'une connexion de LLP, fournir les capacités ci-dessus indépendamment sur chaque flux de LLP et permettre la capacité d'être exporté à l'ULP sur la base du flux de LLP.

1.2 Vue d'ensemble du protocole

DDP prend en charge deux modèles de base de transfert de données : un modèle de transfert de données de mémoire tampon étiquetée et un modèle de transfert de données de mémoire tampon non étiquetée.

Le modèle de transfert de données de mémoire tampon étiquetée exige du collecteur de données qu'il envoie à la source de données un identifiant pour la mémoire tampon d'ULP, appelé une étiquette de pilotage (STag, *Steering Tag*). La STag est

transférée à la source de données en utilisant une méthode définie par l'ULP. Une fois que l'ULP de la source de données a une STag pour une mémoire tampon d'ULP de destination, il peut demander que DDP envoie les données d'ULP à la mémoire tampon d'ULP de destination en spécifiant la STag à DDP. Noter que la mémoire tampon étiquetée n'a pas besoin d'être remplie en commençant par le début de la mémoire tampon d'ULP. La source de données de l'ULP peut fournir un décalage arbitraire dans la mémoire tampon d'ULP.

Le modèle de transfert de données de mémoire tampon non étiquetée permet que le transfert de données se produise sans exiger du collecteur de données qu'il annonce une mémoire tampon d'ULP à la source de données. Le collecteur de données peut mettre en file d'attente une série de mémoires tampon de réception d'ULP. Un message DDP non étiqueté provenant de la source de données consomme une mémoire tampon non étiquetée au collecteur de données. Parce que DDP est en mode message, même si la source de données envoie une charge utile de message DDP plus petite que la mémoire tampon d'ULP de réception, la mémoire tampon d'ULP de réception partiellement remplie est livrée de toutes façons à l'ULP. Si la source de données envoie une charge utile de message DDP plus grande que la mémoire tampon de réception de l'ULP, il en résulte une erreur.

Il y a plusieurs différences clés entre les modèles de mémoire tampon étiquetée et non étiquetée :

* Pour le modèle de mémoire tampon étiquetée, la source de données spécifie quelle mémoire tampon étiquetée de réception va être utilisée pour un message DDP étiqueté spécifique (gestion de mémoire tampon d'ULP fondée sur l'envoyeur). Pour le modèle de la mémoire tampon non étiquetée, le collecteur de données spécifie l'ordre dans lequel les mémoires tampon non étiquetées vont être consommées lorsque des messages DDP non étiquetés sont reçus (gestion de mémoire tampon d'ULP fondée sur le receveur).

* Pour le modèle de mémoire tampon étiquetée, l'ULP au collecteur de données doit annoncer la mémoire tampon d'ULP à la source de données par un mécanisme spécifique de l'ULP avant que le transfert de données puisse intervenir. Pour le modèle de mémoire tampon non étiquetée, le transfert de données peut survenir sans une annonce explicite de bout en bout de mémoire tampon d'ULP. Noter cependant que l'ULP a besoin de traiter les problèmes de contrôle de flux.

* Pour le modèle de mémoire tampon étiquetée, un message DDP peut commencer à un décalage arbitraire au sein de la mémoire tampon étiquetée. Pour le modèle de la mémoire tampon non étiquetée, un message DDP peut seulement commencer au décalage 0.

* Le modèle de mémoire tampon étiquetée permet plusieurs messages DDP ciblés sur une mémoire tampon étiquetée avec une seule annonce de mémoire tampon d'ULP. Le modèle de mémoire tampon non étiquetée exige d'associer une mémoire tampon de réception d'ULP pour chaque message DDP ciblé à une mémoire tampon non étiquetée.

L'un et l'autre modèle de transfert de données place un message d'ULP dans un message DDP. Chaque message DDP est alors découpé en segments DDP qui sont destinés à tenir dans une unité de données de protocole de couche supérieure (MULPDU, *Maximum Upper Layer Protocol Data Unit*) d'un protocole de couche inférieure (LLP, *Lower Layer Protocol*). Donc, l'ULP peut envoyer des messages d'ULP de taille arbitraire, contenant jusqu'à 2^32 - 1 octets de charge utile d'ULP, et DDP découpe le message d'ULP en segments DDP, qui sont réassemblés de façon transparente au collecteur de données.

DDP fournit la livraison en ordre pour l'ULP. Cependant, DDP fait la différence entre livraison des données et placement de données. DDP donne suffisamment d'informations dans chaque segment DDP pour permettre que la charge utile d'ULP dans chaque charge utile de segment DDP entrant soit directement placée dans la mémoire tampon d'ULP correcte, même quand les segments DDP arrivent dans le désordre. Donc, DDP permet que le réassemblage de la charge utile d'ULP contenue dans les segments DDP d'un message DDP en un message d'ULP se produise dans la mémoire tampon d'ULP, éliminant donc la copie traditionnelle de la mémoire tampon de réassemblage dans la mémoire tampon d'ULP.

La charge utile d'un message DDP est livrée à l'ULP quand :
* tous les segments DDP d'un message DDP ont tous été complètement reçus, et que la charge utile du message DDP a été placée dans la mémoire tampon d'ULP associée,
* tous les messages DDP antérieurs ont été placés, et
* toutes les livraisons de message DDP antérieur ont été effectuées.

Le LLP sous DDP peut prendre en charge un seul flux de LLP de données par connexion (par exemple, TCP [RFC0793]) ou plusieurs flux de données de LLP par connexion (par exemple, SCTP [RFC4960]). Mais dans l'un et l'autre cas, DDP est spécifié de telle façon que chaque flux DDP soit indépendant et se transpose en un seul flux de LLP. Au sein d'un flux DDP spécifique, il est exigé du flux de LLP qu'il assure la livraison fiable dans l'ordre. Noter que DDP n'a pas de garantie d'ordre entre les flux DDP.

Un protocole DDP pourrait éventuellement fonctionner sur des LLP de livraison fiables ou non fiables. La présente spécification exige des LLP de livraison fiable dans l'ordre.

1.3 Mise en couche de DDP

DDP est destiné à être indépendant du LLP, sous réserve des exigences définies à la Section 3. Cependant, DDP a été spécifiquement défini comme faisant partie d'une famille de protocoles créés pour bien fonctionner ensemble, comme le montre la Figure 1, "Mise en couche de DDP". Pour les définitions de protocole de chaque LLP, voir "Tramage verrouillé de PDU de marqueur pour la spécification de TCP" [RFC5044] et "Adaptation du placement direct des données (DDP) au protocole de transmission de contrôle de flux (SCTP)" [RFC5043].

DDP permet la capacité du placement direct des données pour tout ULP, mais a été spécifiquement conçu pour bien fonctionner avec le protocole d'accès direct à une mémoire distante (RDMAP, *Remote Direct Memory Access Protocol*) [RFC5040], et fait partie de la suite de protocoles iWARP.

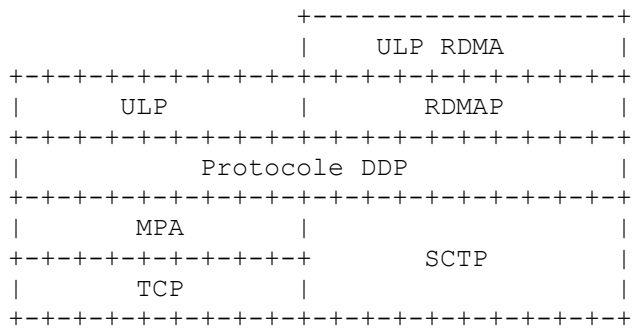

Figure 1 : Mise en couches de DDP

Si DDP est mis en couches en dessous de RDMAP et par dessus MPA et TCP, alors les en-têtes et charges utiles respectifs sont arrangés comme suit : (Note : Pour être clair, l'en-tête MPA et le CRC sont inclus, mais les marqueurs de tramage ne sont pas montrés).

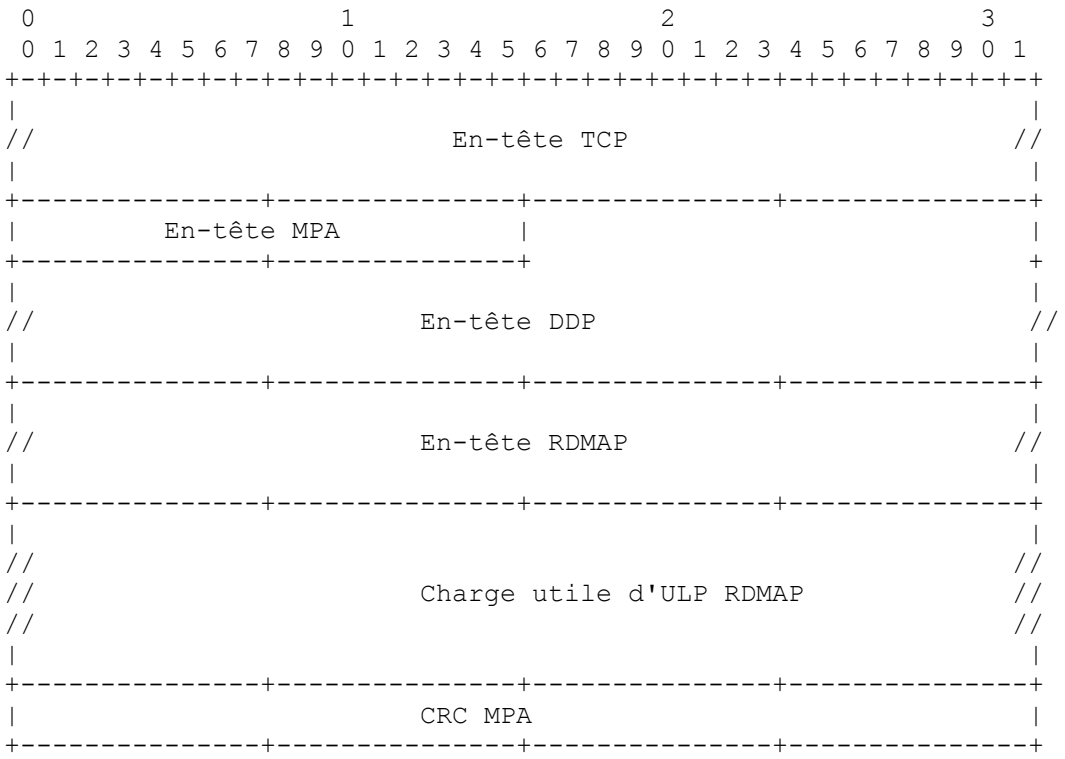

Figure 2 : Alignement d'en-tête MPA, DDP, et RDMAP

2. Glossaire

2.1 Général

Annonce (Annoncée, Annoncer, Annonces, Annonce) : acte d'informer un homologue distant qu'une mémoire tampon RDMA locale lui est disponible. Un nœud rend disponible une mémoire tampon RDMA pour un accès entrant RDMA Read ou RDMA Write en informant son homologue RDMA/DDP des identifiants de mémoire tampon étiquetée (STag, adresse de base, longueur). Cette annonce d'informations de mémoire tampon étiquetée n'est pas définie par RDMA/DDP et est laissée à l'ULP. Une méthode normale serait que l'homologue local incorpore la STage, l'adresse et la longueur de la mémoire tampon étiquetée dans un message Send destiné à l'homologue distant.

Livraison des données (livraison, livré, livre) : livraison est défini comme le processus d'informer l'ULP ou le consommateur que l'usage d'un message particulier est disponible. Ceci est spécifiquement différent du "Placement", qui peut généralement se produire dans n'importe quel ordre, tandis que l'ordre de "livraison" est strictement défini. Voir "placement de données".

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

Source de données : c'est 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 (Livré, Livre) : voir livraison des données ci-dessus.

iWARP : suite de protocoles réseau composée de RDMAP [RFC5040], DDP (la présente spécification), et Tramage verrouillé de PDU de marqueur pour TCP (MPA) [RFC5044]. 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 pour la description d'un échange de protocole ou autre interaction entre deux nœuds.

Nœud : appareil de calcul rattaché à une ou plusieurs liaisons d'un réseau. Un nœud dans ce contexte ne se réfère pas à une instanciation spécifique d'application ou protocole fonctionnant sur l'ordinateur. Un nœud peut consister en un ou plusieurs contrôleurs d'interface réseau à capacité RDMA (RNIC, *RDMA Enabled Network Interface Controller*) installés sur un ordinateur hôte.

Placement (Placé, Place) : voir "Placement de données" au paragraphe 2.3.

Homologue distant : mise en œuvre du protocole RDMA/DDP sur le côté opposé de la connexion. Utilisé pour se référer à l'entité distante pour décrire les échanges de protocole ou autres interactions entre deux nœuds.

RNIC *(RDMA Enabled Network Interface Controller)* : contrôleur d'interface réseau à capacité RDMA. Dans ce contexte, cela va être un adaptateur d'entrée/sortie réseau ou un contrôleur incorporé avec une fonction iWARP.

ULP *(Upper Layer Protocol)* : protocole de couche supérieure. C'est la couche de protocole au dessus de la couche de protocole actuellement référencée. L'ULP pour RDMA/DDP est supposé être un système d'exploitation (OS, *Operating System*) une application, une couche d'adaptation, ou un appareil propriétaire. Les documents RDMA/DDP ne spécifient pas d'ULP ; ils fournissent un ensemble de sens qui permettent qu'un ULP soit conçu pour utiliser RDMA/DDP.

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

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

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 l'adaptation SCTP DDP, 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 de LLP de l'homologue sur deux nœuds.

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

MULPDU *(Maximum Upper Layer Protocol Data Unit)* unité de données maximum de protocole de couche supérieure : 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

Placement de données (Placement, Placé, Place) : 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 receveuse de DDP pour effectuer le placement de données de la charge utile d'ULP du segment DDP. Voir "Livraison des données" et "Placement direct de données".

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

Suppression DDP en douceur : acte de clôture d'un flux DDP de façon à permettre que tous les messages DDP en cours et en instance s'achèvent avec succès.

Champ Contrôle DDP : champ fixe de 8 bits dans l'en-tête DDP.

En-tête DDP : en-tête présent dans tous les segments DDP. L'en-tête DDP contient des champs de contrôle et de placement qui sont utilisés pour définir la localisation finale de placement pour 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 se produire pour diverses raisons, incluant une 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. Elle inclut un en-tête DDP et la charge utile d'ULP (si elle est présente). Un segment DDP devrait être dimensionné pour tenir dans la MULPDU du protocole de couche inférieure.

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 n'a pas de garanties d'ordre entre les flux DDP.

Identifiant de flux DDP : un identifiant pour un 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 l'ULP. Ceci peut se produire même quand les segments DDP arrivent en désordre. La prise en charge du placement en désordre peut exiger que le collecteur de données mette en œuvre le LLP et DDP comme un bloc fonctionnel.

Protocole de placement direct de données (DDP) : c'est aussi un protocole du réseau qui prend en charge le placement direct de données en associant des informations explictes de placement de mémoire tampon avec les unités de charge

utile de LLP.

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

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

Domaine de protection (PD) : mécanisme utilisé pour associer un flux DDP et une STag. Avec ce mécanisme, l'utilisation d'une STag est valide sur un flux DDP si la STag a le même identifiant de domaine de protection (PD ID) que le flux DDP.

Identifiant de domaine de protection (PD ID) : identifiant du domaine de protection.

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

Étiquette de pilotage *(STag, Steering Tag)* : identifiant d'une mémoire tampon étiquetée sur un nœud, valide comme défini dans une spécification de protocole.

Mémoire tampon étiquetée : mémoire tampon qui est explicitement annoncée à l'homologue distant par l'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 d'ULP : mémoire tampon située au dessus de la couche DDP et annoncée à la couche DDP comme une mémoire tampon étiquetée ou comme une mémoire tampon d'ULP non étiquetée.

Longueur de message d'ULP : longueur totale, en octets, de la charge utile d'ULP contenue dans un message DDP.

Mémoire tampon non étiquetée : mémoire tampon qui n'est pas explicitement annoncée à l'homologue distant.

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.

3. Exigences de LLP de livraison fiable

Tout protocole qui peut servir de LLP pour DDP DOIT satisfaire les exigences suivantes :

1. Les LLP DOIVENT exposer les MULPDU et les changements de MULPDU. Ceci est exigé pour que la couche DDP puisse effectuer une segmentation alignée sur la MULPDU et puisse s'adapter lorsque des changements de MULPDU se font jour. Le cas particulier de comment traiter les demandes en instance durant un changement de MULPDU est couvert par les exigences ci dessous.

2. Dans le cas d'un changement de MULPDU, le LLP NE DOIT PAS exiger de DDP qu'il resegmente les segments DDP qui ont été précédemment envoyés au LLP. Noter que dans des conditions extrêmes, le LLP peut changer la MULPDU annoncée plus fréquemment que n'est purgée la file d'attente des demandes de segments DDP précédemment transmises. Dans ces conditions extrêmes, la file d'attente de transmission du LLP peut contenir des messages DDP pour lesquels plusieurs mises à jour de la MULPDU correspondante se sont produites suite à l'envoi de messages. Donc, il ne peut pas y avoir de corrélation entre les segments DDP en file d'attente et la valeur courante de MULPDU du LLP.

3. Le LLP DOIT s'assurer que, si il accepte un segment DDP, il va le transférer de façon fiable au receveur ou le retourner

avec une erreur déclarant que le transfert a échoué à s'achever.

4. Le LLP DOIT préserver les limites de segment et message DDP chez le collecteur de données.

5. Le LLP PEUT fournir les segments entrants dans le désordre pour le placement, mais si il le fait, il DOIT aussi fournir des informations qui spécifient quel ordre a spécifié l'envoyeur.

6. Le LLP DOIT fournir un résumé fort (au moins équivalent au CRC32-C) pour couvrir au moins le segment DDP. On estime que certains des résumés de données existants ne sont pas suffisants, et que la sémantique de transfert direct en mémoire exige un résumé plus fort que, par exemple, une simple somme de contrôle.

7. À réception, le LLP DOIT fournir la longueur du segment DDP reçu. Cela assure que DDP n'a pas à porter un champ de longueur dans son en-tête.

8. Si un LLP ne prend pas en charge la suppression d'un flux de LLP indépendamment des autres flux du LLP, et si une erreur de DDP se produit sur un flux DDP spécifique, alors le LLP DOIT étiqueter le flux de LLP associé comme un flux de LLP erroné et NE DOIT PAS permettre d'autre transfert de données sur ce flux de LLP après que DDP a demandé que le flux DDP associé soit supprimé.

9. Pour un flux de LLP spécifique, le LLP DOIT fournir un mécanisme pour indiquer que le flux de LLP a été supprimé en douceur. Pour une connexion de LLP spécifique, le LLP DOIT fournir un mécanisme pour indiquer que la connexion de LLP a été supprimée en douceur. Noter que si le LLP ne permet pas qu'un flux de LLP soit supprimé indépendamment de la connexion de LLP, les exigences ci-dessus permettent au LLP de notifier à DDP les deux événements en même temps.

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

11. Le LLP NE DOIT PAS passer un segment DDP dupliqué à la couche DDP après qu'il a passé tous les segments DDP précédents et les informations d'ordre associées pour les segments DDP précédents et le segment DDP en cours à la couche DDP.

4. Format d'en-tête

DDP a deux formats d'en-tête différents : un pour le placement de données dans les mémoires tampon étiquetées, et l'autre pour le placement de données dans des mémoires tampon non étiquetées. Voir au paragraphe 5.1 la description des deux modèles.

4.1 Champs de contrôle DDP

Les 8 premiers bits de l'en-tête DDP portent un champ Contrôle DDP qui est commun aux deux formats. Il est montré à la Figure 3, avec un décalage 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.

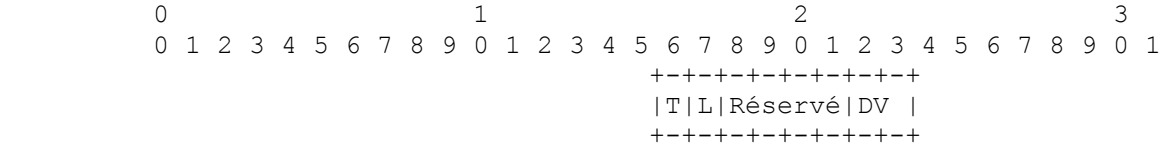

Figure 3: contrôle DDP Field

T *(Tagged)* : fanion Étiqueté : 1 bit. Spécifie le modèle de mémoire tampon étiquetée ou non étiquetée. Si il est réglé à un, la charge utile d'ULP portée dans ce segment DDP DOIT être placée dans une mémoire tampon étiquetée. Si il est réglé à zéro, la charge utile d'ULP portée dans ce segment DDP DOIT être placée dans une mémoire tampon non étiquetée.

L (Last) : fanion Dernier : 1 bit. Spécifie si le segment DDP est le dernier segment d'un message DDP. Il DOIT être réglé à un sur le dernier segment DDP de chaque message DDP. Il NE DOIT PAS être réglé à un sur les autres segments DDP. Le segment DDP avec le bit L établi à 1 DOIT être envoyé au LLP après que tous les autres segments DDP du message

DDP associé ont été envoyés au LLP. Pour un message DDP non étiqueté, le segment DDP avec le bit L réglé à 1 DOIT porter le plus fort MO. Si le fanion Last est réglé à un, la charge utile du message DDP DOIT être livrée à l'ULP après :
- o le placement de tous les segments DDP de ce message DDP et e tous les messages DDP antérieurs , et
- o la livraison de chaque message DDP antérieur.

Si le fanion Last est réglé à zéro, le segment DDP est un segment DDP intermédiaire.

Réservé : 4 bits. Réservé pour une future utilisation par le protocole DDP. Ce champ DOIT être mis à zéro à l'émission, et non vérifié à réception.

DV : Version du protocole de placement direct de données : 2 bits. La version du protocole DDP utilisée. Ce champ DOIT être réglé à un pour indiquer la version de la spécification décrite dans le présent document. La valeur de DV DOIT être la même pour tous les segments DDP transmis ou reçus sur un flux DDP.

4.2 En-tête DDP du modèle de mémoire tampon étiquetée

La Figure 4 montre le format d'en-tête DDP qui DOIT être utilisé dans tous les segments DDP qui ciblent les mémoires tampon étiquetées. Il inclut le champ Contrôle DDP défini au paragraphe 4.1. (Note : dans la Figure 4, 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 n'est présent que si DDP est mis en couche par dessus MPA.)

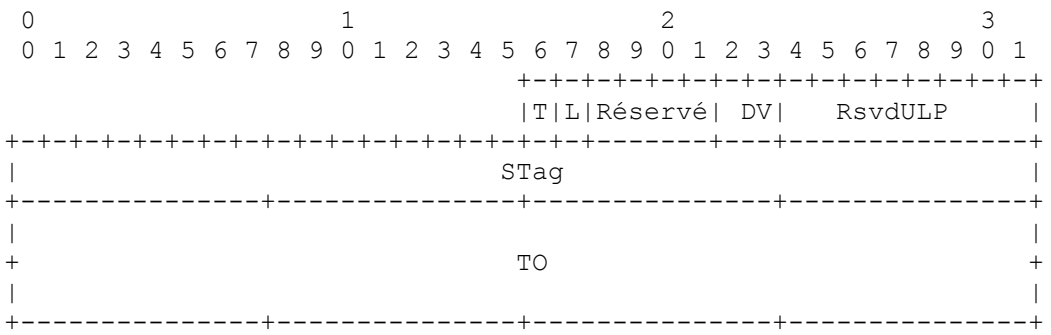

Figure 4 : en-tête DDP de mémoire tampon étiquetée

T est réglé à un.

RsvdULP, réservé à l'usage de l'ULP : 8 bits. Le champ RsvdULP est opaque au protocole DDP et peut être structuré de n'importe quelle façon par l'ULP. À la source de données, DDP DOIT régler le champ RsvdULP à la valeur spécifiée par l'ULP. Il est transféré non modifié de la source de données au collecteur de données. Au collecteur de données, DDP DOIT fournir le champ RsvdULP à l'ULP quand le message DDP est livré. Chaque segment DDP au sein d'un message DDP spécifique DOIT contenir la même valeur pour ce champ. La source de données DOIT s'assurer que chaque segment DDP au sein d'un message DDP spécifique contient la même valeur pour ce champ.

STag, étiquette de pilotage : 32 bits. L'étiquette de pilotage identifie la mémoire tampon étiquetée du collecteur de données. La STag DOIT être valide pour ce flux DDP. La STag est associée au flux DDP par un mécanisme qui sort du domaine d'application de la spécification du protocole DDP. À la source de données, DDP DOIT régler le champ STag à la valeur spécifiée par l'ULP. Au collecteur de données, le DDP DOIT fournir le champ STag quand le message d'ULP est livré. Chaque segment DDP au sein d'un message DDP spécifique DOIT contenir la même valeur pour ce champ et DOIT être la valeur fournie par l'ULP. La source de données DOIT s'assurer que chaque segment DDP au sein d'un message DDP spécifique contient la même valeur pour ce champ.

TO - décalage étiqueté : 64 bits. Le décalage étiqueté spécifie le décalage, en octets, au sein de la mémoire tampon étiquetée du collecteur de données, où commence le placement de la charge utile d'ULP contenue dans le segment DDP. Un message DDP PEUT commencer à un TO arbitraire au sein d'une mémoire tampon étiquetée.

4.3 En-tête DDP du modèle de mémoire tampon non étiquetée

La Figure 5 montre le format de l'en-tête DDP qui DOIT être utilisé dans tous les segments DDP qui ciblent les mémoires tampon non étiquetées. Il inclut le champ Contrôle DDP défini au paragraphe 4.1. (Note : dans la Figure 5, 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 n'est présent que si

DDP est mis en couche par dessus MPA.)

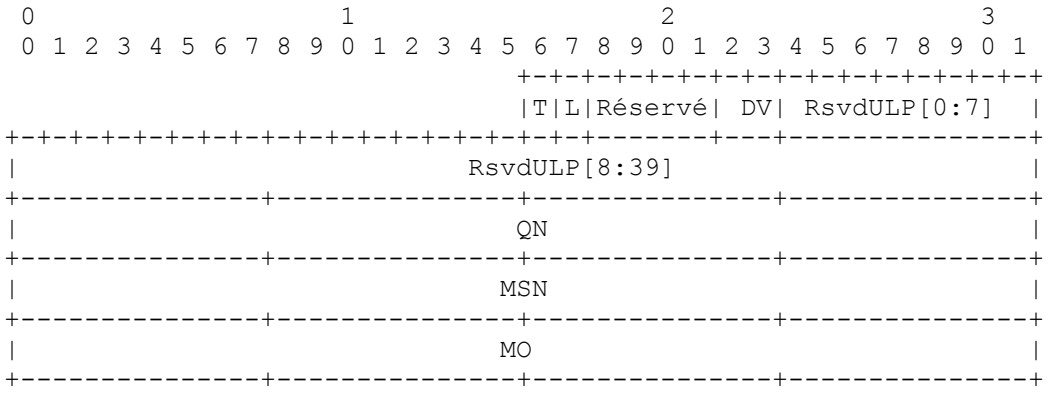

Figure 5 : en-tête DDP de mémoire tampon non étiquetée

T est réglé à zéro.

RsvdULP, réservé à l'usage de l'ULP : 40 bits. Le champ RsvdULP est opaque pour le protocole DDP et peut être structuré de n'imprte quelle façon par l'ULP. À la source de données, DDP DOIT régler le champ RsvdULP à la valeur spécifiée par l'ULP. Elle est transférée non modifiée de la source de données au collecteur de données. Au collecteur de données, DDP DOIT fournir le champ RsvdULP à l'ULP quand le message d'ULP est livré. Chaque segment DDP au sein d'un message DDP spécifique DOIT contenir la même valeur pour le champ RsvdULP. Au collecteur de données, la mise en œuvre de DDP N'EST PAS OBLIGÉE de vérifier que la même valeur est présente dans le champ RsvdULP de chaque segment DDP au sein d'un message DDP spécifique et PEUT fournir la valeur à partir de tout segment DDP reçu à l'ULP quand le message d'ULP est livré.

QN, numéro de file d'attente : 32 bits. Le numéro de file d'attente identifie la file d'attente de mémoire tampon non étiquetée du collecteur de données référencé par cet en-tête. Chaque segment DDP au sein d'un message DDP spécifique DOIT contenir la même valeur pour ce champ et DOIT être la valeur fournie par l'ULP à la source de données. La source de données DOIT s'assurer que chaque segment DDP au sein d'un message DDP spécifique contient la même valeur pour ce champ.

MSN, numéro de séquence de message : 32 bits. Le numéro de séquence de message spécifie un numéro de séquence qui DOIT être augmenté de un (modulo 2^32) à chaque message DDP ciblant le numéro de file d'attente spécifique sur le flux DDP associé à ce segment DDP. La valeur initiale pour le MSN DOIT être un. La valeur de MSN DOIT revenir à 0 après la valeur de 0xFFFFFFFF. Chaque segment DDP au sein d'un message DDP spécifique DOIT contenir la même valeur pour ce champ. La source de données DOIT s'assurer que chaque segment DDP au sein d'un message DDP spécifique contient la même value pour ce champ.

MO,- décalage de mémoire : 32 bits. Le décalage de mémoire spécifie le décalage, en octets, depuis le début du message DDP représenté par le MSN et le numéro de file d'attente sur le flux DDP associé à ce segment DDP. Le MO qui référence le premier octet du message DDP DOIT être réglé à zéro par la couche DDP.

4.4 Format de segment DDP

Chaque segment DDP DOIT contenir un en-tête DDP. Chaque segment DDP peut aussi contenir une charge utile d'ULP. Voici le format de segment DDP :

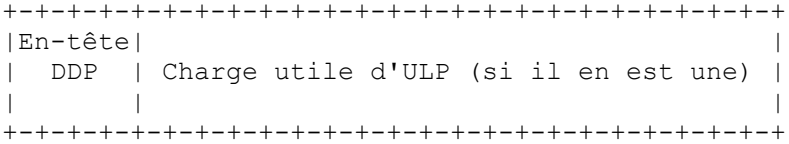

Figure 6 : Format de segment DDP

5. Transfert des données

DDP prend en charge des messages DDP multi-segments. Chaque message DDP est composé de un ou plusieurs segments DDP. Chaque segment DDP contient un en-tête DDP. L'en-tête DDP contient les informations requises par le receveur pour placer toute charge utile d'ULP incluse dans le segment DDP.

5.1 Modèles de mémoire tampon DDP étiquetée ou non étiquetée

DDP utilise deux modèles de base de mémoire tampon pour le placement de la charge utile d'ULP : le modèle de mémoire tampon étiquetée, et le modèle de mémoire tampon non étiquetée.

5.1.1 Modèle de mémoire tampon étiquetée

Le modèle de mémoire tampon étiquetée est utilisé par la source de données pour transférer au collecteur de données un message DDP dans une mémoire tampon étiquetée qui a été préalablement annoncée à la source de données. Une STag identifie une mémoire tampon étiquetée. Pour le placement d'un message DDP utilisant le modèle de mémoire tampon étiquetée, la STag est utilisée pour identifier la mémoire tampon, et le TO est utilisé pour identifier le décalage au sein de la mémoire tampon étiquetée dans laquelle la charge utile d'ULP est transférée. Le protocole utilisé pour annoncer la mémoire tampon étiquetée sort du domaine d'application de la présente spécification (c'est-à-dire, il est spécifique de l'ULP). Un message DDP peut commencer à un TO arbitraire au sein d'une mémoire tampon étiquetée.

De plus, une mémoire tampon étiquetée peut éventuellement être écrite plusieurs fois. Cela pourrait être fait pour une récupération d'erreur ou parce que une mémoire tampon est réutilisée après un mécanisme de synchronisation spécifique de l'ULP.

5.1.2 Modèle de mémoire tampon non étiquetée

Le modèle de la mémoire tampon non étiquetée est utilisé par la source de données pour transférer un message DDP au collecteur de données dans une mémoire tampon mise en file d'attente.

Le numéro de file d'attente DDP est utilisé par l'ULP pour séparer les messages d'ULP dans différentes files d'attente de mémoires tampon de réception. Par exemple, si deux files d'attente sont prises en charge, l'ULP pourrait utiliser une file d'attente pour envoyer les mémoires tampon qui lui sont passées par l'application au-dessus de l'ULP, et il pourrait utiliser l'autre file d'attente pour les mémoires tampon qui sont seulement consommées par des messages de contrôle spécifiques de l'ULP. Cela permet la séparation des messages de contrôle de l'ULP des charges utiles opaques d'ULP quand on utilise des mémoires tampon non étiquetées.

Le numéro de séquence de message DDP peut être utilisé par le collecteur de données pour identifier la mémoire tampon non étiquetée spécifique. Le protocole utilisé pour communiquer comment de nombreuses mémoires tampon ont été mises en file d'attente sort du domaine d'application de la présente spécification. De même, la mise en œuvre exacte de la file d'attente de mémoire tampon sort du domaine d'application de la présente spécification.

5.2 Segmentation et réassemblage d'un message DDP

À la source de données, la couche DDP DOIT segmenter les données contenues dans un message d'ULP en une série de segments DDP, où chaque segment DDP contient un en-tête DDP et une charge utile d'ULP, et DOIT n'être pas plus grand que la valeur de MULPDU annoncée par le LLP. La longueur du message d'ULP DOIT être inférieure à 2^32. À la source de données, la couche DDP DOIT envoyer toutes les données contenues dans le message d'ULP. Au collecteur de données, la couche DDP DOIT placer la charge utile d'ULP contenue dans tous les segments DDP valides entrants associés au message DDP dans la mémoire tampon d'ULP.

La segmentation de message DDP à la source de données est accomplie en identifiant de façon univoque un message DDP (qui correspond de façon bijective à un message d'ULP) et ensuite, pour chaque segment DDP associé d'un message DDP, en spécifiant un décalage d'octet pour la portion du message d'ULP contenue dans le segment DDP.

Pour un message DDP non étiqueté, la combinaison du QN et du MSN identifie de façon univoque un message DDP. Le décalage d'octet pour chaque segment DDP d'un message DDP non étiqueté est le champ MO. Pour chaque segment DDP d'un message DDP non étiqueté, le MO DOIT être réglé au décalage d'octet à partir du premier octet dans le message

d'ULP associé (qui est défini comme étant zéro) au premier octet dans la charge utile d'ULP contenue dans le segment DDP.

Par exemple, si le message d'ULP non étiqueté fait 2048 octets, et si la MULPDU fait 1500 octets, la source de données va générer deux segments DDP, un avec MO = 0, contenant 1482 octets de charge utile d'ULP, et le second avec MO = 1482, contenant 566 octets de charge utile d'ULP. Dans cet exemple, la quantité de charge utile d'ULP pour le premier segment DDP est calculée comme : 1482 = 1500 (MULPDU) - 18 (pour l'en-tête DDP).

Pour un message DDP étiqueté, la STag et le TO, combinés avec les caractéristiques de livraison en ordre du LLP, sont utilisés pour segmenter et réassembler le message d'ULP. Parce que le décalage initial d'octet (le champ TO) peut être différent de zéro, la récupération de la limite originale du message d'ULP ne peut pas être faite dans le cas général sans un message d'ULP supplémentaire.

Note de mise en œuvre : une mise en œuvre, valide pour certains ULP comme RDMAP, ne va pas prendre directement en charge la récupération de la limite de message d'ULP pour un message DDP étiqueté. Par exemple, l'ULP peut souhaiter que l'homologue local utilise de petites mémoires tampon à la source de données même quand l'ULP au collecteur de données a annoncé une seule grande mémoire tampon étiquetée pour ce transfert de données. Dans ce cas, l'ULP peut choisir d'utiliser la même STag pour plusieurs messages d'ULP consécutifs. Donc, un TO initial non zéro et la réutilisation de la STag permettent effectivement à l'ULP de mettre en œuvre la segmentation et le réassemblage du fait de contraintes spécifiques de l'ULP. Voir dans la [RFC5040] les détails de la façon de faire cela. Une mise en œuvre différente d'ULP pourrait utiliser un message DDP non étiqueté (envoyé après le message DDP étiqueté) qui détaille le TO initial pour la STag qui a été utilisée dans le message DDP étiqueté. Et finalement, une autre mise en œuvre d'ULP pourrait choisir de toujours utiliser un TO initial de zéro afin qu'aucun message supplémentaire ne soit nécessaire pour porter le TO initial utilisé dans un message DDP étiqueté.

Sans considérer si l'ULP choisit de récupérer la limite du message d'ULP original au collecteur de données pour un message DDP étiqueté, DDP prend en charge la segmentation et le réassemblage du message DDP étiqueté. La STag est utilisée pour identifier la mémoire tampon d'ULP au collecteur de données, et le TO est utilisé pour identifier le décalage d'octet au sein de la mémoire tampon d'ULP référencée par la STag. L'ULP à la source de données DOIT spécifier la STag et le TO initial quand le message d'ULP est passé à DDP.

Pour chaque segment DDP d'un message DDP étiqueté, le TO DOIT être réglé au décalage d'octet à partir du premier octet dans le message d'ULP associé au premier octet dans la charge utile d'ULP contenue dans le segment DDP, plus le TO alloué au premier octet dans le message d'ULP associé.

Par exemple, si le message d'ULP étiqueté fait 2048 octets avec un TO initial de 16384, et si la MULPDU est de 1500 octets, la source de données va générer deux segments DDP : un avec TO = 16384, contenant le premier 1486 octets de charge utile d'ULP, et un second avec TO = 17870, contenant 562 octets de charge utile d'ULP. Dans cet exemple, la quantité de charge utile d'ULP pour le premier segment DDP est calculée par : 1486 = 1500 (MULPDU) - 14 (pour l'en-tête DDP).

Un message DDP de longueur zéro est permis et DOIT consommer exactement un segment DDP. Seuls les champs Contrôle DDP et RsvdULP DOIVENT être valides pour un segment DDP étiqueté de longueur zéro. Les champs STag et TO NE DOIVENT PAS être vérifiés pour un message DDP étiqueté de longueur zéro.

Pour les messages DDP étiquetés ou non étiquetés, le collecteur de données n'est pas obligé de vérifier que le message d'ULP entier a été reçu.

5.3 Ordre des messages DDP

Les messages passés par DDP DOIVENT se conformer aux règles d'ordre définies dans ce paragraphe.

À la source de données, DDP :
* DOIT transmettre les messages DDP dans l'ordre de leur soumission par la couche DDP,
* DEVRAIT transmettre les segments DDP au sein d'un message DDP en ordre de MO croissant pour les messages DDP non étiquetés, et en ordre de TO croissant pour les messages DDP étiquetés.

Au collecteur de données, DDP (note : les règles suivantes sont motivées par les mises en œuvre de LLP qui séparent le placement et la livraison.) :
* PEUT effectuer le placement de segments DDP déclassés,

* PEUT effectuer plus d'une fois le placement d'un segment DDP,
* DOIT livrer au plus une fois un message DDP à l'ULP,
* DOIT livrer les messages DDP à l'ULP dans l'ordre de leur envoi par la source de données.

5.4 Achèvement et livraison du message DDP

À la source de données, le transfert de message DDP est considéré achevé quand le LLP de transport fiable dans l'ordre a indiqué que le transfert va se produire de façon fiable. Noter que cela n'empêche en aucune façon le LLP de mettre les données en mémoire tampon à la source de données ou au collecteur de données. Donc, à la source de données, l'achèvement d'un message DDP ne signifie pas nécessairement que le collecteur de données a reçu le message.

Au collecteur de données, DDP DOIT livrer un message DDP si et seulement si tout ce qui suit est vrai :
* le dernier segment DDP du message DDP a son fanion Last établi,
* tous les segments DDP du message DDP ont été placés,
* tous les messages DDP précédents ont été placés, et
* chaque message DDP précédent a été livré à l'ULP.

Au collecteur de données, DDP DOIT fournir la longueur du message d'ULP à l'ULP quand un message DDP non étiqueté est livré. La longueur de message d'ULP peut être calculée en ajoutant le MO et la longueur de charge utile d'ULP dans le dernier segment DDP (avec le fanion Last établi) d'un message DDP non étiqueté.

Au collecteur de données, DDP DOIT fournir le champ RsvdULP du message DDP à l'ULP quand le message DDP est livré.

6. Établissement et suppression de flux DDP

Cette Section décrit les questions indépendantes du LLP relatives à l'établissement et à la suppression du flux DDP.

6.1 Établissement de flux DDP

Il est prévu que l'ULP utilise un mécanisme qui sort du domaine d'application de la présente spécification pour établir une connexion de LLP, et que la connexion de LLP va prendre en charge un ou plusieurs flux de LLP (par exemple, MPA/TCP ou SCTP). Après que le LLP a établi le flux de LLP, il va permettre un flux DDP sur un flux de LLP spécifique à un point approprié.

Il est exigé que l'ULP active au même moment les deux points d'extrémité d'un flux de LLP pour le transfert de données DDP, dans les deux directions ; ceci est nécessaire afin que le collecteur de données puisse correctement reconnaître les segments DDP.

6.2 Suppression de flux DDP

DDP NE DOIT PAS initier indépendamment une suppression de flux. Soit DDP répond à une suppression de flux par le LLP, soit il traite une demande de l'ULP de supprimer un flux. La suppression d'un flux DDP désactive les capacités DDP sur les deux points d'extrémité. Pour les LLP en mode connexion, la suppression de flux DDP PEUT résulter en la suppression de la connexion de LLP sous-jacente.

6.2.1 Suppression DDP en douceur

Il appartient à l'ULP de s'assurer que la suppression DDP se produit sur les deux points d'extrémité du flux DDP au même moment ; ceci est nécessaire afin que le collecteur de données cesse d'essayer d'interpréter les segments DDP.

Si l'homologue ULP local indique une suppression en douceur, la couche DDP chez l'homologue local DEVRAIT s'assurer que toutes les données d'ULP vont être transférées avant que le flux de LLP et la connexion sous-jacents soient supprimés, et toutes les demandes de transfert de données suivantes de l'homologue ULP local DOIVENT retourner une erreur.

Si la couche DDP chez l'homologue local reçoit une demande de suppression en douceur du LLP, toutes les données reçues après la demande sont considérées comme une erreur et DOIVENT causer la suppression interruptive du flux DDP.

Si le LLP de l'homologue local prend en charge un flux de LLP à demi fermé, à réception d'une demande de suppression en douceur de LLP du flux DDP, DDP DEVRAIT indiquer l'état demi-clos à l'ULP, et continuer de traiter normalement les demandes de transfert de données sortantes. À la suite de cet événement, quand l'homologue ULP local demande une suppression en douceur, DDP DOIT indiquer au LLP qu'il DEVRAIT effectuer une fermeture en douceur de l'autre moitié du flux de LLP.

Si le LLP de l'homologue local prend en charge un flux de LLP demi-clos, à réception d'une demande de fermeture en douceur d'ULP demi-clos du flux DDP, DDP DEVRAIT garder la réception de données activée sur l'autre moitié du flux de LLP.

6.2.2 Suppression DDP interruptive

Comme mentionné précédemment, DDP ne termine pas un flux DDP de façon indépendante. Donc, toutes les erreurs fatales suivantes sur un flux DDP DOIVENT causer l'indication par DDP à l'ULP qu'une erreur fatale s'est produite :
* la connexion LLP ou le flux de LLP sous jacent est perdu ;
* le LLP sous-jacent rapporte une erreur fatale ;
* l'en-tête DDP a un ou plusieurs champs invalides.

Si le LLP indique à l'ULP qu'une erreur fatale s'est produite, la couche DDP DEVRAIT rapporter l'erreur à l'ULP (voir au paragraphe 7.2, "Numéros d'erreur DDP") et achever toutes les demandes d'ULP en instance avec une erreur. Si le flux de LLP sous-jacent est encore intact, DDP DEVRAIT continuer de permettre à l'ULP de transférer des messages DDP supplémentaires sur la demi connexion sortante après que l'erreur fatale a été indiquée à l'ULP. Cela permet à l'ULP de transférer un syndrome d'erreur à l'homologue distant. Après avoir indiqué à l'ULP qu'une erreur fatale s'est produite, le flux DDP NE DOIT PAS être terminé tant que l'homologue ULP local n'a pas indiqué à la couche DDP que le flux DDP devrait être supprimé de façon interruptive.

7. Sémantique des erreurs

Toutes les erreurs de LLP rapportées à DDP DEVRAIENT être passées à l'ULP.

7.1 Erreurs détectées au collecteur de données

Pour les segments DDP non étiquetés de longueur non zéro, le segment DDP DOIT être validé avant le placement en vérifiant que :

1. Le QN est valide pour ce flux.

2. Le QN et le MSN ont une mémoire tampon associée qui permet le placement de la charge utile.

Note de mise en œuvre : les mises en œuvre DDP DEVRAIENT considérer l'absence d'une mémoire tampon associée comme une faute du système. Les mises en œuvre de DDP PEUVENT essayer de récupérer de la faute du système en utilisant des moyens du LLP d'une façon transparente pour l'ULP. Les mises en œuvre de DDP NE DEVRAIENT PAS permettre que des fautes de système se produisent de façon répétée ou fréquente. Si il n'y a pas de mémoire tampon associée, les mises en œuvre de DDP PEUVENT choisir de désactiver le flux pour la réception et rapporter une erreur à l'ULP au collecteur de données.

3. Le MO tombe dans la gamme légale des décalages associés à la mémoire tampon non étiquetée.

4. La somme de la longueur de la charge utile du segment DDP et du MO tombe dans la gamme légale des décalages associés à la mémoire tampon non étiquetée.

5. Le numéro de séquence de message tombe dans la gamme légale des numéros de séquence de message, pour la file d'attente définie par le QN. La gamme légale est définie comme étant entre la valeur de MSN allouée à la première mémoire tampon disponible pour un QN spécifique et la valeur de MSN allouée à la dernière mémoire tampon disponible pour un QN spécifique.

Note de mise en œuvre : pour un numéro de file d'attente normal, la limite inférieure de numéro de séquence de message

est définie par tous les messages DDP qui ont déjà été achevés. La limite supérieure est définie par le nombre de mémoires tampon de message qui sont actuellement disponibles pour cette file d'attente. Les deux nombres changent de façon dynamique lorsque de nouveaux messages DDP sont reçus et achevés, et que de nouvelles mémoires tampon sont ajoutées. Il appartient à l'ULP de s'assurer que suffisamment de mémoires tampon sont disponibles pour traiter les segments DDP entrants.

Pour les segments DDP étiquetés de longueur non zéro, le segment DOIT être validé avant le placement en vérifiant que :
1. la STag est valide pour ce flux ;
2. la STag a une mémoire tampon associée qui permet le placement de la charge utile ;
3. le TO tombe dans la gamme légale des décalages enregistrée pour la STag ;
4. la somme de la longueur de la charge utile du segment DDP et du TO tombe dans la gamme légale des décalages enregistrée pour la STag ;
5. une somme de 64 bits non signée de la longueur de la charge utile du segment DDP et du TO ne revient pas à zéro.

Si la couche DDP détecte une des erreurs de réception mentionnées dans cette section, elle DOIT cesser de placer le reste du segment DDP et rapporter la ou les erreurs à l'ULP. La couche DDP DEVRAIT inclure dans le rapport d'erreur l'en-tête DDP, le type d'erreur, et la longueur du segment DDP, si disponible. DDP DOIT éliminer en silence tout segment DDP entrant suivant. Comme chacune de ces erreurs représente une défaillance de l'ULP ou protocole envoyeur, DDP DEVRAIT permettre à l'ULP d'envoyer un message DDP supplémentaire avant de terminer le flux DDP.

7.2 Numéros d'erreur DDP

Les numéros d'erreur suivants DOIVENT être utilisés pour les rapports d'erreurs à l'ULP. Ils correspondent aux vérifications énumérées au paragraphe 7.1. Chaque erreur est subdivisée en un type d'erreur de 4 bits et un code d'erreur de 8 bits.

Type d'erreur	Code d'erreur	Description
0x0	0x00	Locale catastrophique
0x1		Erreur de mémoire tampon étiquetée
	0x00	STag invalide
	0x01	violation de base ou de limites
	0x02	STag non associée au flux DDP
	0x03	retour à zéro du TO
	0x04	version DDP invalide
0x2		Erreur de mémoire tampon non étiquetée
	0x01	QN invalide
	0x02	MSN invalide - pas de mémoire tampon disponible
	0x03	MSN invalide - gamme de MSN invalide
	0x04	MO invalide
	0x05	message DDP trop long pour la mémoire tampon disponible
	0x06	version DDP invalide
0x3	Rsvd	réservé pour l'utilisation du LLP

8. Considérations sur la sécurité

Cette Section discute des considérations spécifiques du protocole et des implications de l'utilisation de DDP avec les mécanismes de sécurité existants. Les exigences de sécurité pour la mise en œuvre de DDP sont fournies à la fin de la section. Une analyse plus détaillée des problèmes de sécurité autour de la mise en œuvre et l'utilisation de DDP se trouve dans la [RFC5042].

Les exigences de IPsec pour RDDP se fondent sur la version de IPsec spécifiée dans la [RFC2401] et les RFC qui s'y relatent, comme 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]. Une des applications précoces importantes des protocoles RDDP est leur utilisation avec iSCSI [RFC5046] ; les exigences de IPsec de RDDP suivent celles de IPsec afin de faciliter cet usage en permettant qu'un profil commun de IPsec soit utilisé avec iSCSI et les protocoles RDDP. À l'avenir, la RFC 3723 pourra être mise à jour avec le version plus récente de IPsec ; les exigences de sécurité de IPsec d'une telle mise à jour devraient s'appliquer uniformement à iSCSI et aux protocoles RDDP.

8.1 Considérations sur la sécurité spécifiques du protocole

Les vulnérabilités de DDP à l'interférence active de tiers ne sont pas plus grandes que celles de tout autre protocole fonctionnant par dessus des protocoles de transport comme TCP et SCTP sur IP. Un tiers, en injectant des paquets usurpés dans le réseau qui sont livrés à un collecteur de données DDP, pourrait lancer diverses attaques qui exploitent un comportement spécifique de DDP. Comme DDP expose directement ou indirectement les adressses de mémoire sur le réseau, les informations de placement portées dans chaque segment DDP doivent être validées, incluant la vérification de base et de limite à la granularité de niveau octet et de la validité de STag avant tout placement de données. Par exemple, un adversaire tiers pourrait injecter des paquets aléatoires qui paraissent être des segments DDP valides et corrompent la mémoire sur un collecteur de données DDP. Comme DDP est indépendant du protocole de transport IP, les mécanismes de sécurité de la communication comme IPsec [RFC2401] peuvent être utilisés pour empêcher de telles attaques.

8.2 Association d'un flux STag et d'un flux DDP

Il y a plusieurs mécanismes pour associer une STag et un flux DDP. Deux mécanismes requis pour cette association sont une association de domaine de protection (PD) et une association de flux DDP.

Avec l'association de domaine de protection (PD) un unique identifiant de domaine de protection (PD ID) est créé et utilisé en local pour associer une STag à un ensemble de flux DDP. Avec ce mécanisme, l'utilisation de la STag est seulement permise sur les flux DDP qui ont le même PD ID que la STag. Pour un segment DDP entrant d'un message DDP étiqueté sur un flux DDP, si le PD ID du flux DDP n'est pas le même que le PD ID de la STag ciblée par le message DDP étiqueté, alors le segment DDP n'est pas placé, et la couche DDP DOIT nettoyer une erreur en local à l'ULP. Noter que le PD ID est défini localement et ne peut pas être directement manipulé par l'homologue distant.

Avec l'association de flux DDP, un flux DDP est identifié localement par un unique identifiant de flux DDP (ID). Une STag est associée au flux DDP en utilisant un identifiant de flux DDP. Dans ce cas, pour un segment DDP entrant d'un message DDP étiqueté sur un flux DDP, si l'identifiant de flux DDP du flux DDP n'est pas le même que celui de la STag ciblée par le message DDP étiqueté, alors le segment DDP n'est pas placé et la couche DDP DOIT nettoyer une erreur locale à l'ULP. Noter que l'identifiant de flux DDP est défini localement et ne peut pas être directement manipulé par l'homologue distant.

Un ULP DEVRAIT associer une STag à au moins un flux DDP. DDP DOIT prendre en charge les mécanismes d'association de domaine de protection et d'association de flux DDP pour associer une STag et un flux DDP.

8.3 Exigences pour la sécurité

La [RFC5042] définit le modèle de sécurité et les hypothèses générales pour RDMAP/DDP. On donne ici les exigences de sécurité pour la mise en œuvre de DDP. Pour plus de détails sur le type d'attaques, d'attaquants, les modèles de confiance, et le partage de ressources pour la mise en œuvre de DDP, le lecteur se reportera à la [RFC5042].

DDP a plusieurs mécanismes qui traitent un certain nombre d'attaques. Ces attaques incluent, mais ne se limitent pas à :
1. la connexion à un point d'extrémité non autorisé ou non authentifié ;
2. la capture d'un flux DDP ;
3. la tentative de lire ou écrire à partir de régions de mémoire non autorisées ;
4. l'injection de messages RDMA dans un flux sur un système d'exploitation multi-utilisateurs par une autre application.

DDP s'appuie sur le LLP pour établir le flux de LLP sur lequel les messages DDP vont être portés. DDP lui-même ne fait rien pour authentifier la validité du flux de LLP de l'un ou l'autre des points d'extrémité. Il est de la responsabilité de l'ULP de valider le flux de LLP. Ceci est très souhaitable à cause de la nature de DDP.

La capture d'un flux DDP va exiger que le flux de LLP sous-jacent soit capturé. Cela va exiger la connaissance des mémoires tampon annoncées afin de placer directement les données dans une mémoire tampon d'utilisateur. Donc, ceci est contraint par les mêmes techniques que mentionné pour se garder contre les tentatives de lire ou écrire à partir de régions de mémoire non autorisées.

DDP n'exige pas d'un nœud qu'il ouvre ses mémoires tampon à des attaques arbitraires sur le flux DDP. Il ne peut accéder à la mémoire d'ULP que dans la mesure où l'ULP l'a activé et autorisé à le faire. Le modèle de contrôle d'accès de STag est défini dans la [RFC5042]. Les opérations spécifiques de sécurité incluent que :
1. Les STag sont seulement valides sur la gamme d'octets exacte établie par l'ULP. DDP DOIT fournir un mécanisme pour que l'ULP établisse et révoque la gamme de TO associée à la mémoire tampon d'ULP référencée par la STag.

2. Les STag sont seulement valides pour la durée établie par l'ULP. L'ULP peut les révoquer à tout moment, en accord avec ses propres exigences d'ULP. DDP DOIT fournir un mécanisme pour que l'ULP établisse et révoque la validité de la STag.
3. DDP DOIT fournir un mécanisme pour que l'ULP communique l'association entre une STag et un flux DDP spécifique.
4. Un ULP peut seulement exposer la mémoire à l'accès à distance dans la mesure où il a déjà lui-même accès à cette mémoire.
5. Si une STag n'est pas valide sur un flux DDP, DDP DOIT passer la tentative d'accès invalide à l'ULP. L'ULP peut fournir un mécanisme pour terminer le flux DDP.

De plus, DDP fournit un mécanisme qui place directement les charges utiles entrantes dans les mémoires tampon d'ULP en mode utilisateur. Cela évite les risquess des solutions précédentes qui reposent sur l'exposition des mémoires tampon système pour les charges utiles entrantes.

Pour la mise en œuvre de DDP, deux composants DOIVENT être fournis : un contrôleur d'interface réseau à capacité RDMA (RNIC, *RDMA-enabled NIC*) et un gestionnaire de ressource privilégié (PRM, *Priviledged Resource Manager*).

8.3.1 Exigences pour RNIC

Le RNIC DOIT mettre en œuvre le protocole réseau DDP et la sémantique de sécurité décrits ci-dessous.

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

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

3. Un RNIC DOIT s'assure qu'un homologue distant n'est pas capable d'accéder à la 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 NE DOIT PAS permettre qu'un logiciel soit chargé sur le RNIC directement d'un homologue local ou distant qui n'est pas de confiance, sauf si l'homologue est correctement authentifié (par un mécanisme qui sort du domaine d'application de la présente spécification ; le mécanisme entraîne probablement d'authentifier que l'ULP distant a le droit d'effectuer la mise à jour) et si la mise à jour est faite via un protocole sûr, comme IPsec.

8.3.2 Exigences pour le gestionnaire de ressources privilégiées

Le PRM DOIT mettre en œuvre la sémantique de sécurité décrite ci-dessous.

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 ressource privilégié comme mandataire.

2. Toutes les demandes d'allocation de ressource de l'ULP pour des ressources éparses 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 partielle mutuelle sauf si il y a un mécanisme pour s'assurer que les ULP partagent bien une confiance partielle mutuelle.

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

l'ULP a les droits d'accès pour être associé à une mémoire tampon de données spécifique.

5. Le gestionnaire de ressource privilégié DEVRAIT empêcher un homologue local d'allouer plus que sa part équitable des ressources. Si un RNIC donne la capacité de partager les mémoires tampon de réception sur plusieurs flux DDP, 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 juste part des ressources afin que l'homologue local puisse appliquer des contre mesures pour détecter et prévenir l'attaque.

8.4 Services de sécurité pour DDP

DDP utilise des services réseau fondés sur IP ; donc, tous les segments DDP échangés sont vulnérables à des attaques d'usurpation d'identité, d'altération et de divulgation d'informations. Si un flux DDP peut être soumis à des attaques d'usurpation d'identité, ou des attaques de capture de flux, il est fortement RECOMMANDÉ que le flux DDP soit authentifié, protégé en intégrité, et protégé contre les attaques en répétition. Il PEUT utiliser la protection de la confidentialité pour se protéger contre l'espionnage.

8.4.1 Services de sécurité disponibles

IPsec peut être utilisé pour protéger contre les attaques d'injection de paquets mentionnées précédemment. Parce que IPsec est conçu pour sécuriser des flux arbitraires de paquets IP, incluant des flux où des paquets sont perdus, DDP peut fonctionner par dessus IPsec sans aucun changement.

La sécurité de DDP peut aussi profiter des services de sécurité de SSL ou TLS fournis pour les ULP fondés sur TCP ou SCTP [RFC4346] ainsi que des services de sécurité provenant de DTLS [RFC4347] fournis en dessous du protocole de transport. Voir dans la [RFC5042] la discussion de ces approches et la raison du choix des services de sécurité IPsec pour les protocoles RDDP.

8.4.2. Exigences pour les services IPsec de DDP

Les paquets IPsec sont traités (par exemple, vérification d'intégrité, et éventuellement déchiffrés) dans l'ordre de leur réception, et un collecteur de données DDP va traiter le contenu des segments DDP déchiffrés dans ces paquets de la même manière que le contenu des segments DDP dans des paquets IP non sécurisés.

Le groupe de travail "IP Storage" a défini les exigences normatives de IPsec pour la mémorisation IP [RFC3723]. Des portions de cette spécification sont applicables à DDP. En particulier, une mise en œuvre conforme des services IPsec DOIT satisfaire aux exigences mentionnées au paragraphe 2.3 de la [RFC3723]. Sans reproduire la discussion détaillée 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 d'origine des données par paquet, l'intégrité, et la protection 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 IKE "delete" comme une raison pour supprimer le flux DDP. Comme le matériel d'accélération IPsec peut seulement être capable de traiter un nombre limité d'associations de sécurité (SA) IPsec actives, des SA inactives peuvent être supprimées dynamiquement et une nouvelle SA être montée 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 d'homologue utilisant les méthodes de chiffrement à 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. Le mode principal IKE avec authentification par clé pré-partagée NE DEVRAIT PAS être utilisé quand l'un ou l'autre des

homologues utilise une adresse IP allouée de façon dynamique.

7. L'accès à des informations secrètes mémorisées en local (clé pré-partagée ou privée pour une signature numérique) doit être convenablement encadrée, car la compromission des informations secrètes annule les propriétés de sécurité des protocoles IKE/IPsec.

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

De plus, la mise en œuvre et le déploiement des services IPsec pour DDP devrait 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 remarque suivante est un simple commentaire.

Si DDP était activé à priori pour un ULP en le connectant à un accès bien connu, cet accès bien connu serait enregistré pour DDP par l'IANA. L'enregistrement de l'accès bien connu serait 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)

[RFC5040] R. Recio et autres, "Spécification d'un protocole d'accès direct à une mémoire distante", 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*)*

[RFC5043] C. Bestler et R. Stewart, éd., "Adaptation du placement direct des données (DDP) au protocole de transmission de contrôle de flux (SCTP)", octobre 2007. *(P.S. ; MàJ par* RFC6581, 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 *)*

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)

[RFC4347] E. Rescorla, N. Modadugu, "Sécurité de la couche de transport de datagrammes", avril 2006. *(P.S.)*

[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)

Appendice A. Dimensionnement de la fenêtre de réception

Cet Appendice donne des lignes directrices pour les mises en œuvre de LLP.

Les LLP fiables, suivis, incluent un mécanisme pour annoncer la quantité d'espace de mémoire tampon de réception qu'un envoyeur peut consommer. Ceci est généralement appelé une "fenêtre de réception".

DDP permet que les données soient transférées directement aux mémoires tampon prédéfinies chez le collecteur de données. En conséquence, la taille de la fenêtre de réception de LLP ne doit pas être affectée par la réception d'un segment DDP, si ce segment est placé avant que des segments supplémentaires arrivent.

La mise en œuvre de LLP DEVRAIT maintenir une fenêtre de réception annoncée assez grande pour permettre qu'un nombre raisonnable de segments soient en instance à un instant donné. La quantité à annoncer dépend du débit de données désiré, et du délai d'aller-retour attendu ou réel entre les points d'extrémité.

La quantité de mémoires tampon réelles maintenues pour "sauvegarder" la fenêtre de réception relève de la mise en œuvre. Cette quantité va dépendre du taux auquel les segments DDP peuvent être retirés ; il peut y avoir des cas où le traitement de segment ne peut pas soutenir le taux de paquets entrants. Si cela se produit, un façon raisonnable de ralentir le taux de paquets entrants est de diminuer la fenêtre de réception.

Noter que le LLP devrait veiller à se conformer aux RFC applicables ; par exemple, pour TCP, il est fortement déconseillé aux receveurs de "réduire" la fenêtre de réception (réduire le côté droit de la fenêtre après qu'elle a été annoncée).

Appendice B. Contributeurs

Tous nos remerciements aux personnes suivantes pour leurs contributions.
John Carrier, Cray Inc. ; mél : carrier@cray.com
Hari Ghadia, Gen10 Technology, Inc. ; mél : hghadia@gen10technology.com
Caitlin Bestler, Broadcom Corporation ; mél : caitlinb@Broadcom.com
Uri Elzur, Broadcom Corporation ; mél : uri@broadcom.com
Mike Penna, Broadcom Corporation ; mél : MPenna@Broadcom.com
Patricia Thaler, Broadcom Corporation ; mél : pthaler@broadcom.com
Ted Compton, EMC Corporation ; mél : compton_ted@emc.com
Jim Wendt, Hewlett-Packard Company ; mél : jim_wendt@hp.com
Mike Krause, Hewlett-Packard Company ; mél : krause@cup.hp.com
Dave Minturn, Intel Corporation ; mél : dave.b.minturn@intel.com

Howard C. Herbert, Intel Corporation ; mél : howard.c.herbert@intel.com
Tom Talpey, Network Appliance ; mél : thomas.talpey@netapp.com
Dwight Barron, Hewlett-Packard Company ; mél : Dwight.Barron@Hp.com
Dave Garcia, StanfordAlumni ; mél : Dave.Garcia@StanfordAlumni.org
Jeff Hilland, Hewlett-Packard Company ; mél : jeff.hilland@hp.com
Barry Reinhold, Lamprey Networks ; mél : bbr@LampreyNetworks.com

Adresse des auteurs

Hemal Shah
Broadcom Corporation
5300 California Avenue
Irvine, CA 92617 USA
tél. : +1 (949) 926-6941
mél : hemal@broadcom.com

James Pinkerton
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052 USA
tél. : +1 (425) 705-5442
mél : jpink@microsoft.com

Renato Recio
IBM Corporation
11501 Burnett Road
Austin, TX 78758 USA
tél. : +1 (512) 838-1365
mél : recio@us.ibm.com

Paul R. Culley
Hewlett-Packard Company
20555 SH 249
Houston, TX 77070-2698
tél. : +1 (281) 514-5543
mél : paul.culley@hp.com

Déclaration complète de droits de reproduction