Équipe d'ingénierie de l'Internet ((IETF)
Request for Comments : 7145
RFC rendue obsolète : 5046
Catégorie : En cours de normalisation
ISSN: 2070-1721

M. Ko
A. Nezhinsky, Mellanox

avril 2014
Traduction Claude Brière de L'Isle

Extensions d'interface Internet de système de petit ordinateur (iSCSI) pour la spécification de l'accès direct en mémoire distante (RDMA)

Résumé

Les extensions d'interface Internet de système de petit ordinateur (iSCSI, *Internet Small Computer System Interface*) pour la spécification de l'accès direct en mémoire distante (RDMA, *Remote Direct Memory Access*) fournissent la capacité de transfert de données RDMA à iSCSI en mettant en couches iSCSI par dessus un protocole à capacité RDMA. Un protocole à capacité RDMA fournit des services de lecture et écriture RDMA qui permettent que des données soient transférées directement dans les mémoires tampon d'entrée/sortie iSCSI sans copies de données intermédiaires. Le présent document décrit les extensions au protocole iSCSI pour la prise en charge des services RDMA comme ils sont fournis par un protocole à capacité RDMA.

Le présent document rend obsolète la RFC 5046.

Statut de ce mémoire

Ceci est un document de l'Internet en cours de normalisation.

Le présent document a été produit par l'équipe d'ingénierie de l'Internet (IETF). Il représente le consensus de la communauté de l'IETF. Il a subi une révision publique et sa publication a été approuvée par le groupe de pilotage de l'ingénierie de l'Internet (IESG). Plus d'informations sur les normes de l'Internet sont disponibles à la Section 2 de la [RFC5741].

Les informations sur le statut actuel du présent document, tout errata, et comment fournir des réactions sur lui peuvent être obtenues à http://www.rfc-editor.org/info/rfc7145

Table des matières

1. Introduction

1.1 Motivation

Le protocole iSCSI ([RFC7143]) est une transposition du modèle d'architecture SCSI (voir [SAM5] et la [RFC7144]) sur le protocole TCP. Les commandes SCSI sont portées par les demandes iSCSI, et les réponses et états iSCSI sont portés par les réponses iSCSI. D'autres échanges de protocole iSCSI et de données SCSI sont aussi transportés dans des PDU iSCSI.

Les segments TCP décalés dans le modèle iSCSI traditionnel doivent être mémorisés et réassemblés avant que la couche de protocole iSCSI au sein d'un nœud d'extrémité puisse placer les données dans les mémoires tampon iSCSI. Ce réassemblage est exigé parce que tous les segments TCP ne vont probablement pas contenir un en-tête iSCSI pour permettre leur placement et TCP lui-même n'a pas de mécanisme incorporé pour signaler les limites de message de protocole de niveau supérieur (ULP, *Upper Level Protocol*) pour aider au placement des segments déclassés. Ce réassemblage TCP à de hauts débits du réseau est assez contre productif pour les raisons suivantes : gaspillage de mémoire dans la copie des données, besoin de mémoire de réassemblage, gaspillage de cycles de CPU dans la copie des données, et latence générale du différé du point de vue de l'application.

Le terme générique de protocole à capacité RDMA (RcaP, *RDMA-Capable Protocol*) est utilisé pour se référer aux piles de protocoles qui fournissent la fonction d'accès directe à la mémoire distante (RDMA, *Remote Direct Memory Access*) comme iWARP et InfiniBand.

Avec la disponibilité de conrôleurs à capacité RDMA au sein d'un système hôte, il est approprié que iSCSI soit capable d'exploiter la fonction de placement direct des données du contrôleur à capacité RDMA comme les autres applications.

Les extensions iSCSI pour RDMA (iSER) sont conçues précisément pour tirer parti des technologies génériques RDMA – le but d'iSER est de permettre à iSCSI d'employer le placement direct de données et les capacités RDMA en utilisant un contrôleur générique à capacité RDMA. En résumé, la pile de protocoles iSCSI/iSER est conçue pour permettre de s'adapter aux grandes vitesses en s'appuyant sur un processus générique de placement de données et les technologies et produits RDMA qui permettent le placement direct des données de données aussi bien en ordre que déclassées.

Le présent document décrit iSER comme une extension de protocole à iSCSI, à la fois pour les besoins de la description et aussi parce que c'est vrai au sens strict du protocole. Cependant, on notera que iSER est en réalité une extension de la connexité du protocole iSCSI définie dans la [RFC7143], et le nom "iSER" reflète cette réalité.

Quand le protocole iSCSI défini dans la [RFC7143] (c'est-à-dire, sans les améliorations iSER) est visé dans la suite du document, le terme "iSCSI traditionnel" est utilisé pour rendre l'intention claire.

Le présent document rend obsolète la RFC 5046. Voir à l'Appendice A la liste des changements par rapport à la RFC 5046.

1.2 Mise en couches iSCSI/iSER

Les extensions iSCSI pour RDMA (iSER) sont mises en couches entre la couche iSCSI et la couche RCaP.

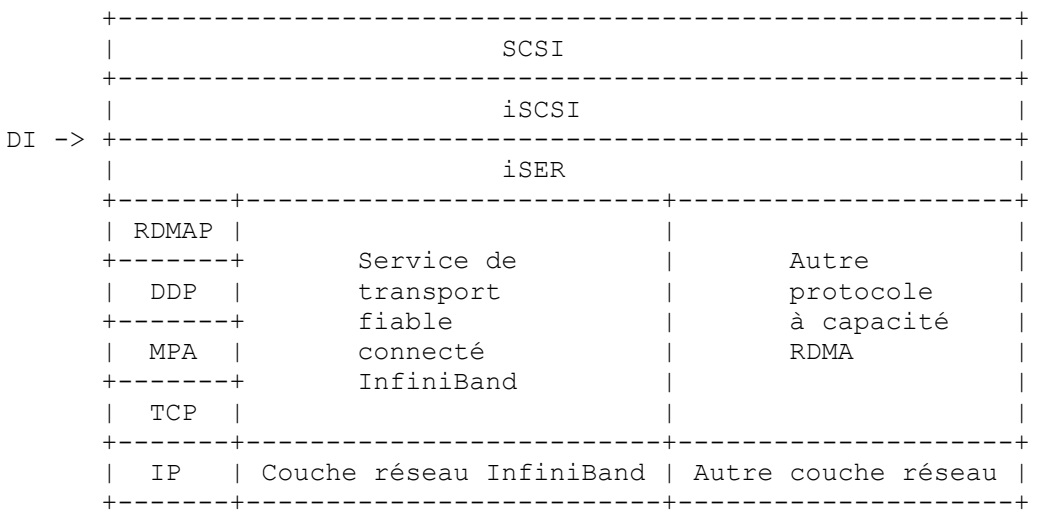


Figure 1 : Exemple de mise en couche iSCSI/iSER en phase de pleines caractéristiques

La Figure 1 montre un exemple des relations entre SCSI, iSCSI, iSER, et les différentes couches RCaP. Pour TCP, le RCaP est iWARP. Pour InfiniBand, le RCaP est le service de transport fiable connecté. Noter que la couche iSCSI telle que décrite ici prend en charge les extensions RDMA comme elles sont utilisées dans iSER.

1.3 Buts architecturaux

Ce paragraphe résume les buts architecturaux qui ont guidé la conception d'iSER.

1. Fournir un modèle de transfert de données RDMA pour iSCSI qui permette le placement direct de données SCSI en ordre ou déclassées dans les mémoires tampon SCSI préallouées tout en maintenant la livraison des données dans l'ordre.

2. Ne pas exiger de changement majeur du modèle d'architecture SCSI [SAM5] et des normes du jeu de commandes SCSI.

3. Utiliser l'infrastructure iSCSI existante (parfois appelée "l'écosystème iSCSI") incluant sans s'y limiter la MIB, l'amorçage, la négociation, la désignation et la découverte, et la sécurité.

4. Permettre à une session de fonctionner sous le mode de transfert de données iSCSI traditionnel si iSER n'est pas pris en charge par l'initiateur ou la cible. (Cela n'exige pas l'interopérabilité de phase de pleines caractéristiques iSCSI entre un nœud d'extrémité qui fonctionne en mode iSCSI traditionnel et un nœud d'extrémité qui fonctionne en mode assisté par iSER.)

5. Permettre aux mises en œuvre d'initiateur et de cible d'utiliser des contrôleurs génériques à capacité RDMA comme des RNIC ou de mettre en œuvre iSCSI et iSER dans le logiciel. (Cela n'exige pas une assistance spécifique de iSCSI ou iSER dans la mise en œuvre de RCaP ou du contrôleur à capacité RDMA.)

6. Mettre en œuvre un protocole léger de déplacement de données pour iSCSI avec une maintenance d'état minimale.

1.4 Vue d'ensemble du protocole

En cohérence avec les buts architecturaux du paragraphe 1.3, le protocole iSER n'exige pas de changement de l'écosystème iSCSI ou des spécifications SCSI qui s'y rapportent. Le protocole iSER définit la transposition des PDU iSCSI en messages RCaP d'une façon telle qu'il soit possible de réaliser des mises en œuvre iSCSI/iSER qui se fondent sur les contrôleurs génériques à capacité RDMA. La couche de protocole iSER exige une maintenance d'état minimale pour assister une connexion durant la phase iSCSI de pleines caractéristiques, sans tenir compte de la notion de session iSCSI. Les aspects cruciaux du protocole iSER peuvent être résumés comme suit :

1. Le mode d'assistance iSER est négocié durant l'établissement iSCSI dans la connexion de tête pour chaque session, et une session iSCSI entière ne peut fonctionner que dans un mode (c'est-à-dire, une connexion dans une session ne peut pas fonctionner en mode à assistance iSER si une connexion différente de la même session est déjà en phase de pleines caractéristiques dans le mode iSCSI traditionnel).

2. Une fois dans le mode à assistance iSER, toutes les interactions iSCSI sur cette connexion utilisent les messages RCaP.

3. Un message Send est utilisé pour porter une PDU iSCSI Type de contrôle précédée d'un en-tête iSER. Voir au paragraphe 7.2 les détails des PDU Type de contrôle iSCSI.

4. Les messages RDMA Écriture, Demande de lecture RDMA, et Réponse de lecture RDMA sont utilisés pour porter les informations de contrôle et toutes les informations de données associées aux PDU Type de données iSCSI (c'est-à-dire, les PDU SCSI Data-In et R2T). iSER n'utilise pas les PDU Data-Out SCSI pour les données sollicitées, et les PDU Data-Out SCSI pour les données non sollicitées ne sont pas traitées comme des PDU Type de données iSCSI par iSER parce que RDMA n'est pas utilisé. Voir au paragraphe 7.1 les détails sur les PDU Type de données iSCSI.

5. La cible pilote tous les transferts de données (à l'exception des données non sollicitées iSCSI) pour les opérations d'écriture et de lecture SCSI, en produisant respectivement des demandes RDMA Lecture et RDMA Écriture.

6. RCaP est chargé d' s'assurer de l'intégrité des données. (Par exemple, iWARP inclut une couche de tramage à CRC amélioré appelée MPA par dessus TCP; et pour InfiniBand, les CRC sont inclus dans le mode de connexion fiable). Pour cette raison, l'en-tête iSCSI et les résumés de données sont négociés à "Aucun" pour les sessions iSCSI/iSER.

7. La hiérarchie iSCSI de récupération d'erreur définie dans la [RFC7143] est pleinement prise en charge par iSER. (Cependant, voir le paragraphe 7.3.11 sur le traitement des PDU Demande de SNACK.)

8. iSER n'exige pas de changement aux mécanismes de négociation de mode de sécurité et texte iSCSI.

Noter que les mises en œuvre iSCSI traditionnelles peuvent devoir être adaptées pour employer iSER. On s'attend à ce que quand l'adaptation est exigée, elle soit centrée sur les exigences d'interface de couche supérieure de iSER (Section 3).

1.5 Services RDMA et iSER

iSER est conçu pour fonctionner sur des piles de protocole de logiciel et/ou matériel fournissant les services de protocole définis dans les documents RCaP tels que la [RFC5040], [IB], etc. Les paragraphes qui suivent décrivent les éléments clés de protocole des services RCaP sur lesquels s'appuie iSER.

1.5.1 STag

Une STag est l'identifiant d'une mémoire tampon I/O unique d'un contrôleur à capacité RDMA que la couche iSER annonce au nœud iSCSI/iSER distant afin de réaliser une entrée/sortie SCSI complète.

Une annonce iSER est l'acte de l'initiateur d'informer la cible qu'une mémoire tampon I/O est disponible chez l'initiateur pour un accès RDMA en lecture ou écriture par la cible. L'initiateur annonce la mémoire tampon I/O en incluant la STag et le décalage de base dans l'en-tête d'un message iSER contenant la PDU Commande SCSI à la cible. La longueur de mémoire tampon est celle spécifiée dans la PDU Commande SCSI.

La couche iSER chez l'initiateur annonce la STag et le décalage de base pour la mémoire tampon I/O de chaque I/O SCSI à la couche iSER chez la cible dans l'en-tête iSER d'un message Send contenant la PDU Commande SCSI, sauf si la I/O peut être complètement satisfaite par des données non sollicitées seules. Le message SendSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP).

La couche iSER chez la cible fournit la STag pour la mémoire tampon d'entrée/sortie qui est le collecteur de données d'une opération de lecture RDMA (paragraphe 1.5.4) à la couche RCaP sur le nœud de l'initiateur -- c'est-à-dire, ceci est complètement transparent à la couche iSER chez l'initiateur.

La couche iSER chez l'initiateur DEVRAIT invalider la STag annoncée à l'achèvement normal de la tâche associée. Le message Send avec Invalider, si il est pris en charge par la couche RCaP (par exemple, iWARP) peut être utilisé pour une invalidation automatique quand il est utilisé pour porter la PDU Réponse SCSI. Il y a deux exceptions à cette invalidation automatique – les commandes bidirectionnelles et l'achèvement anormal d'une commande. La couche iSER chez l'initiateur DEVRAIT explicitement invalider la STag dans ces deux cas. Cette couche iSER DOIT vérifier que l'invalidation de STag s'est produite chaque fois que la réception d'un message Send avec Invalider est le moyen attendu de causer l'invalidation d'une STag, et elle DOIT effectuer l'invalidation de la STag si celle-ci n'a pas été déjà invalidée (par exemple, parce qu'un message Send a été utilisé à la place d'un Send avec Invalider).

Si la STag annoncée n'est pas invalidée comme recommandé dans le paragraphe précédent (par exemple, afin de mettre en antémémoire la STag pour une réutilisation future) la mémoire tampon d'entrée/sortie reste exposée au réseau pour accès par le RCaP. Une telle mémoire tampon d'entrée/sortie est capable d'être lue ou écrite par le RCaP en dehors de l'opération iSCSI pour laquelle elle a été établie à l'origine ; ce fait pose des problèmes de robustesse et de sécurité. Le problème de robustesse est que le système qui contient l'initiateur iSER peut mal réagir à une modification inattendue de sa mémoire. Pour les considérations de sécurité, voir la Section 11.

1.5.2 Send

Send est l'opération RDMA qui n'est pas visée par une mémoire tampon annoncée et utilise des mémoires tampon non étiquetées lorsque le message est reçu.

La couche iSER chez l'initiateur utilise l'opération Send pour transmettre toute PDU Type de contrôle iSCSI à la cible. Par exemple, l'initiateur utilise des opérations Send pour transférer des messages iSER contenant des PDU Commande SCSI à la couche iSER chez la cible.

Une couche iSER chez la cible utilise l'opération Send pour transmettre toute PDU Type de contrôle iSCSI à l'initiateur. Par exemple, la cible utilise les opérations Send pour transférer les messages iSER contenant des PDU Réponse SCSI à la couche iSER chez l'initiateur.

Pour l'interopérabilité, les mises en œuvre iSER DEVRAIENT accepter et traiter correctement les messages SendSE et SendInvSE. Cependant, les messages SendSE et SendInvSE sont à considérer comme des optimisations ou des améliorations du message Send de base, et leur prise en charge peut varier selon le protocole RCaP et les mises en œuvre spécifiques. En général, ces messages NE DEVRAIENT PAS être utilisés, sauf si le RCaP exige leur prise en charge dans toutes les mises en œuvre. Si ces messages sont utilisés, la mise en œuvre DEVRAIT être capable de revenir à l'usage de Send afin de travailler avec un receveur qui ne prend pas en charge ces messages. Tenter d'utiliser ces messages avec un homologue qui ne les prend pas en charge peut résulter en une erreur fatale qui ferme la connexion RcaP. Par exemple, ces messages NE DEVRAIENT PAS être utilisés avec le RCaP InfiniBand parce que InfiniBand n'exige pas leur prise en charge dans tous les cas. Les nouvelles mises en œuvre iSER DEVRAIENT utiliser Send (et non SendSE ou SendInvSE) sauf si il y a des raisons impérieuses pour faire autrement. De même, les mises en œuvre iSER NE DEVRAIENT PAS s'appuyer sur des événements déclenchés par SendSE et SendInvSE, car ces messages peuvent n'être pas utilisés.

1.5.3 RDMA Écriture *(Write)*

RDMA Écriture est l'opération RDMA qui est utilisée pour placer des données dans une mémoire tampon annoncée au collecteur de données. La source des données adresse le message en utilisant une STag et un décalage étiqueté qui sont valides sur le collecteur de données.

La couche iSER chez la cible utilise l'opération Écriture RDMA pour transférer le contenu d'une mémoire tampon d'entrée/sortie locale à une mémoire tampon d'entrée/sortie annoncée chez l'initiateur. La couche iSER chez la cible utilise le RDMA Écriture pour transférer tout ou partie des données requises pour achever la commande Lire SCSI.

La couche iSER chez l'initiateur n'emploie pas de RDMA Écriture.

1.5.4 RDMA Lire *(Read)*

RDMA Lire est l'opération RDMA qui est utilisée pour restituer des données d'une mémoire tampon annoncée chez la source des données. L'envoyeur de la demande de lecture RDMA adresse le message en utilisant une STag et un décalage étiqueté qui sont valides sur la source des données en plus de fournir une STag et un décalage locaux valides qui identifient le collecteur de données.

La couche iSER à la cible utilise l'opération de lecture RDMA pour transférer le contenu d'une mémoire tampon d'entrée/sortie annoncée chez l'initiateur à une mémoire tampon d'entrée/sortie locale à la cible. La couche iSER à la cible utilise le RDMA Lire pour aller chercher tout ou partie des données requises pour achever une commande Écriture SCSI.

La couche iSER chez l'initiateur n'emploie pas les RDMA Lire.

1.6 Vue d'ensemble de l'opération Lire SCSI

La couche iSER chez l'initiateur reçoit la PDU Commande SCSI de la couche iSCSI. La couche iSER chez l'initiateur génère une STag pour la mémoire tampon d'entrée/sortie du Lire SCSI et annonce la mémoire tampon en incluant la STag et le décalage de base au titre de l'en-tête iSER pour la PDU. Le message iSER est transféré à la cible en utilisant un message Send. Le message SendSE devrait être utilisé si il est accepté par la couche RCaP (par exemple, iWARP).

La couche iSER chez la cible utilise un ou plusieurs Écriture RDMA pour transférer les données requises pour achever le Lire SCSI.

La couche iSER chez la cible utilise un message Send pour retransférer la PDU Réponse SCSI à la couche iSER chez l'initiateur. La couche iSER chez l'initiateur invalide la STag et notifie à la couche iSCSI la disponibilité de la PDU Réponse SCSI. Le message Send avec Invalider, si il est pris en charge par la couche RCaP (par exemple, iWARP) peut être utilisé pour l'invalidation automatique de la STag.

1.7 Vue d'ensemble de l'opération Write SCSI

La couche iSER chez l'initiateur reçoit la PDU Commande SCSI de la couche iSCSI. Si un transfert de données sollicitées est impliqué, la couche iSER chez l'initiateur génère une STag pour la mémoire tampon d'entrée/sortie du Écriture SCSI et annonce la mémoire tampon en incluant la STag et le décalage de base au titre de l'en-tête iSER pour la PDU. Le message iSER est transféré à la cible en utilisant un message Send. Le message SendSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP).

La couche iSER chez l'initiateur peut facultativement envoyer une ou plusieurs PDU de données non sollicitées non immédiates à la cible en utilisant des messages Send.

Si un transfert de données sollicitées est impliqué, la couche iSER à la cible utilise un ou plusieurs Lire RDMA pour transférer les données requises pour achever le Écriture SCSI.

La couche iSER chez la cible utilise un message Send pour retransférer la PDU Réponse SCSI à la couche iSER chez l'initiateur. La couche iSER chez l'initiateur invalidé la STag et notifie à la couche iSCSI la disponibilité de la PDU Réponse SCSI. Le message Send avec Invalider, si il est pris en charge par la couche RCaP (par exemple, iWARP) peut être utilisé pour l'invalidation automatique de la STag.

2. Définitions et acronymes

2.1 Définitions

Annonce (annoncé, annoncer, annonces) – Acte d'informer une couche d'extensions iSCSI pour RDMA (iSER, *iSCSI Extensions for RDMA*) distante que la mémoire tampon d'un nœud local lui est disponible. Un nœud rend une mémoire tampon disponible pour l'accès d'un message de demande de lecture RDMA entrant ou d'un message en écriture RDMA entrant en informant la couche iSER distante des identifiants de mémoirer tampon étiquetée (STag, décalage de base, et longueur de mémoire tampon). Noter que cette annonce des informations de mémoire tampon étiquetée est de la responsabilité de la couche iSER à l'une et l'autre extrémité et n'est pas définie par le protocole à capacité RDMA. Une méthode normale serait que la couche iSER incorpore la STag, le décalage de base et la longueur de mémoire tampon de la mémoire tampon étiquetée dans un message destiné à la couche iSER distante.

Décalage de base *(Base Offset)* : valeur qui, ajoutée au décalage de mémoire tampon, forme le décalage étiqueté.

Achèvement *(Completion)* (achevé) : l'achèvement est défini comme le processus par lequel la couche de protocole à capacité RDMA informe la couche iSER qu'une opération RDMA particulière a effectué toutes les fonctions spécifiées pour l'opération RDMA.

Connexion : une connexion est un canal bidirectionnel de communication logique entre l'initiateur et la cible, par exemple, une connexion TCP. La communication entre l'initiateur et la cible se produit sur une ou plusieurs connexions. Les connexions portent les messages de contrôle, les commandes SCSI, les paramètres, et les données au sein des unités de données de protocole iSCSI (des PDU iSCSI).

Bride de connexion *(Connection Handle)* : élément d'information qui identifie la connexion iSCSI particulière et est unique pour une certaine couche iSCSI et la couche iSER sous-jacente. Chaque invocation d'une primitive opérationnelle est qualifiée avec la bride de connexion.

Collecteur de données *(Data Sink)* : l'homologue qui reçoit une charge utile de données. Noter que le collecteur de données peut être obligé de recevoir et d'envoyer des messages RCaP (protocole à capacité RDMA) 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 RCaP pour transférer une charge utile de données.

Interface Datamover (DI) : interface entre la couche iSCSI et la couche Datamover comme décrite dans la [RFC5047].

Couche Datamover : couche qui est directement en dessous de la couche iSCSI et au dessus des couches de transport sous-jaacentes. Cette couche expose et utilise un ensemble de primitives opérationnelles indépendantes du transport pour la communication entre la couche iSCSI et elle-même. La couche Datamover, fonctionnant en conjonction avec les couches transport, déplace les informations de contrôle et de données sur la connexion iSCSI. Dans la présente spécification, la couche iSER est la couche Datamover.

Protocole Datamover : c'est le protocole réseau qui est défini pour réaliser la fonction de couche Datamover. Dans la présente spécification, le protocole iSER est le protocole Datamover.

Profondeur de file d'attente de lecture RDMA entrante (IRD, *Inbound RDMA Read Queue Depth*) : nombre maximum de demandes en instance de lecture RDMA entrantes que le contôleur à capacité RDMA peut traiter sur un flux RCaP particulier à la source des données. Pour certaines couches de protocole à capacité RDMA, le terme "IRD" peut être connu sous un nom différent. Par exemple, pour InfiniBand, l'équivalent de IRD est "Ressources de réponse".

mémoire tampon d'entrée/sortie : mémoire tampon qui est utilisée dans une opération SCSI de lecture ou d'écriture de telle sorte que les données SCSI peuvent être envoyées de ou reçues dans cette mémoire tampon.

iSCSI : le protocole iSCSI défini dans la [RFC7143] est une transposition du modèle d'architecture SCSI de SAM-5 sur TCP.

PDU Type de contrôle iSCSI : toute PDU iSCSI qui n'est pas une PDU Type de données iSCSI ni une PDU Data-Out SCSI portant des données sollicitées est définie comme PDU Type de contrôle iSCSI. Précisément, on notera que les PDU Data-Out SCSI pour les données non sollicitées sont définies comme PDU Type de contrôle iSCSI.

PDU Type de données iSCSI : une PDU Type de données iSCSI est définie comme une PDU iSCSI qui cause un transfert de données via des opérations RDMA à la couche iSER, transparente à la couche iSCSI distante, qui a lieu entre les nœuds iSCSI homologues sur une connexion iSCSI en phase de pleines caractéristiques. Une PDU Type de données iSCSI, quand elle est demandée pour la transmission par la couche iSCSI de l'envoyeur, résulte en le transfert de données associées sans la participation de la couche iSCSI distante, c'est-à-dire que la PDU elle-même n'est pas livrée telle qu'elle à la couche iSCSI distante. Les PDU iSCSI suivantes constituent l'ensemble des PDU Type de données iSCSI : PDU Data-In SCSI et PDU R2T.

Couche iSCSI : couche de la mise en œuvre de la pile de protocoles au sein d'un nœud d'extrémité qui met en œuvre le protocole iSCSI et fait l'interface avec la couche iSER via l'interface Datamover.

PDU iSCSI *(iSCSI Protocol Data Unit)* : la couche iSCSI chez l'initiateur et la couche iSCSI chez la cible divisent leurs communications en messages. Le terme de "unité de données de protocole iSCSI" (PDU iSCSI) est utilisé pour ces messages.

Connexion iSCSI/iSER : connexion iSCSI assistée par iSER. Une connexion iSCSI qui n'est pas assistée par iSER se transpose toujours en une connexion TCP au niveau transport. Mais une connexion iSCSI assistée par iSER peut n'avoir pas une connexion TCP sous-jacente. Pour certaines mises en œuvre RcaP (par exemple, iWARP), une connexion iSCSI assistée par iSER a une connexion TCP sous-jacente. Pour les autres mises en œuvre RcaP (par exemple, InfiniBand) il n'y a pas de connexion TCP sous-jacente. (Dans l'exemple spécifique de InfiniBand [IB], une connexion iSCSI assistée par iSER se transpose directement en le canal InfiniBand fondé sur une connexion fiable (RC).)

Session iSCSI/iSER : session iSCSP assistée par iSER. Toutes les connexions d'une session iSCSI/iSER sont des connexions iSCSI/iSER.

iSER : extensions iSCSI pour RDMA, le protocole défini dans le présent document.

Assisté par iSER : terme généralement utilisé pour décrire le fonctionnement de iSCSI quand la fonction iSER est aussi activée en dessous de la couche iSCSI pour la connexion iSCSI/iSER en question.

iSER-IRD : cette variable représente le nombre maximum de demandes de lecture RDMA entrantes en instance que la couche iSER chez l'initiateur accorde sur un flux RCaP particulier.

iSER-ORD : cette variable représente le nombre maximum de demandes de lecture RDMA en instance que la couche iSER peut initier sur un flux RCaP particulier. Cette variable n'est conservée que par la couche iSER chez la cible.

Couche iSER : couche qui met en œuvre le protocole d'extensions iSCSI pour RDMA (iSER).

iWARP : suite de protocoles réseau incluant les [RFC5040], [RFC5041], et [RFC5044] quand elle est mise en couche par dessus la [RFC0793]. Les [RFC5040] et [RFC5041] peuvent être mises en couche par dessus SCTP ou d'autres protocoles de transport.

Transposition locale : enregistrement d'état de tâche tenu par la couche iSER qui associe l'étiquette de tâche d'initiateur aux étiquettes d'état local (STag). Les spécificités de la structure de l'enregistrement dépendent de la mise en œuvre.

Homologue local : mise en œuvre du protocole à capacité RDMA sur l'extrémité locale de la connexion. Utilisé pour se référer à l'entité locale lors de la description des échanges de protocole ou autres interactions 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 application spécifique ou instanciation de protocole fonctionnant sur l'ordinateur. Un nœud peut consister en un ou plusieurs contrôleurs à capacité RDMA installés dans un ordinateur hôte.

Primitive opérationnelle : c'est une procédure d'interface fonctionnelle abstraite qui demande à une autre couche d'effectuer une action spécifique au nom du demandeur ou qui notifie un événement à l'autre couche. L'interface Datamover entre une couche iSCSI et une couche Datamover au sein d'un nœud d'extrémité iSCSI utilise un ensemble de primitives opérationnelles pour définir l'interface fonctionnelle entre les deux couches. Noter que toute invocation d'une primitive opérationnelle ne peut pas provoquer une réponse de la couche demandée. On trouvera une discussion complète de la sémantique des types et demandes/réponse de primitives opérationnelles disponibles pour iSCSI et iSER dans la [RFC5047].

Profondeur de file d'attente de lecture RDMA sortante (ORD, *Outbound RDMA Read Queue Depth*) : nombre maximum de demandes de lecture RDMA en instance que le contrôleur à capacité RDMA peut initier sur un flux RCaP particulier au collecteur de données. Pour certaines couches de protocole à capacité RDMA, le terme "ORD" peut avoir un nom différent. Par exemple, pour InfiniBand, l'équivalent de ORD est la profondeur d'initiateur.

Collapsus de phase : se réfère à l'optimisation dans iSCSI où l'état SCSI est transféré avec la PDU Data-In SCSI finale d'une cible. Voir au paragraphe 4.2 de la [RFC7143].

Message RCaP : un ou plusieurs paquets de la couche réseau qui constituent une seule opération RDMA ou une partie d'une opération de lecture RDMA du protocole à capacité RDMA. Pour iWARP, un message RCaP est appelé un message RDMAP.

Flux RCaP : une seule association bidirectionnelle entre les couches de protocole à capacité RDMA homologues sur deux nœuds sur un seul flux de niveau transport. Pour iWARP, un flux RCaP est appelé un flux RDMAP, et l'association est créée à la suite d'une phase d'établissement réussie durant laquelle la prise en charge d'iSER est négociée.

Protocole à capacité RDMA (RcaP, *RDMA-Capable Protocol*) : protocole ou suite de protocoles qui fournit une fonctionnalité de transport RDMA fiable, par exemple, iWARP, InfiniBand, etc.

Contrôleur à capacité RDMA : adaptateur d'entrée/sortie réseau ou contrôleur incorporé avec fonctionnalité RDMA. Par exemple, pour iWARP, ce pourrait être un RNIC, et pour InfiniBand, ce pourrait être un adaptateur de canal hôte (HCA, *Host Channel Adapter*) ou un adaptateur de canal cible (TCA, *Target Channel Adapter*).

Contrôleur d'interface réseau à capacité RDMA (RNIC, *RDMA-enabled Network Interface Controller*) : adaptateur d'entrée/sortie réseau ou contrôleur incorporé avec fonctionnalité iWARP.

Opération RDMA : séquence de messages RCaP, 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 : opération d'écriture RDMA, opération de lecture RDMA, et opération Send.

Protocole RDMA (RDMAP) : protocole réseau sui prend en charge les opérations RDMA pour transférer des données ULP entre un homologue local et l'homologue distant, comme décrit dans la [RFC5040].

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

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

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

Opération d'écriture RDMA : opération RDMA utilisée par la source des données pour transférer le contenu d'une mémoire tampon de source de données de l'homologue local à une mémoire tampon de collecteur de données chez l'homologue distant. Le message d'écriture RDMA ne décrit que la mémoire tampon de collecteur de données.

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

Transposition distante : enregistrement d'état de tâche tenu par la couche iSER qui associe l'étiquette de tâche d'initiateur à la ou les STag annoncées et au ou aux décalages de base. Les spécificités de la structure de l'enregistrement dépendent de la mise en œuvre.

Homologue distant : mise en œuvre du protocole à capacité RDMA sur l'extrémité opposée de la connexion. Utilisé pour se référer à l'entité distante lors de la description des échanges de protocole ou autres interactions entre deux nœuds.

Couche SCSI : cette couche construit/reçoit des blocs de descripteur de commande SCSI (CDB, *Command Descriptor Block*) et les envoie/reçoit avec le reste des paramètres d'exécution de commande [SAM5] de/vers la couche iSCSI.

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

Message SendInvSE : un Send avec événement sollicité et message Invalider.

Message SendSE : un Send avec message d'événement sollicité.

Numéro de séquence (SN, *Sequence Number*) : DataSN pour une PDU SCSI Data-In et R2TSN pour une PDU R2T. La sémantique de ces deux types de numéros de séquence est définie dans la [RFC7143].

Session, session iSCSI : le groupe de connexions qui relient un accès d'initiateur SCSI à un accès de cible SCSI forme une session iSCSI (équivalente à un nexus Initiateur-Cible (I-T) SCSI). Les connexions peuvent être ajoutées à, et supprimées d'une session même lorsque le nexus I-T est intact. À travers les connexions au sein d'une session, un initiateur voit une seule et même cible.

Étiquette de pilotage (STag, *Steering Tag*) : identifiant d'une mémoire tampon étiquetée sur un nœud (local ou distant) comme défini dans les [RFC5040] et [RFC5041]. Pour les autres protocoles à capacité RDMA, l'étiquette de pilotage peut être appelée différemment mais sera appelée ici STag. Par exemple, pour InfiniBand, une STag distante est appelée R-Key, et une STag locale une L-Key, et toutes deux sont considérées comme des STag.

Mémoire tampon étiquetée *(Tagged Buffer)* : mémoire tampon qui est explicitement annoncée à la couche iSER chez le nœud distant par l'échange d'une STag, du décalage de base, et de la longueur.

Décalage étiqueté (*Tagged Offset*) : décalage au sein d'une mémoire tampon étiquetée.

iSCSI traditionnel : se réfère au protocole iSCSI défini dans la [RFC7143] (c'est-à-dire, sans les améliorations iSER).

Mémoire tampon non étiquetée *(Untagged Buffer)* : mémoire tampon qui n'est pas explicitement annoncée à la couche iSER chez le nœud distant.

2.2 Acronymes

AHS *(Additional Header Segment)* : segment d'en-tête supplémentaire
BHS *(Basic Header Segment)* : segment d'en-tête de base
CO *(Connection Only)* : seulement en connexion
CRC *(Cyclic Redundancy Check)* : contrôle de redondance cyclique
DDP *(Direct Data Placement Protocol)* : protocole de placement direct des données
DI *(Datamover Interface)* : interface Datamover
HCA *(Host Channel Adapter*) : adaptateur de canal hôte
IANA *(Internet Assigned Numbers Authority)* : Autorité d'allocation des numéros de l'Internet
IB *(InfiniBand)* : InfiniBand
IETF *(Internet Engineering Task Force)* : équipe d'ingénierie de l'Internet
I/O *(Input – Output)* : entrée/sortie
IO *(Initialize Only)* : seulement en initialisation
IP *(Internet Protocol)* : protocole Internet
IPoIB *(IP over InfiniBand)* : IP sur InfiniBand
IPsec *(Internet Protocol Security)* : sécurité du protocole Internet
iSER *(iSCSI Extensions for RDMA)* : extension iSCSI pour RDMA
ITT *(Initiator Task Tag)* : étiquette de tâche d'initiateur
LO *(Leading Only)* : seulement de tête
MPA *(Marker PDU Aligned Framing for TCP)* : marqueur de tramage aligné sur la PDU pour TCP
NOP *(No Operation)* : non fonctionnement

NSG *(Next Stage)* : prochaine étape (durant la phase d'établissement iSCSI)
PDU *(Protocol Data Unit)* :unité de données de protocole
R2T *(Ready To Transfer)* : prêt au transfert
R2TSN *(Ready To Transfer Sequence Number)* : numéro de séquence de prêt au transfert
RCaP *(RDMA-Capable Protocol)* : protocole à capacité RDMA
RDMA *(Remote Direct Memory Access)* : accès direct à la mémoire distante
RDMAP *(Remote Direct Memory Access Protocol)* : protocole d'accès direct à la mémoire distante
RFC *(Request For Comments)* : demande de commentaires
RNIC *(RDMA-enabled Network Interface Controller)* : contrôleur d'interface réseau à capacité RDMA
SAM5 *(SCSI Architecture Model – 5)* : modèle d'architecture SCSI version 5
SCSI *(Small Computer System Interface)* : interface de système de petit ordinateur
SNACK *(Selective Negative Acknowledgment)* : accusé de réception négatif sélectif – aussi accusé de réception de numéro de séquence pour des données
STag *(Steering Tag)* : étiquette de pilotage
SW *(Session Wide)* : à l'échelle de la session
TCA *(Target Channel Adapter)* : adaptateur de canal cible
TCP *(Transmission Control Protocol)* : protocole de contrôle de transmission
TMF *(Task Management Function)* : fonction de gestion de tâche
TTT *(Target Transfer Tag)* : étiquette de transfert de cible
ULP *(Upper Level Protocol)* : protocole de niveau supérieur

2.3 Conventions

Les mots clés "DOIT", "NE DOIT PAS", "EXIGE", "DEVRA", "NE DEVRA PAS", "DEVRAIT", "NE DEVRAIT PAS", "RECOMMANDE", "PEUT", et "FACULTATIF" dans ce document sont à interpréter comme décrit dans la [RFC2119].

3. Exigences pour l'interface de couche supérieure

Cette section expose les exigences d'interface de couche supérieure sous la forme d'un modèle abstrait des interactions requises entre la couche iSCSI et la couche iSER. Le modèle abstrait utilisé ici est dérivé du modèle architectural décrit dans la [RFC5047]. La [RFC5047] fournit aussi une vue d'ensemble fonctionnelle des interactions entre la couche iSCSI et la couche Datamover comme prévu dans l'architecture Datamover.

Les exigences d'interface sont spécifiées par les primitives opérationnelles. Une primitive opérationnelle est une procédure abstraite d'interface fonctionnelle entre la couche iSCSI et la couche iSER qui demande à une couche d'effectuer une action spécifique au nom de l'autre couche ou notifie à l'autre couche un événement. Chaque fois qu'une primitive opérationnelle est invoquée, le qualificatif Connection_Handle est utilisé pour identifier une connexion iSCSI particulière. Pour certaines primitives opérationnelles, un descripteur de données *(Data_Descriptor)* est utilisé pour identifier la mémoire tampon de données iSCSI/SCSI associée à l'opération demandée ou achevée.

Le modèle abstrait et les primitives opérationnelles définis dans cette section facilitent la description du protocole iSER. Dans le reste de la spécification iSER, les déclarations de confomité relatives à l'utilisation de ces primitives opérationnelles sont seulement pour les besoins des interactions requises entre la couche iSCSI et la couche iSER. Noter que les déclarations de conformité relatives aux primitives opérationnelles dans le reste de cette spécification ne rendent obligatoire que l'équivalence fonctionnelle des mises en œuvre, mais ne posent aucune exigence sur les spécificités de mise en œuvre de l'interface entre la couche iSCSI et la couche iSER.

Chaque primitive opérationnelle est invoquée avec un ensemble de qualificatifs qui spécifient le contexte d'information pour effectuer l'action spécifique demandée à la primitive opérationnelle. Bien que les qualificatifs soient exigés, la méthode de réalisation des qualificatifs (par exemple, en les passant synchrones avec l'invocation, ou en les restituant du contexte de tâche, ou en les restituant d'une mémoire partagée, etc.) dépend de la mise en œuvre.

3.1 Primitives de fonctionnement offertes par iSER

La couche de protocole iSER DOIT prendre en charge les primitives opérationnelles suivantes à utiliser par la couche de protocole iSCSI.

3.1.1 Send_Control

Qualificatifs d'entrée : Connection_Handle, BHS et AHS (si il en est) de la PDU iSCSI, qualificatifs spécifiques de PDU.

Résultat retouné : Non spécifié

Ceci est utilisé par les couches iSCSI chez l'initiateur et la cible pour demander le transfert sortant d'une PDU Type de contrôle iSCSI (voir au paragraphe 7.2). Les qualificatifs qui ne s'appliquent qu'à une PDU particulière de type de contrôle sont appelés des qualificatifs spécifiques de PDU, par exemple, ImmediateDataSize pour une commande Écriture SCSI. Pour les détails sur les qualificatifs spécifiques de PDU, voir au paragraphe 7.3. La couche iSCSI peut seulement invoquer la primitive opérationnelle Send_Control quand la connexion est en mode à assistance iSER.

3.1.2 Put_Data

Qualificatifs d'entrée : Connection_Handle, contenu d'un en-tête de PDU Data-In SCSI, Data_Descriptor, Notify_Enable

Résultat retourné : non spécifié

Ceci est utilisé par la couche iSCSI chez la cible pour demander le transfert sortant de données pour une PDU Data-In SCSI à partir de la mémoire tampon identifiée par le qualificatif Data_Descriptor. La couche iSCSI peut seulement invoquer la primitive opérationnelle Put_Data quand la connexion est en mode à assistance iSER.

Le qualificatif Notify_Enable est utilisé pour indiquer à la couche iSER si elle devrait ou non générer une éventuelle notification d'achèvement local à la couche iSCSI. Voir au paragraphe 3.2.2 les détails sur Data_Completion_Notify.

3.1.3 Get_Data

Qualificatifs d'entrée : Connection_Handle, contenu d'une PDU R2T, Data_Descriptor, Notify_Enable

Résultat retourné : non spécifié

Ceci est utilisé par la couche iSCSI chez la cible pour demander le transfert entrant des données sollicitées demandées par une PDU R2T dans la mémoire tampon identifiée par le qualificatif Data_Descriptor. La couche iSCSI ne peut invoquer la primitive opérationnelle Get_Data que quand la connexion est en mode à assistance iSER.

Le qualificatif Notify_Enable est utilisé pour indiquer à la couche iSER si elle devrait ou non générer la notification éventuelle d'achèvement local à la couche iSCSI. Voir au paragraphe 3.2.2 les détails de Data_Completion_Notify.

3.1.4 Allocate_Connection_Resources

Qualificatifs d'entrée : Connection_Handle, Resource_Descriptor (facultatif)

Résultat retourné : État

Ceci est utilisé par les couches iSCSI chez l'initiateur et la cible pour demander l'allocation de toutes les ressources de connexion nécessaires pour prendre en charge RCaP pour une connexion iSCSI/iSER opérationnelle. La couche iSCSI peut facultativement spécifier les exigences de ressource spécifiques de la mise en œuvre pour la connexion iSCSI en utilisant le qualificatif Resource_Descriptor.

Un résultat retourné de État=succès signifie que l'invocation a réussie, et un résultat retourné de État=échec signifie que l'invocation a échoué. Si l'invocation est pour une Connection_Handle pour laquelle une invocation antérieure avait réussi, la demande sera ignorée par la couche iSER et le résultat de État=succès sera retourné. Une seule invocation de la primitive opérationnelle Allocate_Connection_Resources peut être en instance à tout instant pour une Connection_Handle donnée.

3.1.5 Deallocate_Connection_Resources

Qualificatifs d'entrée : Connection_Handle

Résultat retourné : non spécifié

Ceci est utilisé par les couches iSCSI chez l'initiateur et la cible pour demander la désallocation de toutes les ressources de connexion qui ont été allouées antérieurement par suite d'une invocation réussie de la primitive opérationnelle Allocate_Connection_Resources.

3.1.6 Enable_Datamover

Qualificatifs d'entrée : Connection_Handle, Transport_Connection_Descriptor, Final_Login_Response_PDU (facultatif)

Résultat retourné : non spécifié

Ceci est utilisé par les couches iSCSI chez l'initiateur et la cible pour demander que le mode à assistance iSER soit utilisé pour la connexion. Le qualificatif Transport_Connection_Descriptor est utilisé pour identifier la connexion spécifique associée à la Connection_Handle. La couche iSCSI ne peut invoquer la primitive opérationnelle Enable_Datamover que quand il y a eu une allocation de ressources correspondantes antérieure.

Le qualificatif d'entrée Final_Login_Response_PDU n'est applicable que pour une cible et contient la PDU Réponse finale d'établissement qui conclut la phase d'établissement iSCSI.

3.1.7 Connection_Terminate

Qualificatifs d'entrée : Connection_Handle

Résultat retourné : non spécifié.

Ceci est utilisé par les couches iSCSI chez l'initiateur et la cible pour demander qu'une connexion iSCSI/iSER spécifiée soit terminée et que toutes les connexions et ressources de tâche associées soient libérées. Quand cette invocation de primitive opérationnelle retourne à la couche iSCSI, la couche iSCSI peut supposer la pleine possession de toutes les ressources de niveau iSCSI, par exemple, des mémoires tampon d'entrée/sortie, associées à la connexion.

3.1.8 Notice_Key_Values

Qualificatifs d'entrée : Connection_Handle, nombre de clés, liste des paires de clé-valeur.

Résultat retourné : non spécifié.

Ceci est utilisé par les couches iSCSI chez l'initiateur et la cible pour demander à la couche iSER de prendre note des paires clé-valeur spécifiées qui ont été négociées par les homologues iSCSI pour la connexion.

3.1.9 Deallocate_Task_Resources

Qualificatifs d'entrée : Connection_Handle, ITT.

Résultat retourné : non spécifié.

Ceci est utilisé par les couches iSCSI chez l'initiateur et la cible pour demander la désallocation de toutes les ressources spécifiques de RCaP allouées par la couche iSER pour la tâche identifiée par le qualificatif ITT. La couche iSER peut exiger un certain nombre de ressources spécifiques de RCaP associées à l'ITT pour chaque nouvelle tâche iSCSI. Dans le cours normal d'exécution, ces ressources au niveau de la tâche dans la couche iSER sont supposées être allouées de façon transparente sur chaque initiation de tâche et désallouées à la conclusion de chaque tâche comme approprié. Dans les scénarios d'exception où la tâche ne se conclut pas avec une PDU Réponse SCSI, la couche iSER a besoin d'une notification des terminaisons individuelles de tâche pour aider à sa gestion des ressources au niveau de la tâche. Cette primitive opérationnelle est utilisée à cette fin et n'est pas nécessaire lorsque une PDU Réponse SCSI conclut normalement une tâche. Noter que les ressources de tâche spécifiques de RCaP sont désallouées par la couche iSER quand une PDU Réponse SCSI conclut normalement une tâche, même si l'état SCSI n'était pas un succès.

3.2 Primitives de fonctionnement utilisées par iSER

La couche iSER DOIT utiliser les primitives opérationnelles suivantes, offertes par la couche de protocole iSCSI lorsque la connexion est en mode à assistance iSER.

3.2.1 Control_Notify

Qualificatifs d'entrée : Connection_Handle, une PDU Type de contrôle iSCSI.

Résultat retourné : non spécifié.

Ceci est utilisé par les couches iSER chez l'initiateur et la cible pour notifier à la couche iSCSI la disponibilité d'une PDU Type de contrôle iSCSI entrante. Une PDU est décrite comme "disponible" à la couche iSCSI quand la couche iSER notifie à la couche iSCSI la réception de cette PDU entrante, avec une indication spécifique de la mise en œuvre de l'endroit où se trouve la PDU reçue.

3.2.2 Data_Completion_Notify

Qualificatifs d'entrée : Connection_Handle, ITT, SN.

Résultat retourné : non spécifié.

Ceci n'est utilisé par la couche iSER pour notifier à la couche iSCSI l'achèvement du transfert de données sortant qui était demandé par la couche iSCSI que si l'invocation de la primitive opérationnelle Put_Data (voir au paragraphe 3.1.2) était qualifiée avec Notify_Enable établi. SN se réfère au DataSN associé à la PDU Data-In SCSI.

Ceci n'est utilisé par la couche iSER pour notifier à la couche iSCSI l'achèvement du transfert de données entrant qui était demandé par la couche iSCSI que si l'invocation de la primitive opérationnelle Get_Data (voir au paragraphe 3.1.3) était qualifiée avec Notify_Enable établi. SN se réfère au R2TSN associé à la PDU R2T.

3.2.3 Data_ACK_Notify

Qualificatif d'entrée : Connection_Handle, ITT, DataSN.

Résultat retourné : non spécifié.

Ceci est utilisé par la couche iSER chez la cible pour notifier à la couche iSCSI l'arrivé de l'accusé de réception des données (comme défini dans la [RFC7143]) demandé antérieurement par la couche iSCSI pour le transfert de données sortant via une invocation de la primitive opérationnelle Put_Data où le bit A dans la PDU Data-In SCSI est réglé à un. Voir au paragraphe 7.3.5. DataSN se réfère au DataSN attendu de la prochaine PDU Data-In SCSI qui suit immédiatement la PDU Data-In SCSI avec le bit A établi à laquelle cette notification correspond ; sa sémantique est définie dans la [RFC7143].

3.2.4 Connection_Terminate_Notify

Qualificatifs d'entrée : Connection_Handle.

Résultat retourné : non spécifié.

Ceci est utilisé par les couches iSER chez l'initiateur et la cible pour notifier à la couche iSCSI la terminaison non sollicitée ou la défaillance d'une connexion iSCSI/iSER. La couche iSER DOIT désallouer la connexion et les ressources de tâche associées à la connexion terminée avant l'invocation de cette primitive opérationnelle. Noter que la primitive opérationnelle Connection_Terminate_Notify n'est pas invoquée quand la terminaison de la connexion avait été demandée antérieurement par la couche iSCSI locale.

3.3 Exigences pour l'utilisation du protocole iSCSI

Pour fonctionner en mode à assistance iSER, les couches iSCSI chez l'initiateur et la cible DOIVENT négocier la clé RDMAExtensions (voir au paragraphe 6.3) à "Oui" sur la connexion de tête. Si la clé RDMAExtensions n'est pas négociée à "Oui", le mode à assistance iSER NE DOIT alors PAS être utilisé. Si la clé RDMAExtensions est négociée à "Oui", mais si l'invocation de la primitive opérationnelle Allocate_Connection_Resources à la couche iSER échoue, la couche iSCSI DOIT faire échouer le processus d'établissement iSCSI ou terminer la connexion comme approprié. Voir les détails au paragraphe 10.1.3.1.

Si la clé RDMAExtensions est négociée à "Oui", la couche iSCSI DOIT satisfaire aux exigences d'utilisation de protocole suivantes du protocole iSER :

1. La couche iSCSI chez l'initiateur DOIT régler ExpDataSN à zéro dans les demandes de fonction de gestion de tâche pour la réallocation d'allégeance pour les commandes de lecture/bidirectionnelles, afin d'amener la cible à envoyer toutes les données de lecture non acquittées.
2. La couche iSCSI chez la cible DOIT toujours retourner l'état SCSI dans une PDU Réponse SCSI séparée pour les commandes de lecture, c'est-à-dire, il NE DOIT PAS y avoir un "collapsus de phase" dans la conclusion d'une commande Read SCSI.
3. Les couches iSCSI chez l'initiateur et la cible DOIVENT prendre en charge les clés comme défini à la Section 6 sur les clés opérationnelles Login/Text. Si elles sont utilisées comme spécifié, on NE DOIT PAS répondre à ces clé par un NonCompris, et la sémantique définie DOIT être respectée pour chaque connexion assistée par iSER.
4. La couche iSCSI chez l'initiateur NE DOIT PAS produire de SNACK pour les PDU.

4. Exigences pour l'interface de couche inférieure

4.1 Interactions avec la couche RCaP

La couche de protocole iSER est mise en couche par dessus une couche RCaP (voir la Figure 1) et les caractéristiques clés suivantes sont supposées être prises en charge par toute couche RcaP :

* La couche RCaP prend en charge toutes les opérations RDMA de base, incluant l'opération Écriture RDMA, l'opération Lire RDMA, et l'opération Send.

* La couche RCaP fournit la livraison fiable, dans l'ordre des messages et le placement direct des données.

* Lorsque la couche iSER initie une opération de lecture RDMA à la suite d'une opération Écriture RDMA sur un flux RCaP, le traitement du message de réponse de lecture RDMA sur le nœud distant ne commencera qu'après que la précédante charge utile de message Écriture RDMA est placée dans la mémoire du nœud distant.

* La couche RCaP encapsule un seul message iSER dans un seul message RCaP du côté de la source des données. La couche RCaP désencapsule le message iSER avant de le livrer à la couche iSER sur le côté collecteur des données.

* Pour une couche RCaP qui prend en charge le message Send avec Invalider (par exemple, iWARP) quand la couche iSER fournit la STag à invalider à distance à la couche RCaP pour un message Send avec Invalider, la couche RCaP utilise cette STag comme STag à invalider dans le message Send avec Invalider.

* La couche RCaP utilise la STag et le décalage étiqueté fournis par la couche iSER pour les messages Écriture RDMA et demande de lecture RDMA.

* Lorsque la couche RCaP livre le contenu d'un message Send RDMA à la couche iSER, la couche RCaP fournit la longueur du message Send RDMA. Cela assure que la couche iSER n'a pas à porter un champ Longueur dans l'en-tête iSER.

* Lorsque la couche RCaP livre le message Send à la couche iSER, elle le notifie à la couche iSER avec le mécanisme fourni sur cette interface.

* Pour une couche RCaP qui prend en charge le message Send avec Invalider (par exemple, iWARP) quand la couche RCaP livre un message Send avec Invalider à la couche iSER, elle passe la valeur de la STag qui a été invalidée.

* La couche RCaP propage tous les états et les indications d'erreur à la couche iSER.

* Pour une couche de transport qui opère en mode flux d'octets comme TCP, la mise en œuvre RCaP prend en charge l'activation du mode RDMA après l'établissement de la connexion et l'échange des paramètres d'établissement en mode de flux d'octets. Pour une couche transport qui fournit une capacité de livraison de message comme [IB], la mise en œuvre RCaP prend en charge l'utilisation directe de la capacité de messagerie par la couche iSCSI pour la phase Établissement après l'établissement de connexion et avant d'activer le mode à assistance iSER. (Dans l'exemple spécifique de InfiniBand [IB], la couche iSCSI utilise les messages IB pour transférer des PDU iSCSI pour la phase Établissement après l'établissement de connexion et avant d'activer le mode à assistance iSER.)

* Chaque fois que la couche iSER termine le flux RCaP, la couche RCaP termine la connexion associée.

4.2 Interactions avec la couche Transport

Après l'établissement de la connexion iSER, la couche RCaP et la couche de transport sous-jacente sont responsables du maintien de la connexion et du rapport à la couche iSER de toute défaillance de connexion.

5. Établissement et terminaison de connexion

5.1 Établissement de connexion iSCSI/iSER

Durant l'établissement de connexion, la couche iSCSI chez l'initiateur est responsable de l'établissement d'une connexion avec la cible. Après l'établissement de la connexion, les couches iSCSI chez l'initiateur et la cible entrent dans la phase Établissement en utilisant les mêmes règles que mentionnées dans la [RFC7143]. La connexion passe à la phase de pleines caractéristiques iSCSI en mode à assistance iSER à la suite d'une négociation d'établissement réussie entre l'initiateur et la cible dans laquelle le mode à assistance iSER est négocié et les ressources de connexion nécessaires pour prendre en charge

RCaP ont été allouées à l'initiateur et à la cible. La même connexion DOIT être utilisée pour la phase d'établissement iSCSI et pour la phase de pleines caractéristiques à assistance iSER qui suit.

Pour une couche transport qui fonctionne en mode flux d'octets comme TCP, la mise en œuvre RCaP prend en charge l'activation du mode RDMA après l'établissement de la connexion et l'échange des paramètres d'établissement en mode flux d'octets. Pour une couche transport qui fournit la capacité de livraison de message comme [IB], la mise en œuvre RCaP prend en charge l'utilisation de la capacité de messagerie par la couche iSCSI directement pour la phase Établissement après l'établissement de la connexion et avant d'activer le mode à assistance iSER.

Le mode à assistance iSER NE DOIT PAS être activé tant qu'il n'est pas négocié sur la connexion de tête durant l'étape LoginOperationalNegotiation de la phase d'établissement iSCSI. Le mode à assistance iSER est négocié en utilisant la clé RDMAExtensions=<valeur booléenne>. L'initiateur et la cible DOIVENT échanger la clé RDMAExtensions avec la valeur réglée à "Oui" pour activer le mode à assistance iSER. Si l'initiateur et la cible échouent à négocier la clé RDMAExtensions réglée à "Oui", la connexion DOIT alors continuer avec la sémantique d'établissement définie dans la [RFC7143]. Si la clé RDMAExtensions n'est pas négociée à Oui, pour certaines mises en œuvre RCaP (comme [IB]) la connexion existante peut devoir être supprimée et une nouvelle connexion peut devoir être établie en mode à capacité TCP. (Pour InfiniBand, cela va exiger une connexion comme dans la [RFC4391].)

Le mode à assistance iSER n'est défini que pour une session normale, et la clé RDMAExtensions NE DOIT PAS être négociée pour une session de découverte. Les sessions de découverte sont toujours conduites en utilisant la couche transport comme décrit dans la [RFC7143].

Il n'est pas exigé d'un nœud à capacité iSER qu'il initie l'échange de clé RDMAExtensions si il préfère le mode iSCSI traditionnel. La clé RDMAExtensions, si elle est offerte, DOIT être envoyée dans la première PDU Réponse d'établissement ou Demande d'établissement disponible dans l'étape LoginOperationalNegotiation. Ceci est dû au fait que la valeur de certains paramètres d'établissement peut dépendre de si le mode à assistance iSER est activé ou non.

Le mode à assistance iSER est un attribut de niveau session. Si l'initiateur et la cible négocient tous deux RDMAExtensions="Oui" sur la connexion de tête d'une session, toutes les connexions suivantes de la même session DOIVENT alors activer le mode à assistance iSER sans avoir à échanger les clés RDMAExtensions durant la phase d'établissement iSCSI. À l'inverse, si l'initiateur et la cible ont tous deux échoué à négocier RDMAExtensions à "Oui" sur la connexion de tête d'une session, la clé RDMAExtensions NE DOIT alors PAS être négociée sur une connexion supplémentaire suivante de la session.

Lorsque la clé RDMAExtensions est négociée à "Oui", les clés HeaderDigest et DataDigest DOIVENT être négociées à "Aucune" sur toutes les connexions iSCSI/iSER qui participent à cette session iSCSI. C'est parce que, pour une connexion iSCSI/iSER, RCaP est chargé de fournir la détection d'erreur qui fait un CRC de 32 bits pour tous les messages iSER. De plus, toutes les données en lecture SCSI sont envoyées en utilisant des messages Écriture RDMA au lieu des PDU SCSI Data-In, et toutes les données Écriture SCSI sollicitées sont envoyées en utilisant des messages Réponse de lecture RDMA au lieu de PDU SCSI Data-Out. Les HeaderDigest et DataDigest qui s'appliquent à des PDU iSCSI ne seraient pas appropriées pour les opérations RDMA Lecture et Écriture utilisées avec iSER.

5.1.1 Comportement de l'initiateur

Si le résultat de la négociation iSCSI est d'activer le mode à assistance iSER, alors du côté de l'initiateur, avant d'envoyer la demande d'établissement avec le bit T (Transit) établi à un et le champ NSG (Prochaine étape) réglé à FullFeaturePhase, la couche iSCSI DEVRAIT demander à la couche iSER d'allouer les ressources de connexion nécessaires pour prendre en charge RCaP en invoquant la primitive opérationnelle Allocate_Connection_Resources. Les ressources de connexion requises sont définies par la mise en œuvre et sortent du domaine d'application de la présente spécification. La couche iSCSI peut invoquer la primitive opérationnelle Notice_Key_Values avant d'invoquer la primitive opérationnelle Allocate_Connection_Resources pour demander à la couche iSER de prendre note des valeurs négociées des clés iSCSI pour la connexion. Les clés spécifiques à passer comme qualificatifs d'entrée dépendent de la mise en œuvre. Elles peuvent inclure, mais ne s'y limitent pas, MaxOutstandingR2T et ErrorRecoveryLevel.

Parmi les ressources de connexion allouées chez l'initiateur il y a la profondeur de file d'attente de lecture RDMA entrante (IRD, *Inbound RDMA Read Queue Depth*). Comme décrit au paragraphe 9.5.1, les R2T sont transformés par la cible en opérations Lire RDMA. IRD limite le nombre maximum de demandes Lecture RDMA simultanément en instance d'entrée par flux RCaP de la cible à l'initiateur. La valeur requise de IRD sort du domaine d'application de la spécification iSER. La couche iSER chez l'initiateur DOIT régler IRD à 1 ou plus si des R2T vont être utilisés dans la connexion. Cependant, la couche iSER chez l'initiateur PEUT régler IRD à zéro sur la base de la configuration de la mise en œuvre ; régler IRD à zéro indique qu'aucun R2T ne va être utilisé sur cette connexion. Initialement, la valeur d'IRD iSER chez l'initiateur DEVRAIT être réglée à la valeur d'IRD de chez l'initiateur et NE DOIT PAS être plus que la valeur d'IRD.

D'un autre côté, la profondeur de file d'attente de lecture RDMA sortante (ORD, *Outbound RDMA Read Queue Depth*) PEUT être réglée à zéro car la couche iSER chez l'initiateur ne produit pas de demandes Lecture RDMA à la cible.

L'échec de l'allocation des ressources de connexion demandées localement résulte en un échec d'établissement, et son traitement est décrit au paragraphe 10.1.3.1.

La couche iSER DOIT retourne un état de succès à la couche iSCSI en réponse à la primitive opérationnelle Allocate_Connection_Resources.

Après que la cible a retourné la réponse Établissement avec le bit T réglé à un et le champ NSG réglé à FullFeaturePhase, et une classe d'état de 0x00 (Succès) la couche iSCSI DOIT invoquer la primitive opérationnelle Enable_Datamover avec les qualificatifs suivants (voir au paragraphe 10.1.4.6 le cas où la classe d'état n'est pas Succès) :

a. Connection_Handle qui identifie la connexion iSCSI,
b. Transport_Connection_Descriptor qui identifie la connexion de transport associée à la Connection_Handle.

La couche iSER NE DOIT envoyer le message iSER Hello comme premier message iSER que si iSERHelloRequired est négocé à "Oui". Voir au paragraphe 5.1.3 les échanges iSER Hello.

Si la couche iSCSI du côté de l'initiateur alloue les ressources de connexion pour prendre en charge RCaP seulement après qu'elle a reçu la PDU Réponse d'établissement finale de la cible, elle peut alors n'être pas capable de traiter le nombre de PDU Type de contrôle iSCSI non attendues (comme déclaré par la clé MaxOutstandingUnexpectedPDU provenant de l'initiateur) qui peuvent être envoyées par la cible avant que les ressources de mémoire tampon soient allouées du côté de l'initiateur. Dans ce cas, la clé iSERHelloRequired DEVRAIT être négociée à "Oui" afin que l'initiateur puisse allouer les ressources de connexion avant d'envoyer le message Hello iSER. Voir les détails au paragraphe 5.1.3.

5.1.2 Comportement de cible

Si le résultat de la négociation iSCSI est d'activer le mode à assistance iSER, akirs du côté de la cible, avant d'envoyer la réponse Établissement avec le bit T (Transit) réglé à un et le champ NSG (prochaine étape) réglé à FullFeaturePhase, la couche iSCSI DOIT demander à la couche iSER d'allouer les ressources nécessaires pour prendre en charge RCaP en invoquant la primitive opérationnelle Allocate_Connection_Resources. Les ressources de connexion requises sont définies par la mise en œuvre et sortent du domaine d'application de la présente spécification. Facultativement, la couche iSCSI peut invoquer la primitive opérationnelle Notice_Key_Values avant d'invoquer la primitive opérationnelle Allocate_Connection_Resources pour demander à la couche iSER de prendre note des valeurs négociées des clés iSCSI pour la connexion. Les clés spécifiqes à passer comme qualificatifs d'entrée dépendent de la mise en œuvre. Elles peuvent inclure, sans s'y limiter, MaxOutstandingR2T et ErrorRecoveryLevel.

Une allocation prématurée des ressources de connexion RCaP peut exposer une cible iSER à une attaque d'épuisement de ressources sur ces ressources via plusieurs connexions iSER qui ne progressent qu'au point auquel la mise en œuvre alloue les ressources de connexion RCaP. La contre mesure à cette attaque est l'authentification de l'initiateur ; la couche iSCSI NE DOIT PAS demander à la couche iSER d'allouer les ressources de connexion nécessaires pour prendre en charge RCaP jusqu'à ce que la couche iSCSI soit suffisamment avancée dans la phase d'établissement iSCSI pour qu'elle soit raisonnablement certaine que le côté de l'homologue n'est pas un agresseur. En particulier, si la phase Établissement inclut une étape SecurityNegotiation, la couche iSCSI DOIT différer l'allocation des ressources de connexion (c'est-à-dire, invoquer la primitive opérationnelle Allocate_Connection_Resources) jusqu'à l'étape LoginOperationalNegotiation ([RFC7143]) afin que l'allocation de ressources se produise après l'achèvement de la phase d'authentification.

Parmi les ressources de connexion allouées à la cible se trouve la profondeur de file d'attente de lecture RDMA sortante (ORD). Comme décrit au paragraphe 9.5.1, les R2T sont transformés par la cible en opérations de lecture RDMA. La ORD limite le nombre maximum de demandes de lecture RDMA simultanément en instance par flux RCaP de la cible à l'initiateur. Initialement, la valeur iSER-ORD à la cible DEVRAIT être réglée à la valeur de ORD de la cible.

D'un autre côté, l'IRD à la cible PEUT être réglée à zéro car la couche iSER à la cible ne s'attend pas à ce que des demandes de lecture RDMA soient produites par l'initiateur.

L'échec d'allocation des ressources de connexion demandées localement résulte en un échec d'établissement, et son traitement est décrit au paragraphe 10.1.3.1.

Si la couche iSER à la cible réussit à allouer les ressources de connexion nécessaires pour la prise en charge de RCaP, les événements suivants DOIVENT se produire dans l'ordre spécifié :

1. La couche iSER DOIT retourner un état de succès à la couche iSCSI en réponse à la primitive opérationnelle Allocate_Connection_Resources.

2. La couche iSCSI DOIT invoquer la primitive opérationnelle Enable_Datamover avec les qualificatis suivants :
   a. Connection_Handle qui identifie la connexion iSCSI,
   b. Transport_Connection_Descriptor qui identifie la connexion de transport associée au Connection_Handle,
   c. le message final de couche transport (par exemple, TCP) contenant la réponse Établissement avec le bit T réglé à un et le champ NSG réglé à FullFeaturePhase.

3. La couche iSER DOIT envoyer la PDU Réponse d'établissement finale dans le mode de transport natif pour conclure la phase d'établissement iSCSI. Si le transport sous-jacent est TCP, la couche iSER DOIT alors envoyer la PDU Réponse d'établissement final en mode de flux d'octets.

4. Après la réception du message Hello iSER de l'initiateur, la couche iSER DOIT répondre avec le message iSER HelloReply à envoyer comme premier message iSER si iSERHelloRequired est négocié à "Oui". Si la couche iSER reçoit un message Hello iSER quand iSERHelloRequired est négocié à "Non", ceci DOIT alors être traité comme une erreur de protocole iSER. Voir au paragraphe 5.1.3 les détails de l'échange Hello iSER.

Note : Dans la séquence ci-dessus, les opérations décrites aux éléments 3 et 4 DOIVENT être effectuées de façon atomique pour les connexions iWARP. Manquer à le faire peut résulter en des conditions de compétition.

5.1.3 Échange de Hello iSER

Si iSERHelloRequired est négocié à "Oui", le premier message iSER envoyé par la couche iSER de l'initiateur à la cible DOIT être le message Hello iSER. Le message Hello iSER est utilisé par la couche iSER chez l'initiateur pour déclarer les paramètres iSER à la cible. Voir au paragraphe 9.3 le format d'en-tête iSER pour le message Hello iSER. À l'inverse, si iSERHelloRequired est négocié à "Non", la couche iSER chez l'initiateur NE DOIT alors PAS envoyer de message Hello iSER.

En réponse au message Hello iSER, la couche iSER à la cible DOIT retourner le message iSER HelloReply comme premier message iSER envoyé par la cible si iSERHelloRequired est négocié à "Oui". Le message HelloReply iSER est utilisé par la couche iSER à la cible pour déclarer les paramètres iSER à l'initiateur. Voir au paragraphe 9.4 sur le format d'en-tête iSER pour le message HelloReply iSER. Si la couche iSER reçoit un message Hello iSER quand iSERHelloRequired est négocié à "Non", ceci DOIT alors être traité comme une erreur de protocole iSER. Voir au paragraphe 10.1.3.4 les détails sur les erreurs de protocole iSER.

Dans le message Hello iSER, la couche iSER chez l'initiateur déclare la valeur d'IRD iSER à la cible.

À réception du message Hello iSER, la couche iSER à la cible DOIT régler la valeur de l'IRD iSER au minimum de la valeur de ORD iSER à la cible et de la valeur de IRD iSER déclarée parl'initiateur. Afin, de libérer les ressources inutilisées, la couche iSER à la cible PEUT ajuster (diminuer) sa valeur de ORD pour correspondre à la valeur de ORD iSER si la valeur de ORD iSER est inférieure à la valeur de ORD à la cible.

Dans le message HelloReply iSER, la couche iSER à la cible déclare la valeur de ORD iSER à l'initiateur.

À réception du message HelloReply iSER, la couche iSER chez l'initiateur PEUT ajuster (diminuer) sa valeur d'IRD pour qu'elle corresponde à la valeur de ORD iSER afin de libérer les ressources inutilisées, si la valeur de ORD iSER déclarée par la cible est inférieure à la valeur de iSER-IRD déclarée par l'initiateur.

C'est un échec de négociation de niveau iSER si les paramètres iSER déclarés dans le message Hello iSER par l'initiateur ne sont pas acceptables à la cible. Cela inclut :
* que la valeur d'iSER-IRD déclarée par l'initiateur soit supérieure à 0, et que la valeur d'iSER-ORD déclarée par la cible soit 0 ;
* que les versions de protocole iSER prises en charge par l'initiateur et par la cible ne se recouvrent pas.

Voir au paragraphe 10.1.3.2 le traitement de la situation d'erreur.

Un initiateur qui se conforme à la [RFC5046] alloue les ressources de connexion avant d'envoyer la demande d'établissement avec le bit T (Transit) réglé à un et le champ NSG (prochaine étape) réglé à FullFeaturePhase. (Pour abréger, on appelle cela une allocationt de connexion "précoce".) La spécification iSER actuelle relâche cette exigence pour permettre à un initiateur d'allouer des ressources de connexion après qu'il a reçu la PDU Réponse d'établissement finale de la cible. (Pour abréger, on appelle cela une allocation de connexion "tardive".) Un initiateur qui emploie l'allocation de connexion "tardive" peut rencontrer des problèmes (par exemple, clôture de connexion RCaP) avec une cible qui envoie des PDU iSCSI inattendues immédiatement après la transition à la phase de pleines caractéristiques, comme permis par la valeur négociée de la clé MaxOutstandingUnexpectedPDUs. La seule façon d'empêcher cette situation est en général

d'utiliser les messages Hello iSER, car ils permettent à l'initiateur d'allouer ses ressources de connexion avant d'envoyer son message Hello iSER. La clé iSERHelloRequired est utilisée par l'initiateur pour déterminer si il traite avec une cible qui prend en charge les échanges Hello iSER. Heureusement, les mises en œuvre connues de cible iSER ne tirent pas pleinement parti du nombre de PDU inattendues permises immédiatement après la transition dans la phase de pleines caractéristiques, permettant donc un biais à l'initiateur qui implique une plus petite quantité de ressources de connexion avant la phase de pleines caractéristiques, comme on l'explique plus en détails ci-dessous.

Dans le résumé qui suit, où on pratique l'allocation de connexion "tardive", un initiateur qui suit la [RFC5046] est appelé un "vieil" initiateur ; autrement, il est appelé un "nouvel" initiateur. De même, une cible qui ne prend pas en charge la clé iSERHelloRequired (et répond avec "NonCompris" lors de la négociation de la clé iSERHelloRequired) est appelée une "vieille" cible ; autrement, on l'appelle une "nouvelle" cible. Noter qu'une "vieille" cible peut quand même prendre en charge les échanges Hello iSER, mais ce fait n'est pas connu de l'initiateur. Une "nouvelle" cible peut aussi répondre par "Non" lors de la négociation de la clé iSERHelloRequired. Dans ce cas, son comportement par rapport à l'allocation de connexion "tardive" est similaire à celui d'une "vieille" cible.

Un "nouvel" initiateur va bien interagir avec une "nouvelle" cible.

Pour un "vieil" initiateur et une "vieille" cible, l'échec de l'initiateur à traiter le nombre de PDU Type de contrôle iSCSI inattendues qui sont envoyées par la cible avant que les ressources de mémoire tampon soient allouées chez l'initiateur peut résulter en l'échec de la session iSER causée par la clôture de la connexion RcaP sous-jacente. Pour la "vieille" cible, il y a une mise en œuvre connue qui envoie une PDU Type de contrôle iSCSI inattendue après l'envoi de la réponse Établissement finale et puis attend un peu avant d'envoyer la suivante. Cela tend à alléger un peu le problème de l'allocation de mémoire tampon chez l'initiateur.

Pour un "nouvel" initiateur et une "vieille" cible, l'échec de l'initiateur à traiter le nombre de PDU Type de contrôle iSCSI inattendues qui sont envoyées par la cible avant que les ressources de mémoire tampon soient allouées chez l'initiateur peut résulter en l'échec de la session iSER causée par la clôture de la connexion RCaP sous-jacente. Un "nouvel" initiateur PEUT choisir de terminer la connexion ; autrement, il DEVRAIT faire une des choses suivantes :

1. Allouer les ressources de connexion avant d'envoyer la PDU Demande d'établissement finale.
2. Allouer une ou plusieurs mémoires tampon pour recevoir des PDU Type de contrôle inattendues de la cible avant d'envoyer la PDU Demande d'établissement finale. Cela réduit la possibilité que les PDU Type de contrôle inattendues causent la fermeture de la connexion RCaP avant que les ressources de connexion aient été allouées.

Pour un "vieil" initiateur et une "nouvelle" cible, si la clé iSERHelloRequired n'est pas négociée, une "nouvelle" cible DOIT quand même répondre avec le message HelloReply iSER quand elle reçoit le message Hello iSER. Si la clé iSERHelloRequired est négociée à "Non" ou "NonCompris", une "nouvelle" cible PEUT choisir de terminer la connexion ; autrement, elle DEVRAIT retarder l'envoi de toute PDU Type de contrôle inattendue jusqu'à ce qu'un des événements suivants se soit produit :

1. une PDU est reçue de l'initiateur après l'envoi de la PDU Réponse d'établissement finale ;
2. une période de temporisation configurable par le système (disons, une seconde) est arrivée à expiration.

5.2 Terminaison de connexion iSCSI/iSER

5.2.1 Terminaison normale de connexion chez l'initiateur

La couche iSCSI chez l'initiateur termine une connexion iSCSI/iSER normalement en invoquant la primitive opérationnelle Send_Control qualifiée avec la PDU Demande de désétablissement. La couche iSER chez l'initiateur DOIT utiliser un message Send pour envoyer la PDU Demande de désétablissement à la cible. Le message SendSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP). Après que la couche iSER chez l'initiateur a reçu le message Send contenant la PDU Réponse de désétablissement de la cible, elle DOIT le notifier à la couche iSCSI en invoquant la primitive opérationnelle Control_Notify qualifiée avec la PDU Réponse de désétablissement.

Après l'achèvement du processus de désétablissement, la couche iSCSI chez la cible est chargée de clore la connexion iSCSI/iSER comme décrit au paragraphe 5.2.2. Après que la couche RCaP chez l'initiateur a rapporté que la connexion a été close, la couche iSER chez l'initiateur DOIT désallouer toutes les ressources de connexion et de tâche (si il en est) associées à la connexion, et invalider les transpositions locales (si il en est) avant de le notifier à la couche iSCSI en invoquant la primitive opérationnelle Connection_Terminate_Notify.

5.2.2 Terminaison normale de connexion chez la cible

À réception du message Send contenant la PDU Demande de désétablissement, la couche iSER à la cible DOIT le notifier à la couche iSCSI de la cible en invoquant la primitive opérationnelle Control_Notify qualifiée avec la PDU Demande de

désétablissement. La couche iSCSI achève le processus de désétablieesment en invoquant la primitive opérationnelle Send_Control qualifiée avec la PDU Réponse de désétablissement. La couche iSER de la cible DOIT utiliser un message Send pour envoyer la PDU Réponse de désétablissement à l'initiateur. Le message SendSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP). Après l'achèvement du processus de désétablissement iSCSI, la couche iSCSI de la cible DOIT demander à la couche iSER de la cible de terminer le flux RCaP en invoquant la primitive opérationnelle Connection_Terminate.

Au titre du processus de terminaison, la couche RCaP DOIT clore la connexion. Lorsque la couche RCaP le notifie à la couche iSER après que le flux RCaP et la connexion associée sont terminés, la couche iSER DOIT désallouer toutes les ressources de connexion et de tâche (si il en est) associées à la connexion, et invalider les transpositions locale et distante (si il en est).

5.2.3 Terminaison sans PDU Demande/Réponse de désétablisement

5.2.3.1 Terminaison de connexion initiée par la couche iSCSI

La primitive opérationnelle Connection_Terminate PEUT être invoquée par la couche iSCSI pour demander à la couche iSER de terminer le flux RCaP sans avoir antérieurement échangé les PDU Demande/Réponse de désétablissement entre les deux nœuds iSCSI/iSER. Au titre du processus de terminaison, la couche RCaP va clore la connexion. Lorsque la couche RCaP le notifie à la couche iSER après que le flux RCaP et la connexion associée sont terminés, la couche iSER DOIT effectuer les actions suivantes.

Si la primitive opérationnelle Connection_Terminate est invoquée par la couche iSCSI de la cible, la couche iSER de la cible DOIT alors désallouer toutes les ressources de connexion et de tâche (si il en est) associées à la connexion, et invalider les transpositions locale et distante (si il en est).

Si la primitive opérationnelle Connection_Terminate est invoquée par la couche iSCSI chez l'initiateur, la couche iSER chez l'initiateur DOIT alors désallouer toutes les ressources de connexion et de tâche (si il en est) associées à la connexion, et invalider les transpositions locales (si il en est).

5.2.3.2 Notification de terminaison de connexion à la couche iSCSI

Si la connexion iSCSI/iSER est terminée sans l'invocation de Connection_Terminate provenant de la couche iSCSI, la couche iSER DOIT notifier à la couche iSCSI que la connexion iSCSI/iSER a été terminée en invoquant la primitive opérationnelle Connection_Terminate_Notify.

Avant d'invoquer Connection_Terminate_Notify, la couche iSER de la cible DOIT désallouer toutes les ressources de connexion et de tâche (si il en est) associées à la connexion, et invalider les transpositions locale et distante (si il en est).

Avant d'invoquer Connection_Terminate_Notify, la couche iSER chez l'initiateur DOIT désallouer toutes les ressources de connexion et de tâche (si il en est) associées à la connexion, et invalider les transpositions locales (si il en est).

Si le nœud iSCSI/iSER distant a initié la clôture de la connexion (par exemple, en envoyant un TCP FIN ou TCP RST) la couche iSER DOIT le notifier à la couche iSCSI après que la couche RCaP a rapporté que la connexion est close en invoquant la primitive opérationnelle Connection_Terminate_Notify.

Un autre exemple de terminaison de connexion sans désétablissement préalable est quand la couche iSCSI chez l'initiateur fait un désétablissement implicite (réinstallation de connexion).

6. Clés de fonctionnement Login/Text

Certaines clés opérationnelles iSCSI login/text ont un usage restrient dans iSER, et des clés supplémentaires sont utilisées pour prendre en charge la fonctionnalité de protocole iSER. Toutes les autres clés définies dans la [RFC7143] et non discutées dans cette section peuvent être utilisées sur les connexions iSCSI/iSER avec la même sémantique.

6.1 HeaderDigest et DataDigest

Non pertinent quand RDMAExtensions=Oui

Les négociations qui résultent en RDMAExtensions=Oui pour une session impliquent HeaderDigest=Aucun et DataDigest=Aucun pour toutes les connexions de cette session et outrepassent les réglages, qu'ils soient par défaut ou configurés.

6.2 MaxRecvDataSegmentLength

Pour une connexion iSCSI appartenant à une session dans laquelle RDMAExtensions=Oui a été négocié sur la connexion de tête de la session, MaxRecvDataSegmentLength n'a pas besoin d'être déclaré dans la phase Établissement, et DOIT être ignoré si il est déclaré. Les clés InitiatorRecvDataSegmentLength (comme décrit au paragraphe 6.5) et TargetRecvDataSegmentLength (comme décrit au paragraphe 6.4) sont plutôt négociées. Les valeurs de la MaxRecvDataSegmentLength locale et distante sont déduites des clés InitiatorRecvDataSegmentLength et TargetRecvDataSegmentLength.

Dans la phase de pleines caractéristiques, l'initiateur DOIT considérer la valeur de sa MaxRecvDataSegmentLength locale (qu'il devrait avoir déclaré à la cible) comme ayant la valeur de InitiatorRecvDataSegmentLength, et la valeur de la MaxRecvDataSegmentLength distante (qui devrait avoir été déclarée par la cible) comme ayant la valeur de TargetRecvDataSegmentLength. De même, la cible DOIT considérer la valeur de sa MaxRecvDataSegmentLength locale (qu'il devrait avoir déclarée à l'initiateur) comme ayant la valeur de TargetRecvDataSegmentLength, et la valeur de la MaxRecvDataSegmentLength distante (qui devrait avoir été déclarée par l'initiateur) comme ayant la valeur de InitiatorRecvDataSegmentLength.

Noter que la RFC 3720 exige que quand une cible reçoit une demande NOP-Out avec une étiquette de tâche d'initiateur valide, elle réponde avec un NOP-In avec la même étiquette de tâche d'initiateur qui était fournie dans la demande NOP-Out. De plus, elle retournes les premiers octets MaxRecvDataSegmentLength de données de Ping fournies par l'initiateur. Comme il n'y a pas de MaxRecvDataSegmentLength commune à l'initiateur et à la cible dans iSER, la longueur des données envoyées avec la demande NOP-Out NE DOIT PAS excéder InitiatorMaxRecvDataSegmentLength.

La clé MaxRecvDataSegmentLength n'est applicable que pour des PDU Type de contrôle iSCSI.

6.3 RDMAExtensions

Utilisation : LO (seulement de tête)
Envoyeurs : initiateur et cible
Portée : SW (de session)
RDMAExtensions=<valeur booléenne>
Non pertinent quand : SessionType=Discovery
Par défaut : Non
Fonction résultante est ET

Cette clé est utilisée par l'initiateur et la cible pour négocier la prise en charge du mode à assistance iSER. Pour activer l'utilisation du mode à assistance iSER, l'initiateur et la cible DOIVENT tous deux échanger RDMAExtensions=Oui. Le mode à assistance iSER NE DOIT PAS être utilisé si l'initiateur ou la cible offre RDMAExtensions=Non.

Un nœud à capacité iSER n'est pas obligé d'initier l'échange de clés RDMAExtensions si il préfère fonctionner dans le mode iSCSI traditionnel. Cependant, si la clé RDMAExtensions doit être négociée, un initiateur DOIT offrir la clé dans la première PDU Demande d'établissement dans l'étape LoginOperationalNegotiation de la connexion de tête, et une cible DOIT offrir la clé dans la première PDU Réponse d'établissement avec laquelle il lui est permis de le faire (c'est-à-dire, la première PDU Réponse d'établissement produite après la première PDU Demande d'établissement avec le bit C réglé à zéro) dans l'étape LoginOperationalNegotiation de la connexion de tête. En réponse à la paire clé=valeur offerte de RDMAExtensions=Oui, un initiateur DOIT répondre dans la prochaine PDU Demande d'établissement avec laquelle il lui est permis de faire ainsi, et une cible DOIT répondre dans la prochaine PDU Réponse d'établissement dans laquelle il lui est permis de le faire.

Négocier d'abord la clé RDMAExtensions permet à un nœud de négocier la valeur optimale pour les autres clés. Certaines clés iSCSI comme MaxBurstLength, MaxOutstandingR2T, ErrorRecoveryLevel, InitialR2T, ImmediateData, etc., peuvent être négociées différemment selon que la connexion est en mode iSCSI traditionnel ou en mode à assistance iSER.

6.4 TargetRecvDataSegmentLength

Utilisation : IO (seulement en initialisation)
Envoyeurs : initiateur et cible
Portée : CO (seulement connexion)

Non pertinent quand : RDMAExtensions=Non
TargetRecvDataSegmentLength=<valeur numérique de 512 à (2**24-1)>
Par défaut : 8192 octets
Fonction résultante : minimum

Cette clé n'est pertinente que pour la connexion iSCSI d'une session iSCSI si RDMAExtensions=Oui a été négocié sur la connexion de tête de la session. Elle est utilisée par l'initiateur et la cible pour négocier la taille maximum des segments de données qu'un initiateur peut envoyer à la cible dans une PDU Type de contrôle iSCSI dans la phase de pleines caractéristiques. Pour des PDU Commande SCSI et Data-Out SCSI contenant des données non sollicitées non immédiates à envoyer par l'initiateur, l'initiateur DOIT envoyer toutes les PDU non finales avec une taille de segment de données de exactement TargetRecvDataSegmentLength chaque fois que les PDU constituent une séquence de données dont la taille est supérieure à TargetRecvDataSegmentLength.

6.5 InitiatorRecvDataSegmentLength

Utilisation : IO (seulement initialisation)
Envoyeurs : initiateur et cible
Portée : CO (seulement connexion)
Non pertinent quand : RDMAExtensions=Non
InitiatorRecvDataSegmentLength=<valeur numérique de 512 à (2**24-1)>
Par défaut : 8192 octets
Fonction résultante : minimum

Cette clé n'est pertinente que pour la connexion iSCSI d'une session iSCSI si RDMAExtensions=Oui a été négocié sur la connexion de tête de la session. Elle est utilisée par l'initiateur et la cible pour négocier la taille maximum du segment de données qu'une cible peut envoyer à l'initiateur dans une PDU Type de contrôle iSCSI dans la phase de pleines caractéristiques.

6.6 OFMarker et IFMarker

Non pertinent quand : RDMAExtensions=Oui
Les négociations résultant en RDMAExtensions=Oui pour une session impliquent OFMarker=Non et IFMarker=Non pour toutes les connexions dans cette session et outrepassent les réglages, par défaut ou configurés.

6.7 MaxOutstandingUnexpectedPDUs

Utilisation : LO (seulement de tête), Déclarative
Envoyeurs : initiateur et cible
Portée : SW (pour la session)
Non pertinent quand : RDMAExtensions=Non
MaxOutstandingUnexpectedPDUs= <valeur numérique de 2 à (2**32-1) | 0>
Par défaut : 0

Cette clé est utilisée par l'initiateur et la cible pour déclarer le nombre maximum de PDU en instance "non attendues" Type de contrôle iSCSI qu'il peut recevoir dans la phase de pleines caractéristiques. Elle est destinée à permettre au côté receveur de déterminer la quantité de ressources de mémoire tampon nécessaires au delà du mécanisme normal de contrôle de flux disponible dans iSCSI. Un initiateur ou cible devrait choisir une valeur telle qu'elle n'imposera pas de contraintes inutiles à la couche iSCSI dans des circonstances normales. La valeur 0 est définie comme indiquant que le déclarant n'a pas de limite pour le nombre maximum de PDU en instance "non attendues" Type de contrôle iSCSI qu'il peut recevoir. Voir aux paragraphes 8.1.1 et 8.1.2 l'usage de cette clé. Noter que les messages iSER Hello et HelloReply ne sont pas des PDU Type de contrôle iSCSI et ne sont pas affectés par cette clé.

Pour l'interopérabilité avec les mises en œuvre fondées sur la [RFC5046], cette clé DEVRAIT être négociée parce que la valeur par défaut de 0 dans la [RFC5046] est problématique pour la plupart des mises en œuvre car elle n'impose pas de limite aux ressources consommables par les PDU inattendues.

6.8 MaxAHSLength

Utilisation : LO (seulement de tête), Déclarative
Envoyeurs : initiateur et cible
Portée : SW (pour la session)

Non pertinent quand : RDMAExtensions=Non
MaxAHSLength=<valeur numérique de 2 à (2**32-1) | 0>
Par défaut : 256

Cette clé est utilisée par l'initiateur et la cible pour déclarer la taille maximum de AHS dans une PDU Type de contrôle iSCSI qu'il peut recevoir dans la phase de pleines caractéristiques. Elle est destinée à permettre au côté receveur de déterminer la quantité de ressources nécessaires pour la mémoire tampon de réception. Un initiateur ou cible devrait choisir une valeur telle qu'elle n'impose pas de contrainte inutile à la couche iSCSI dans des circonstances normales. La valeur 0 est définie comme indiquant que le déclarant n'a pas de limite pour la taille maximum de AHS dans les PDU Type de contrôle iSCSI qu'il peut recevoir.

Pour l'interopérabilité avec les mises en œuvre fondées sur la [RFC5046], un initiateur ou cible PEUT terminer la connexion si il prévoit que MaxAHSLength sera supérieure à 256 et que la clé n'est pas comprise par son homologue.

6.9 TaggedBufferForSolicitedDataOnly

Utilisation : LO (seulement de tête), Déclarative
Envoyeurs : initiateur
Portée : SW (pour la session)
RDMAExtensions=<valeur booléenne>
Non pertinent quand : RDMAExtensions=Non
Par défaut : Non

Cette clé est utilisée par l'initiateur pour déclarer à la cible l'usage du décalage de base en écriture dans l'en-tête iSER d'une PDU Type de contrôle iSCSI. Lorsque réglé à Non, le décalage de base est associé à une mémoire tampon d'entrée/sortie qui contient toutes les données d'écriture, incluant des données sollicitéeset non sollicitées. Lorsque réglé à Oui, le décalage de base est associé à une mémoire tampon d'entrée/sortie qui contient seulement des données sollicitées.

6.10 iSERHelloRequired

Utilisation : LO (seulement de tête), Déclarative
Envoyeurs : initiateur
Portée : SW (pour la session)
RDMAExtensions=<valeur booléenne>
Non pertinent quand : RDMAExtensions=Non
Par défaut : Non

Cette clé n'est pertinente que pour la connexion iSCSI d'une session iSCSI si RDMAExtensions=Oui a été négocié sur la connexion de tête de la session. Elle est utilisée par l'initiateur pour déclarer à la cible si l'échange Hello iSER est exigé. Lorsque réglé à Oui, les couches iSER DOIVENT effectuer l'échange Hello iSER comme décrit au paragraphe 5.1.3. Lorsque réglé à Non, les couches iSER NE DOIVENT PAS effectuer l'échange Hello iSER.

7. Considérations sur les PDU iSCSI

Lorsque une connexion est en mode à assistance iSER, deux types de transferts de message sont permis entre la couche iSCSI (chez l'initiateur) et la couche iSCSI (à la cible). Ce sont les PDU Type de données iSCSI et les PDU Type de contrôle iSCSI, et ces termes sont décrits dans les paragraphes qui suivent.

7.1 PDU Type de données iSCSI

Une PDU Type de données iSCSI se définit comme une PDU iSCSI qui cause le transfert de données, transparent à la couche iSCSI distante, qui a lieu entre les nœuds iSCSI homologues dans la phase de pleines caractéristiques d'une connexion iSCSI/iSER. Une PDU Type de données iSCSI, quand elle est demandée pour la transmission par la couche iSCSI dans le nœud d'envoi, résulte en le transfert des données sans la participation des couches iSCSI aux nœuds envoyeur et receveur. Ceci est dû au fait que la PDU elle-même n'est pas livrée telle qu'elle à la couche iSCSI dans le nœud receveur. Les opérations de transfert des données sont plutôt transformées en les opérations appropriées de RDMA, qui sont traitées par le contrôleur à capacité RDMA. L'ensemble des PDU Type de données iSCSI consiste en des PDU SCSI Data-In et R2T.

Si l'invocation de la primitive opérationnelle par la couche iSCSI pour demander à la couche iSER de traiter une PDU Type de données iSCSI est qualifiée avec Notify_Enable établi, à l'achèvement de l'opération RDMA, la couche iSER à la cible DOIT alors le notifier à la couche iSCSI de la cible en invoquant la primitive opérationnelle Data_Completion_Notify qualifiée avec la ITT et le SN. Il n'y a pas de notification d'achèvement des données chez l'initiateur car les opérations RDMA sont entièrement traitées par le contrôleur à capacité RDMA chez l'initiateur et la couche iSER chez l'initiateur n'est pas impliquée dans le transfert des données associées aux PDU Type de données iSCSI.

Si l'invocation de la primitive opérationnelle par la couche iSCSI pour demander à la couche iSER de traiter une PDU Type de données iSCSI est qualifiée avec Notify_Enable à zéro, à l'achèvement de l'opération RDMA, la couche iSER de la cible NE DOIT alors PAS le notifier à la couche iSCSI de la cible et NE DOIT PAS invoquer la primitive opérationnelle Data_Completion_Notify.

Si une opération associée à une PDU Type de données iSCSI échoue pour une raison quelconque, le contenu des mémoires tampon du collecteur de données associées à l'opération est considéré comme indéterminé.

7.2 PDU Type de contrôle iSCSI

Toute PDU iSCSI qui n'est pas une PDU Type de données iSCSI ni une PDU SCSI Data-Out portant des données sollicitées est définie comme PDU Type de contrôle iSCSI. La couche iSCSI invoque la primitive opérationnelle Send_Control pour demander à la couche iSER de traiter une PDU Type de contrôle iSCSI. Les PDU Type de contrôle iSCSI sont transférées en utilisant les messages Send de RCaP. Précisément, on notera que les PDU SCSI Data-Out qui portent des données non sollicitées sont définies comme des PDU Type de contrôle iSCSI. Voir au paragraphe 7.3.4 le traitement des PDU SCSI Data-Out.

Lorsque la couche iSER reçoit une PDU Type de contrôle iSCSI, elle DOIT le notifier à la couche iSCSI en invoquant la primitive opérationnelle Control_Notify qualifiée avec la PDU Type de contrôle iSCSI.

7.3 PDU iSCSI

Cette section décrit le traitement de chacun des types de PDU iSCSI par la couche iSER. La couche iSCSI demande à la couche iSER de traiter la PDU iSCSI en invoquant la primitive opérationnelle appropriée. Un Connection_Handle DOIT qualifier chacune de ces invocations. De plus, le BHS et le AHS facultatif de la PDU iSCSI comme définis dans la [RFC7143] DOIVENT qualifier chacune des invocations. Le Connection_Handle qualificatif, le BHS, et le AHS ne sont pas mentionnés explicitement dans les paragraphes qui suivent.

7.3.1 Commande SCSI

Type : PDU Type de contrôle

Qualificatifs spécifiques de PDU (pour commande SCSI Écriture ou Bidirectionnelle) : ImmediateDataSize, UnsolicitedDataSize, DataDescriptorOut

Qualificatifs spécifiques de PDU (pour commande SCSI Lecture ou Bidirectionnelle) : DataDescriptorIn

La couche iSER chez l'initiateur DOIT envoyer la commande SCSI dans un message Send à la cible. Le message SendSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP).

Pour une commande SCSI Écriture ou bidirectionnelle, la couche iSCSI chez l'initiateur DOIT invoquer la primitive opérationnelle Send_Control comme suit :

* Si il y a des données immédiates à transférer pour la commande SCSI Écriture ou Bidirectionnelle, le qualificatif ImmediateDataSize DOIT être utilisé pour définir le nombre d'octets de données non sollicitées immédiates à envoyer avec la commande d'écriture ou bidirectionnelle, et le qualificatif DataDescriptorOut DOIT être utilisé pour définir la mémoire tampon d'entrée/sortie de l'initiateur qui contient les données d'écriture SCSI.

* Si il y a des données non sollicitées à transférer pour la commande SCSI Écriture ou Bidirectionnelle, le qualificatif UnsolicitedDataSize DOIT être utilisé pour définir le nombre d'octets de données immédiates et non immédiates non sollicitées pour la commande. La couche iSCSI va produire une ou plusieurs PDU SCSI Data-Out pour les données non sollicitées non immédiates. Voir au paragraphe 7.3.4 les détails sur SCSI Data-Out.

* Si il y a des données sollicitées à transférér pour la commande SCSI Écriture ou Bidirectionnelle, comme c'est indiqué lorsque la longueur attendue de transfert de données dans la PDU Commande SCSI excède la valeur de UnsolicitedDataSize, la couche iSER chez l'initiateur DOIT faire ce qui suit :
  a. elle DOIT allouer une STag Écriture pour la mémoire tampon d'entrée/sortie définie par le qualificatif DataDescriptorOut. DataDescriptorOut décrit la mémoire tampon I/O commençant par les données non sollicitées immédiates (si il en est), suivies par les données non sollicitées non immédiates (si il en est) et les données sollicitées. Lorsque TaggedBufferForSolicitedDataOnly est négocié à Non, le décalage de base est associé à cette mémoire tampon d'entrée/sortie. Lorsque TaggedBufferForSolicitedDataOnly est négocié à Oui, le décalage de base est associé à une mémoire tampon d'entrée/sortie qui contient seulement des données sollicitées.
  b. elle DOIT établir une transposition locale *(Local Mapping)* qui associe l'étiquette de tâche d'initiateur (ITT) à la STag Écriture.
  c. elle DOIT annoncer la STag Écriture et le décalage de base à la cible en les envoyant dans l'en-tête iSER du message iSER (la charge utile du message Send de RCaP) contenant la PDU Commande SCSI Écriture ou Bidirectionnelle. Le message SendSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP). Voir au paragraphe 9.2 le format d'en-tête iSER pour la PDU Type de contrôle iSCSI.

Pour une commande SCSI Lecture ou Bidirectionnelle, la couche iSCSI chez l'initiateur DOIT invoquer la primitive opérationnelle Send_Control qualifiée avec DataDescriptorIn, qui définit la mémoire tampon d'entrée/sortie de l'initiateur pour recevoir les données de lecture SCSI. La couche iSER chez l'initiateur DOIT faire ce qui suit :

a. elle DOIT allouer une STag Lecture pour la mémoire tampon d'entrée/sortie et noter le décalage de base pour cette mémoire tampon d'entrée/sortie.
b. elle DOIT établir une transposition locale qui associe l'étiquette de tâche d'initiateur (ITT) à la STag Lecture.
c. elle DOIT annoncer la STag Lecture et le décalage de base à la cible par leur envoi dans l'en-tête iSER du message iSER (la charge utile du message Send de RCaP) contenant la PDU Lecture ou Bidirectionnelle. Le message SendSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP). Voir au paragraphe 9.2 le format d'en-tête iSER pour la PDU Type de contrôle iSCSI.

Si la quantité de données non sollicitées à transférer dans une commande SCSI excède TargetRecvDataSegmentLength, la couche iSCSI chez l'initiateur DOIT alors segmenter les données en plusieurs PDU Type de contrôle iSCSI, avec la longueur des segments de données dans toutes les PDU générées (excepté la dernière) ayant exactement la taille TargetRecvDataSegmentLength. La longueur du segment de données de la dernière PDU Type de contrôle iSCSI portant les données non sollicitées peut aller jusqu'à TargetRecvDataSegmentLength.

Lorsque la couche iSER de la cible reçoit la commande SCSI, elle DOIT établir une transposition distante qui associe la ITT au ou aux décalages de base et les STag annoncés dans l'en-tête iSER. La STag Écritue est utilisée par la couche iSER de la cible pour traiter le transfert des données associées à la ou aux PDU R2T, comme décrit au paragraphe 7.3.6. La STag Lecture est utilisée pour traiter la ou les PDU Data-In SCSI provenant de la couche iSCSI de la cible comme décrit au paragraphe 7.3.5.

7.3.2 Réponse SCSI

Type : PDU Type de contrôle
Qualificatifs spécifiques de PDU : DataDescriptorStatus

La couche iSCSI de la cible DOIT invoquer la primitive opérationnelle Send_Control qualifiée avec DataDescriptorStatus, qui définit la mémoire tampon contenant les informations de sens et de réponse. La couche iSCSI de la cible DOIT toujours retourner l'état SCSI pour une commande SCSI dans une PDU Réponse SCSI séparée. "Collapsus de phase" NE DOIT PAS être utilisé pour transférer l'état SCSI dans une PDU Data-In SCSI. La couche iSER de la cible envoie la PDU Réponse SCSI conformément aux règles suivantes :

* Si aucune STag n'a été annoncée par l'initiateur dans le message iSER contenant la PDU Commande SCSI, la couche iSER de la cible DOIT alors envoyer un message Send contenant la PDU Réponse SCSI. Le message SendSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP).

* Si l'initiateur a annoncé une STag Lecture dans le message iSER contenant la PDU Commande SCSI, la couche iSER de la cible DOIT alors envoyer un message Send contenant la PDU Réponse SCSI. L'en-tête du message Send DOIT porter la STag Lecture à invalider chez l'initiateur. Le message Send avec Invalider, si il est pris en charge par la couche RCaP (par exemple, iWARP) peut être utilisé pour l'invalidation automatique de la STag.

* Si l'initiateur a annoncé seulement la STag Écriture dans le message iSER contenant la PDU Commande SCSI, la couche iSER à la cible DOIT alors envoyer un message Send contenant la PDU Réponse SCSI. L'en-tête du message

Send DOIT porter la STag Écriture à invalider chez l'initiateur. Le message Send avec Invalider, si il est pris en charge par la couche RCaP (par exemple, iWARP) peut être utilisé pour l'invalidation automatique de STag.

Lorsque la couche iSCSI de la cible invoque la primitive opérationnelle Send_Control pour envoyer la PDU Réponse SCSI, la couche iSER à la cible DOIT invalider la transposition distante avant de transférer la PDU Réponse SCSI à l'initiateur.

À réception d'un message Send contenant la PDU Réponse SCSI de la cible, la couche iSER chez l'initiateur DOIT invalider la ou les STag spécifiées dans l'en-tête. (Si un message Send avec Invalider est pris en charge par la couche RCaP (par exemple, iWARP) et est utilisé pour porter la PDU Réponse SCSI, la couche RCaP chez l'initiateur va invalider la STag. La couche iSER chez l'initiateur DOIT s'assurer que la STag correcte est invalidée. Si les deux STag Lecture et Écriture ont été annoncées précédemment par l'initiateur, la couche iSER chez l'initiateur DOIT alors invalider explicitement la STag Écriture à réception du message Send avec Invalider parce que l'en-tête du message Send avec Invalider peut seulement porter une STag (dans ce cas, la STag Écriture) à invalider.)

La couche iSER chez l'initiateur DOIT s'assurer de l'invalidation de la ou des STag utilisées dans une commande avant de la notifier à la couche iSCSI chez l'initiateur en invoquant la primitive opérationnelle Control_Notify qualifiée avec la réponse SCSI. Cela empêche la possibilité d'utiliser les STag après l'achèvement de la commande ; une telle utilisation causerait la corruption des données.

Lorsque la couche iSER chez l'initiateur reçoit un message Send contenant la PDU Réponse SCSI, elle DEVRAIT invalider la transposition locale. La couche iSER DOIT s'assurer que toutes les STag locales associées à l'ITT sont invalidées avant de le notifier à la couche iSCSI de la PDU Réponse SCSI en invoquant la primitive opérationnelle Control_Notify qualifiée avec la PDU Réponse SCSI.

7.3.3 Demande/réponse de fonction de gestion de tâche

Type : PDU Type de contrôle
Qualificatifs spécifiques de PDU (pour demande TMF) : DataDescriptorOut, DataDescriptorIn

La couche iSER DOIT utiliser un message Send pour envoyer la PDU Demande/réponse de fonction de gestion de tâche. Le message SendSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP).

Pour la demande de fonction de gestion de tâche avec la fonction Réallocation de tâche, la couche iSER chez l'initiateur DOIT faire ce qui suit :

* Elle DOIT utiliser l'ITT comme spécifié dans l'étiquette de tâche référencée provenant de la PDU Demande de fonction de gestion de tâche pour localiset les STag existantes (si il en est) dans les transpositions locales.
* Elle DOIT invalider les STag existantes (si il en est) et les transpositions locales.
* Elle DOIT allouer une STag Lecture pour la mémoire tampon d'entrée/sortie et noter le décalage de base associé à la mémoire tampon d'entrée/sortie comme défini par le qualificatif DataDescriptorIn si l'invocation de la primitive opérationnelle Send_Control est qualifiée avec DataDescriptorIn.
* Elle DOIT allouer une STag Écriture pour la mémoire tampon d'entrée/sortie et noter le décalage de base associé à la mémoire tampon d'entrée/sortie comme défini par le qualificatif DataDescriptorOut si l'invocation de la primitive opérationnelle Send_Control est qualifiée avec DataDescriptorOut.
* Si des STag sont allouées, elle DOIT établir de nouvelles transpositions locales qui associent la ITT aux STag allouées.
* Elle DOIT annoncer les STag et les décalages de base, si ils sont alloués, à la cible dans l'en-tête iSER du message Send qui porte la PDU iSCSI, comme décrit au paragraphe 9.2. Le message SendSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP).

Pour la demande de fonction de gestion de tâche avec la fonction Réallocation de tâche pour une commande SCSI Lecture ou Bidirectionnelle, la couche iSCSI chez l'initiateur DOIT régler ExpDataSN à zéro car le transfert des données et les accusés de réception se produisent de façon transparente pour la couche iSCSI chez l'initiateur. Cela donne de la souplesse à la iSCSI de la cible pour demander la transmission des seules données non acquittées comme spécifié dans la [RFC7143].

Lorsque la couche iSER de la cible reçoit la demande de fonction de gestion de tâche avec la fonction Réallocation de tâche, elle DOIT faire ce qui suit :

* Elle DOIT utiliser l'ITT comme spécifié dans l'étiquette de tâche référencée provenant de la PDU Demande de fonction de gestion de tâche pour localiser les transpositions locales et distantes (si il en est).
* Elle DOIT invalider les STag locales (si il en est) associées à l'ITT.
* Elle DOIT remplacer les décalages de base et les STag annoncées dans la transposition distante par les décalages de base et les STag annoncés dans l'en-tête iSER. La STag Écriture est utilisée dans le traitement des PDU R2T provenant de la couche iSCSI de la cible comme décrit au paragraphe 7.3.6. La STag Lecture est utilisée dans le traitement des PDU Data-In SCSI provenant de la couche iSCSI de la cible comme décrit au paragraphe 7.3.5.

7.3.4 Data-Out SCSI

Type : PDU Type de contrôle
Qualificatifs spécifiques de PDU : DataDescriptorOut

La couche iSCSI chez l'initiateur DOIT invoquer la primitive opérationnelle Send_Control qualifiée avec DataDescriptorOut, qui définit la mémoire tampon d'entrée/sortie de l'initiateur qui contient les données d'écriture non sollicitées SCSI.

Si la quantité de données non sollicitées à transférer comme SCSI Data-Out excède TargetRecvDataSegmentLength, la couche iSCSI chez l'initiateur DOIT alors segmenter les données en plusieurs PDU Type de contrôle iSCSI, où la DataSegmentLength a la valeur de TargetRecvDataSegmentLength dans toutes les PDU générées sauf la dernière. La DataSegmentLength de la dernière PDU Type de contrôle iSCSI portant les données non sollicitées peut faire jusqu'à TargetRecvDataSegmentLength. La couche iSCSI de la cible DOIT effectuer la fonction de réassemblage pour les données non sollicitées.

Pour les données non sollicitées, la couche iSER chez l'initiateur DOIT utiliser un message Send pour envoyer la PDU SCSI Data-Out. Si le bit F est réglé à 1, le message SendSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP).

Noter que pour les données sollicitées, les PDU SCSI Data-Out ne sont pas utilisées car les PDU R2T ne sont pas livrées à la couche iSCSI chez l'initiateur ; les PDU R2T sont plutôt transformées par la couche iSER de la cible en opérations Lire RDMA. (Voir au paragraphe 7.3.6.)

7.3.5 Data-In SCSI

Type : PDU Type de données
Qualificatifs spécifiques de PDU : DataDescriptorIn

Lorsque la couche iSCSI de la cible est prête à retourner les données de lecture SCSI à l'initiateur, elle DOIT invoquer la primitive opérationnelle Put_Data qualifiée avec DataDescriptorIn, qui définit la mémoire tampon SCSI Data-In. Voir au paragraphe 7.1 les exigences générales sur le traitement des PDU Type de données iSCSI. Les PDU Data-In SCSI sont utilisées au transfert de données Lecture SCSI comme décrit au paragraphe 9.5.2.

La couche iSER de la cible DOIT faire ce qui suit pour chaque invocation de la primitive opérationnelle Put_Data :

1. Elle DOIT utiliser la ITT dans la PDU Data-In SCSI pour localiser la STag Lecture distante et le décalage de base dans la transposition distante. La transposition distante a été établie antérieurement par la couche iSER de la cible quand la commande Read SCSI a été reçue de l'initiateur.

2. Elle DOIT générer et envoyer un message Écriture RDMA contenant les données réelles à l'initiateur.
   a. Elle DOIT utiliser la STag Lecture distante comme STag de collecteur de données du message Écriture RDMA.
   b. Elle DOIT ajouter le décalage de mémoire tampon de la PDU Data-In SCSI au décalage de base de la transposition distante comme décalage étiqueté de collecteur de données du message Écriture RDMA.
   c. Elle DOIT utiliser la DataSegmentLength de la PDU Data-In SCSI pour déterminer la quantité de données à envoyer dans le message Écriture RDMA.

3. Elle DOIT associer le DataSN et l'ITT de la PDU Data-In SCSI à l'opération Écriture RDMA. Si l'invocation de la primitive opérationnelle Put_Data était qualifiée avec Notify_Enable établi, alors quand la couche iSER de la cible reçoit un achèvement de la couche RCaP pour le message Écriture RDMA, la couche iSER de la cible DOIT le notifier à la couche iSCSI en invoquant la primitive opérationnelle Data_Completion_Notify qualifiée avec le DataSN et l'ITT. À l'inverse, si l'invocation de la primitive opérationnelle Put_Data était qualifiée avec Notify_Enable à zéro, alors la couche iSER de la cible NE DOIT PAS le notifier à la couche iSCSI à l'achèvement et NE DOIT PAS invoquer la primitive opérationnelle Data_Completion_Notify.

Lorsque le bit A est établi à un dans la PDU Data-In SCSI, la couche iSER de la cible DOIT le notifier à la couche iSCSI de la cible quand le transfert des données est terminé chez l'initiateur. Pour effectuer cette fonction supplémentaire, la couche iSER de la cible peut tirer parti du niveau de récupération d'erreur opérationnel si il a été divulgué préalablement par la couche iSCSI via une invocation antérieure de la primitive opérationnelle Notice_Key_Values. Il y a deux approches qui peuvent être suivies :

1. Si la couche iSER de la cible sait que le niveau de récupération d'erreur opérationnel est 2, ou si la couche iSER de la cible ne sait pas le niveau de récupération d'erreur opérationnel, la couche iSER de la cible DOIT alors produire un message Demande de lecture RDMA de longueur zéro à la suite du message Écriture RDMA. Lorsque la couche iSER de la cible reçoit un achèvement pour le message Demande de lecture RDMA de la couche RCaP, impliquant que le contrôleur à capacité RDMA chez l'initiateur a terminé le traitement du message Écriture RDMA du fait de la sémantique d'ordre d'achèvement de RCaP, la couche iSER de la cible DOIT le notifier à la couche iSCSI de la cible en invoquant la primitive opérationnelle Data_ACK_Notify qualifiée avec l'ITT et le DataSN (voir au paragraphe 3.2.3).

2. Si la couche iSER de la cible sait que le niveau de récupération d'erreur opérationnel est 1, alors la couche iSER de la cible DOIT faire une des choses suivantes :
   a. Elle DOIT le notifier à la couche iSCSI de la cible en invoquant la primitive opérationnelle Data_ACK_Notify qualifiée avec l'ITT et le DataSN (voir au paragraphe 3.2.3) quand elle reçoit l'achèvement local de la couche RCaP pour le message Écriture RDMA. C'est permis parce que les erreurs de résumé ne se produisent pas dans iSER (voir au paragraphe 10.1.4.2) et une erreur de CRC va causer la terminaison de la connexion et la tâche sera terminée de toutes façons. L'achèvement local d'écriture RDMA provenant de la couche RCaP garantit que la couche RCaP ne va pas accéder à nouveau à la mémoire tampon d'entrée/sortie pour transférer les données associées à cette opération Écriture RDMA.
   b. Autrement, elle DOIT utiliser pour traiter l'achèvement du transfert des données chez l'initiateur la même procédure que pour ErrorRecoveryLevel 2.

On devrait noter que la couche iSCSI de la cible ne peut pas établir le bit A à 1 si ErrorRecoveryLevel=0.

L'état SCSI DOIT toujours être retourné dans une PDU Réponse SCSI séparée. Le bit S dans la PDU Data-In SCSI DOIT toujours être réglé à zéro. Il NE DOIT PAS y avoir de "collapsus de phase" dans la PDU Data-In SCSI.

Comme le message Écriture RDMA ne transfère que la portion des données de la PDU Data-In SCSI mais pas les informations de contrôle de l'en-tête, comme ExpCmdSN, si la mise à jour à temps de telles informations est cruciale, la couche iSCSI chez l'initiateur PEUT produire des PDU NOP-Out pour demander à la couche iSCSI de la cible de répondre avec les informations en utilisant les PDU NOP-In.

7.3.6 Prêt au transfert (R2T, *Ready To Transfer*)

Type : PDU Type de données
Qualificatifs spécifiques de PDU : DataDescriptorOut

Afin d'envoyer une PDU R2T, la couche iSCSI de la cible DOIT invoquer la primitive opérationnelle Get_Data qualifiée avec DataDescriptorOut, qui définit la mémoire tampon d'entrée/sortie pour recevoir les données d'écriture SCSI de l'initiateur. Voir au paragraphe 7.1 les exigences générales sur le traitement des PDU Type de données iSCSI.

La couche iSER de la cible DOIT faire ce qui suit pour chaque invocation de la primitive opérationnelle Get_Data :

1. Elle DOIT assurer une STag locale valide pour la mémoire tampon d'entrée/sortie et une transposition locale valide. Cela peut impliquer d'allouer une STag locale valide et d'établir une transposition locale.

2. Elle DOIT utiliser l'ITT dans le R2T pour localiser la STag Écriture distante et le décalage de base dans la transposition distante. La transposition distante a été établie précédemment par la couche iSER à la cible quand le message iSER contenant la STag Écriture annoncée, le décalage de base, et la PDU Commande SCSI pour une écriture SCSI ou une commande bidirectionnelle, a été reçu de l'initiateur.

3. Si la valeur de ORD iSER à la cible est réglée à zéro, la couche iSER de la cible DOIT terminer la connexion et libérer les ressources associées à la connexion (comme décrit au paragraphe 5.2.3) si elle a reçu la PDU R2T de la couche iSCSI à la cible. À la terminaison de la connexion, la couche iSER de la cible DOIT le notifier à la couche iSCSI de la cible en invoquant la primitive opérationnelle Connection_Terminate_Notify.

4. Si la valeur de ORD iSER à la cible est réglée à plus que 0, la couche iSER à la cible DOIT transformer la PDU R2T en un message Demande de lecture RDMA. Tout en transformant la PDU R2T, la couche iSER de la cible DOIT s'assurer que le nombre de messages de demande de lecture RDMA en instance n'excède par la valeur de la ORD iSER. Pour transformer la PDU R2T, la couche iSER à la cible :
   a. DOIT déduire la STag locale et le décalage étiqueté local du DataDescriptorOut qui qualifiait l'invocation de Get_Data.
   b. DOIT utiliser la STag locale comme STag de collecteur de données du message Demande de lecture RDMA.

c. DOIT utiliser le décalage étiqueté local comme décalage étiqueté de collecteur de données du message Demande de lecture RDMA.
d. DOIT utiliser la longueur de transfert de données désirée provenant de la PDU R2T comme taille de message Lecture RDMA du message Demande de lecture RDMA.
e. DOIT utiliser la STag Écriture distante comme STag Souce des données du message Demande de lecture RDMA.
f. DOIT ajouter le décalage de mémoire tampon de la PDU R2T au décalage de base de la transposition distante comme décalage étiqueté de source des données du message Demande de lecture RDMA.

5. Elle DOIT associer le R2TSN et l'ITT provenant de la PDU R2T à l'opération Lire RDMA. Si l'invocation de la primitive opérationnelle Get_Data était qualifiée avec Notify_Enable établi, alors quand la couche iSER à la cible reçoit un achèvement de la couche RCaP pour l'opération Lire RDMA, la couche iSER à la cible DOIT le notifier à la couche iSCSI en invoquant la primitive opérationnelle Data_Completion_Notify qualifiée avec le R2TSN et l'ITT. À l'inverse, si l'invocation de la primitive opérationnelle Get_Data était qualifiée avec Notify_Enable à zéro, la couche iSER à la cible NE DOIT alors PAS notifier l'achèvement à la couche iSCSI et NE DOIT PAS invoquer la primitive opérationnelle Data_Completion_Notify.

Lorsque la couche RCaP chez l'initiateur reçoit un message Demande de lecture RDMA valide, elle va retourner un message de réponse de lecture RDMA contenant les données d'écriture sollicitées à la cible. Lorsque la couche RCaP à la cible reçoit le message de réponse de lecture RDMA de l'initiateur, elle va placer les données sollicitées dans la mémoire tampon d'entrée/sortie référencée par la STag Collecteur de données dans le message Réponse de lecture RDMA.

Comme le message Demande de lecture RDMA de la cible ne transfère pas les informations de contrôle dans la PDU R2T comme ExpCmdSN, si la mise à jour à temps de telles informations est cruciale, la couche iSCSI chez l'initiateur PEUT produire des PDU NOP-Out pour demander à la couche iSCSI de la cible de répondre avec les informations en utilisant des PDU NOP-In.

De même, comme le message de réponse de lecture RDMA de l'initiateur ne transfère que les données mais pas les informations de contrôle qu'on trouve normalement dans la PDU SCSI Data-Out, comme ExpStatSN, si la mise à jour à temps de telles informations est cruciale, la couche iSCSI de la cible PEUT produire des PDU NOP-In pour demander à la couche iSCSI chez l'initiateur de répondre avec les informations en utilisant les PDU NOP-Out.

7.3.7 Message asynchrone

Type : PDU Type de contrôle
Qualificatifs spécifiques de PDU : DataDescriptorSense

La couche iSCSI DOIT invoquer la primitive opérationnelle Send_Control qualifiée avec DataDescriptorSense, qui définit la mémoire tampon contenant les informations de sens et d'événement iSCSI. La couche iSER DOIT utiliser un message Send pour envoyer la PDU Message asynchrone. Le message SendSE devrait être utilisé si il est supporté par la couche RCaP (par exemple, iWARP).

7.3.8 Demande et réponse Text

Type : PDU Type de contrôle
Qualificatifs spécifiques de PDU : DataDescriptorTextOut (pour demande Text), DataDescriptorIn (pour réponse Text)

La couche iSCSI DOIT invoquer la primitive opérationnelle Send_Control qualifiée avec DataDescriptorTextOut (ou DataDescriptorIn), qui définit la mémoire tampon Demande Text (ou Réponse Text). La couche iSER DOIT utiliser les messages Send pour envoyer les PDU Demande Text (ou Réponse Text). Le message SendSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP).

7.3.9 Demande et réponse d'établissement *(Login)*

Durant la négociation d'établissement, la couche iSCSI interagit directement avec la couche transport, et la couche iSER n'est pas impliquée. Voir au paragraphe 5.1 l'établissement de connexion iSCSI/iSER. Si le transport sous-jacent est TCP, Les PDU Demande d'établissement et les PDU Réponse d'établissement sont échangées quand la connexion entre l'initiateur et la cible est encore en mode de flux d'octets.

La couche iSCSI NE DOIT PAS envoyer une PDU Demande d'établissement (ou Réponse d'établissement) durant la phase de pleines caractéristiques. Une PDU Demande d'établissement (ou Réponse d'établissement) si elle est utilisée, DOIT être traitée comme erreur de protocole iSCSI. La couche iSER PEUT rejeter une telle PDU de la couche iSCSI avec un code

d'erreur approprié. Si une PDU Demande d'établissement est reçue par la couche iSCSI de la cible, elle DOIT répondre pau une PDU Rejet avec un code de cause de "erreur de protocole".

7.3.10 Demande et réponse de désétablissement *(Logout)*

Type : PDU Type de contrôle
Qualificatifs spécifiques de PDU : Aucun

La couche iSER DOIT utiliser un message Send pour envoyer la PDU Demande de désétablissement ou Réponse de désétablissement. Le message SendSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP). Les paragraphes 5.2.1 et 5.2.2 décrivent le traitement de la demande de désétablissement et réponse de désétablissement chez l'initiateur et la cible et les interactions entre l'initiateur et la cible pour terminer une connexion.

7.3.11 Demande SNACK

Comme HeaderDigest et DataDigest doivent être négociés à "Aucun", il n'y a pas d'erreur de résumé quand la connexion est en mode à assistance iSER. Aussi, comme RCaP livre tous les messages dans l'ordre d'envoi, il n'y a pas d'erreurs de séquence quand la connexion est en mode à assistance iSER. Donc, la couche iSCSI NE DOIT PAS envoyer de PDU Demande de SNACK. Si une PDU Demande de SNACK est utilisée, elle DOIT être traitée comme une erreur de protocole. La couche iSER PEUT rejeter une telle PDU de la couche iSCSI avec un code d'erreur approprié. Si une PDU Demande de SNACK est reçue par la couche iSCSI de la cible, elle DOIT répondre avec une PDU Rejet avec un code de cause de "erreur de protocole".

7.3.12 Rejet

Type : PDU Type de contrôle
Qualificatifs spécifiques de PDU : DataDescriptorReject

La couche iSCSI DOIT invoquer la primitive opérationnelle Send_Control qualifiée avec DataDescriptorReject, qui définit la mémoire tampon Rejet. La couche iSER DOIT utiliser un message Send pour envoyer la PDU Rejet. Le message SendSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP).

7.3.13 NOP-Out et NOP-In

Type : PDU Type de contrôle
Qualificatifs spécifiques de PDU : DataDescriptorNOPOut (pour NOP-Out), DataDescriptorNOPIn (pour NOP-In)

La couche iSCSI DOIT invoquer la primitive opérationnelle Send_Control qualifiée avec DataDescriptorNOPOut (ou DataDescriptorNOPIn) qui définit la mémoire tampon de donnéees de Ping (ou Retour de Ping). La couche iSER DOIT utiliser les messages Send pour envoyer la PDU NOP-Out (ou NOP-In). Le message SendSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP).

8. Contrôle de flux et gestion de STag

8.1 Contrôle de flux pour messages Send RDMA

Les messages Send dans RCaP sont utilisés par la couche iSER pour transférer des PDU Type de contrôle iSCSI. Chaque message Send dans RCaP consomme une mémoire tampon non étiquetée chez le collecteur de données. Cependant, ni la couche RCaP ni la couche iSER ne fournissent un mécanisme explicite de contrôle de flux pour les messages Send. Donc, la couche iSER DEVRAIT provisionner suffisamment de mémoires tampon non étiquetées pour traiter les messages Send entrants de façon à empêcher l'épuisement de mémoire tampon à la couche RCaP. Si l'épuisement de mémoire tampon se produit, il peut en résulter la terminaison de la connexion.

Une mise en œuvre peut choisir de satisfaire les exigences de mémoire tampon en utilisant un réservoir commun de mémoires tampon partagé sur plusieurs connexions, avec des limites d'usage par connexion et des limites d'usage sur le réservoir de mémoire tampon lui-même. Dans une telle mise en œuvre, excéder la limite d'usage de mémoire tampon pour une connexion ou du réservoir de mémoire tampon lui-même peut déclencher des interventions de la couche iSER pour abonder le réservoir de mémoire tampon et/ou isoler la connexion qui cause problème.

iSER fournit aussi la clé MaxOutstandingUnexpectedPDUs à utiliser par l'initiateur et la cible pour déclarer le nombre maximum de PDU Type de contrôle "inattendues" en instance qu'il peut recevoir. Elle est destinée à permettre au côté

receveur de déterminer la quantité de ressources de mémoire tampon nécessaire au delà du mécanisme normal de contrôle de flux disponible dans iSCSI.

Les ressources de mémoire tampon requises chez l'initiateur et la cible par suite des PDU Type de contrôle envoyées par l'initiateur sont décrites au paragraphe 8.1.1. Les ressources de mémoire tampon requises chez l'initiateur et la cible par suite des PDU Type de contrôle envoyées par la cible sont décrites au paragraphe 8.1.2.

8.1.1 Contrôle de flux pour PDU Type de contrôle de l'initiateur

Les PDU Type de contrôle qui sont envoyées par un initiateur à une cible peuvent être groupées dans les catégories suivantes :

1. Régulée : les PDU Type de contrôle de cette catégorie sont régulées par le mécanisme de fenêtre iSCSI CmdSN et le fanion Immédiat n'est pas établi.
2. Non régulée mais attendue : les PDU Type de contrôle de cette catégorie ne sont pas régulées par le mécanisme de fenêtre iSCSI CmdSN mais sont attendues par la cible.
3. Non régulée et non attendue : les PDU Type de contrôle de cette catégorie ne sont pas régulées par le mécanisme de fenêtre iSCSI CmdSN et ne sont pas "attendues" par la cible.

8.1.1.1 PDU Type de contrôle de l'initiateur dans la catégorie Régulée

Les PDU Type de contrôle qui peuvent être envoyées par l'initiateur dans cette catégorie sont régulées par le mécanisme de fenêtre iSCSI CmdSN, et le fanion Immédiat n'est pas établi.

La capacité de mise en file d'attente requise de la couche iSCSI à la cible est décrite au paragraphe 4.2.2.1 de la [RFC7143]. Pour chacune des PDU Type de contrôle qui peut être envoyée par l'initiateur dans cette catégorie, l'initiateur DOIT provisionner les ressources de mémoire tampon requises pour la PDU Type de contrôle correspondante envoyée comme réponse de la cible. Suit une liste des PDU qui peuvent être envoyées par l'initiateur et des PDU qui peuvent être envoyées par la cible en réponse :

a. Lorsque un initiateur envoie une PDU Commande SCSI, il attend une PDU Réponse SCSI de la cible.
b. Lorsque l'initiateur envoie une PDU Demande de fonction de gestion de tâche, il attend une PDU Réponse de fonction de gestion de tâche de la cible.
c. Lorsque l'initiateur envoie une PDU Demande Text, il attend une PDU Réponse Text de la cible.
d. Lorsque l'initiateur envoie une PDU Demande de désétablissement, il attend une PDU Réponse de désétablissement de la cible.
e. Lorsque l'initiateur envoie une PDU NOP-Out comme demande de ping avec ITT != 0xffffffff et TTT = 0xffffffff, il attend une PDU NOP-In de la cible avec les mêmes ITT et TTT que dans une demande de ping.

La réponse de la cible pour toute PDU mentionnée ici peut autrement être de la forme d'une PDU Rejet envoyée avant que la tâche soit active, comme décrit au paragraphe 7.3 de la [RFC7143].

8.1.1.2 PDU Type de contrôle de l'initiateur dans la catégorie non régulée mais attendue

Pour les PDU Type de contrôle dans la catégorie non régulée mais attendue, la quantité de ressources de mémoire tampon requise à la cible peut être prédéterminée. Voici une liste des PDU de cette catégorie :

a. Les PDU SCSI Data-Out sont utilisées parl'initiateur pour envoyer des données non sollicitées. La quantité de ressources de mémoire tampon requise par la cible peut être déterminée en utilisant FirstBurstLength. Noter que les PDU SCSI Data-Out ne sont pas utilisées pour les données sollicitées car la PDU R2T, qui est utilisée pour la sollicitation est transformée en opérations de lecture RDMA par la couche iSER de la cible. Voir au paragraphe 7.3.4.
b. Une PDU NOP-Out avec TTT != 0xffffffff est envoyée comme réponse de ping par l'initiateur à la PDU NOP-In envoyée comme demande de ping par la cible.

8.1.1.3 PDU Type de contrôle de l'initiateur dans la catégorie non régulée et non attendue

Les PDU de la catégorie non régulée et non attendue sont des PDUqui ont le fanion Immédiat établi. Le nombre de PDU qui sont dans cette catégorie et peuvent être envoyées par un initiateur est contrôlé par la valeur de MaxOutstandingUnexpectedPDUs déclaré par la cible (voir au paragraphe 6.7). Après l'envoi d'une PDU de cette catégorie par l'initiateur, elle est en instance jusqu'à ce qu'elle soit retirée. À tout moment, le nombre de PDU non attendues en instance NE DOIT PAS excéder la valeur de MaxOutstandingUnexpectedPDUs déclaré par la cible.

La cible utilise la valeur de MaxOutstandingUnexpectedPDUs qui est déclarée pour déterminer la quantité de ressources de mémoire tampon requise pour les PDU Type de contrôle de cette catégorie qui peuvent être envoyées par un initiateur. Pour

l'initiateur, pour chacune des PDU Type de contrôle qui peut être envoyée dans cette catégorie, l'initiateur DOIT provisionner les ressources de mémoire tampon qui sont requises pour les PDU Type de contrôle correspondantes qui peuvent être envoyées comme réponse de la cible.

Une PDU en instance de cette catégorie est retirée comme suit. Si le CmdSN de la PDU envoyée par l'initiateur dans cette catégorie est x, la PDU est en instance jusqu'à ce que l'initiateur envoie une PDU Type de contrôle non immédiate sur la même connexion avec CmdSN = y (où y est au moins x) et que la cible réponde avec une PDU Type de contrôle sur une connexion où ExpCmdSN est au moins y+1.

Lorsque le nombre de PDU Type de contrôle non attendues en instance égale MaxOutstandingUnexpectedPDUs, la couche iSCSI chez l'initiateur NE DOIT PAS générer d'autre PDU non attendue, qu'elle aurait autrement généré, même si la PDU non attendue est destinée à une livraison immédiate.

8.1.2 Contrôle de flux pour PDU Type de contrôle de la cible

Les PDU Type de contrôle qui peuvent être envoyées par une cible et sont attendues par l'initiateur sont mentionnées dans la catégorie Régulée. (Voir au paragraphe 8.1.1.1.)

Pour les PDU Type de contrôle qui peuvent être envoyées par une cible et ne sont pas attendues par l'initiateur, leur nombre est contrôlé par le MaxOutstandingUnexpectedPDUs déclaré par l'initiateur (voir au paragraphe 6.7). Après l'envoi d'une PDU de cette catégorie par une cible, elle est en instance juqu'à ce qu'elle soit retirée. À tout moment, le nombre de PDU non attendues en isntance NE DOIT PAS excéder la valeur de MaxOutstandingUnexpectedPDUs déclarée par l'initiateur. L'initiateur utilise la valeur de MaxOutstandingUnexpectedPDUs qui est déclarée pour déterminer la quantité de ressources de mémoire tampon requise pour les PDU Type de contrôle de cette catégorie qui peut être envoyée par une cible. Voici une liste des PDU de cette catégorie et des conditions pour retirer les PDU en instance :

a. Pour une PDU Message asynchrone avec StatSN = x, la PDU est en instance jusqu'à ce que l'initiateur envoie une PDU Type de contrôle avec ExpStatSN réglé au moins à x+1.
b. Pour une PDU Rejet avec StatSN = x, qui est envoyée après qu'une tâche est activée, la PDU est en instance jusqu'à ce que l'initiateur envoie une PDU Type de contrôle avec ExpStatSN réglé au moins à x+1.
c. Pour une PDU NOP-In avec ITT = 0xffffffff et StatSN = x, la PDU est en instance jusqu'à ce que l'initiateur réponde avec une PDU Type de contrôle sur la même connexion où ExpStatSN est au moins x+1. Mais si la PDU NOP-In est envoyée comme une demande de ping avec TTT != 0xffffffff, la PDU peut aussi être retirée lorsque l'initiateur envoie une PDU NOP-Out avec les mêmes ITT et TTT que dans la demande de ping. Noter que quand la cible envoie une PDU NOP-In comme demande de ping, elle doit provisionner une mémoire tampon pour la PDU NOP-Out envoyée comme réponse de ping de l'initiateur.

Lorsque le nombre de PDU Type de contrôle non attendues en instance égale MaxOutstandingUnexpectedPDUs, la couche iSCSI de la cible NE DOIT PAS générer d'autre PDU inattendue, qu'elle aurait générée autrement, même si son intention est d'indiquer une condition d'erreur iSCSI (par exemple, message asynchrone, rejet). Les fins de temporisation de tâche, comme lorsque l'initiateur attend l'achèvement d'une commande ou autres exceptions de niveau connexion et session, vont assurer qu'un comportement de fonctionement correct va résulter dans ce cas en dépit de la non génération de la PDU. Cette règle outrepasse toutes les autres exigences qui par ailleurs exigent qu'une PDU Rejet DOIVE être envoyée.

Note de mise en œuvre : La fin de temporisation et la récupération de tâche SCSI peut être un processus lent et donc DEVRAIT être évité par un provisionne ment approprié des ressources.

Note de mise en œuvre : Pour assurer que l'initiateur a un moyen pour informer la cible que des PDU en instance ont été retirées, la cible devrait réserver la dernière PDU Type de contrôle non attendue permise par la valeur de MaxOutstandingUnexpectedPDUs déclarée par l'initiateur pour l'envoi d'une demande de ping NOP-In avec TTT != 0xffffffff pour permettre à l'initiateur de retourner la réponse de ping NOP-Outavec le ExpStatSN en cours.)

8.2 Contrôle de flux pour ressources Lecture RDMA

Si iSERHelloRequired est négocié à "Oui", le nombre total d'opérations Lire RDMA qui peuvent être actives simultanément sur une connexion iSCSI/iSER dépend alors de la quantité de ressources allouées comme déclaré dans l'échange iSER Hello décrit au paragraphe 5.1.3. Excéder le nombre d'opérations Lire RDMA permises sur une connexion va résulter en la terminaison de la connexion par la couche RCaP. La couche iSER de la cible tient la ORD iSER pour garder trace du nombre maximum de demandes Lecture RDMA qui peuvent être produites par la couche iSER sur un flux RcaP particulier.

Durant l'établissement de connexion (voir au paragraphe 5.1), la IRD iSER est connue chez l'initiateur et la ORD iSER est connue à la cible après que les couches iSER chez l'initiateur et la cible ont respectivement alloué les ressources de connexion nécessaires pour prendre en charge RCaP, comme indiqué par la primitive opérationnelle Allocate_Connection_Resources provenant de la couche iSCSI avant la fin de la phase d'établissement iSCSI. Dans la phase de pleines caractéristiques, si iSERHelloRequired est négocié à "Oui", le premier message envoyé par l'initiateur est alors le message Hello iSER (voir au paragraphe 9.3) qui contient la valeur de la IRD iSER. En réponse au message Hello iSER, la cible envoie le message HelloReply iSER (voir au paragraphe 9.4) qui contient la valeur de ORD iSER. La couche iSER chez l'initiateur et chez la cible PEUT ajuster (diminuer) les ressources associées respectivement à la IRD et la ORD iSER, pour correspondre à la valeur de ORD iSER déclarée dans le message HelloReply. La couche iSER à la cible DOIT contrôler le flux de messages Demande de lecture RDMA afin qu'il n'excède pas la valeur de la ORD iSER de la cible.

Si iSERHelloRequired est négocié à "Non", le nombre maximum d'opérations Lire RDMA qui peuvent être actives est alors négocié via d'autres moyens qui sortent du domaine d'application du présent document. Par exemple, dans InfiniBand, l'établissement de connexion iSER utilise des datagrammes de gestion (MAD) de gestionnaire de connexion InfiniBand, avec des informations iSER supplémentaires échangées dans les données privées.

8.3 Gestion de STag

Une STag est un identifiant de mémoire tampon étiquetée utilisée dans une opération RDMA. Si les STag sont exposées sur le réseau en étant annoncées dans l'en-tête iSER ou déclarées dans l'en-tête d'un message RCaP, l'allocation et l'invalidation subséquente des STag sont comme spécifié dans le présent document.

8.3.1 Allocation des STag

Lorsque la couche iSCSI chez l'initiateur invoque la primitive opérationnelle Send_Control pour demander à la couche iSER chez l'initiateur de traiter une commande SCSI, zéro, une, ou deux STag peuvent être allouées par la couche iSER. Voir les détails au paragraphe 7.3.1. Le nombre de STag allouées dépend de si la commande est unidirectionnelle ou bidirectionnelle et si le transfert de données en écriture sollicitées ou non sollicitées est impliqué.

Lorsque la couche iSCSI chez l'initiateur invoque la primitive opérationnelle Send_Control pour demander à la couche iSER chez l'initiateur de traiter une demande de fonction de gestion de tâche avec la fonction Réallocation de tâche, en plus d'allouer zéro, une, ou deux STag, la couche iSER DOIT invalider les STag existantes (si il en est) associées à l'ITT. Voir les détails au paragraphe 7.3.3.

La couche iSER de la cible alloue une STag Collecteur de données locale quand la couche iSCSI de la cible invoque la primitive opérationnelle Get_Data pour demander à la couche iSER de traiter une PDU R2T. Voir les détails au paragraphe 7.3.6.

8.3.2 Invalidation des STag

L'invalidation des STag chez l'initiateur à l'achèvement d'une commande unidirectionnelle ou bidirectionnelle quand la PDU Réponse SCSI associée est envoyée par la cible est décrite au paragraphe 7.3.2.

Lorsque une commande unidirectionnelle ou bidirectionnelle se conclut sans que la PDU associée Réponse SCSI soit envoyée par la cible, la couche iSCSI chez l'initiateur DOIT demander à la couche iSER chez l'initiateur d'invalider les STag en invoquant la primitive opérationnelle Deallocate_Task_Resources qualifiée avec l'ITT. En réponse, la couche iSER chez l'initiateur DOIT localiser les STag (si il en est) dans la transposition locale. La couche iSER chez l'initiateur DOIT invalider les STag (si il en est) et la transposition locale.

Pour une opération Lire RDMA utilisée pour réaliser un transfert de données d'écriture SCSI, la couche iSER de la cible DEVRAIT invalider la STag Collecteur de données à la conclusion de l'opération Lire RDMA en référençant la STag Collecteur de données (pour permettre la réutilisation immédiate des ressources de mémoire tampon).

Pour une opération Écriture RDMA utilisée pour réaliser un transfert de données de lecture SCSI, la STag Souce des données de la cible n'est pas déclarée à l'initiateur et n'est pas exposée sur le réseau. L'invalidation de la STag n'est donc pas spécifiée.

Lorsque une commande unidirectionnelle ou bidirectionnelle se conclut sans que soit envoyée par la cible la PDU Réponse SCSI associée, la couche iSCSI de la cible DOIT demander à la couche iSER de la cible d'invalider les STag en invoquant la primitive opérationnelle Deallocate_Task_Resources qualifiée avec l'ITT. En réponse, la couche iSER de la cible DOIT localiser les STag locales (si il en est) dans la transposition locale. La couche iSER de la cible DOIT invalider les STag locales (si il en est) et la transposition locale.

9. Transfert de commandes et données iSER

Pour les PDU Type de données iSCSI (voir au paragraphe 7.1) la couche iSER utilise les opérations Lire et Écriture RDMA pour transférer les données sollicitées. Pour des PDU Type de contrôle iSCSI (voir au paragraphe 7.2) la couche iSER utilise les messages Send de RCaP.

9.1 Format d'en-tête iSER

Un en-tête iSER DOIT être présent dans chaque message Send de RCaP. L'en-tête iSER est situé dans les 28 premiers octets de la charge utile du message Send de RCaP, comme le montre la Figure 2.

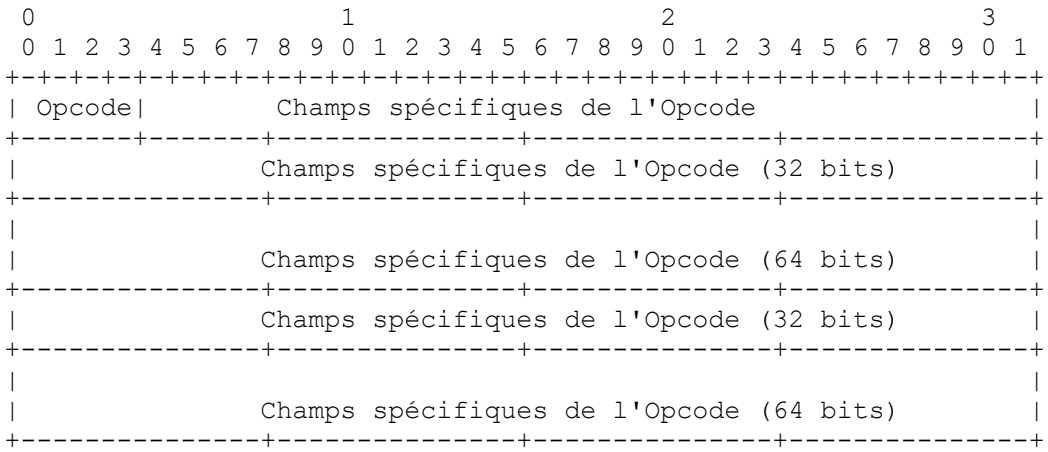


Figure 2 : Format d'en-tête iSER

Opcode (Code d'opération) : 4 bits. Le champ Opcode identifie le type du message iSER :
0001b = PDU Type de contrôle iSCSI
0010b = message Hello iSER
0011b = message HelloReply iSER
Aucun autre Opcode n'est alloué.

9.2 Format d'en-tête iSER pour PDU Type de contrôle iSCSI

La couche iSER utilise les messages Send de RCaP pour transférer des PDU Type de contrôle iSCSI (voir au paragraphe 7.2). La charge utile du message de chaque message Send de RCaP utilisé pour transférer un message iSER contient un en-tête iSER suivi par une PDU Type de contrôle iSCSI.
L'en-tête iSER dans un message Send de RCaP portant une PDU Type de contrôle iSCSI DOIT avoir le format décrit à la Figure 3.

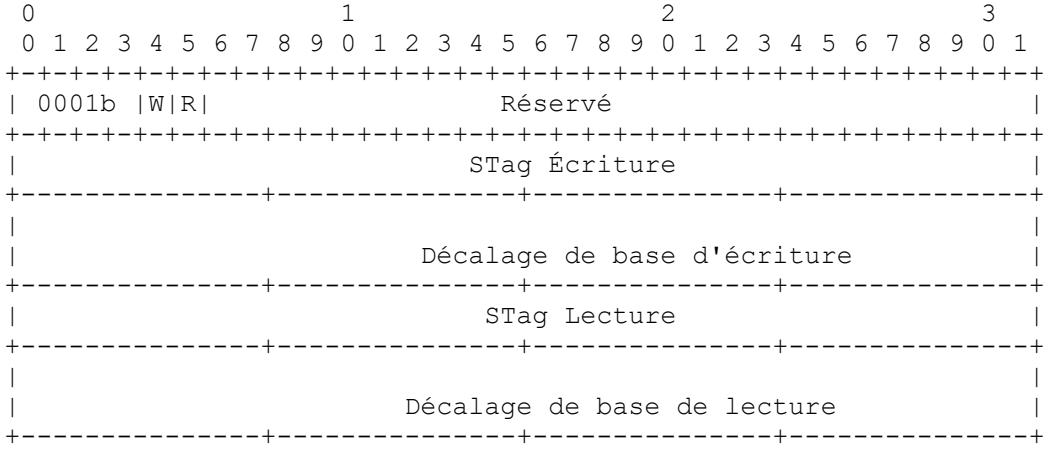


Figure 3 : Format d'en-tête iSER pour PDU Type de contrôle iSCSI

Fanion W (STag Écriture valide) : 1 bit
Ce fanion indique la validité du champ STag Écriture et du champ décalage de base d'écriture dans l'en-tête iSER. Si il est réglé à un, le champ STag Écriture et le champ Décalage de base d'écriture dans cet en-tête iSER sont valides. Si il est réglé à zéro, les champs STag Écriture et Décalage de base d'écriture dans cet en-tête iSER DOIVENT être ignorés par le receveur. Le fanion STag d'écriture valide est réglé à un quand il y a des données sollicitées à transférer pour une commande SCSI Écritue ou Bidirectionnelle, ou quand il y a des données non sollicitées non immédiates et des données sollicitées à transférer pour la tâche référencée spécifiée dans une demmande de fonction de gestion de tâche avec la fonction Réallocation de tâche.

Fanion R (STag de Lecture valide) : 1 bit.
Ce fanion indique la validité du champ STag Lecture et du champ Décalage de base de lecture de l'en-tête iSER. Si il est réglé à un, le champ STag Lecture et champ Décalage de base de lecture dans cet en-tête iSER sont valides. Si il est réglé à zéro, le champ STag Lecture et le champ Décalage de base de lecture dans cet en-tête iSER DOIVENT être ignorés du receveur. Le fanion STag Lecture valide est réglé à un pour une commande SCSI Lecture ou Bidirectionnel, ou une demande de fonction de gestion de tâche avec la fonction Réallocation de tâche.

STag Écriture – étiquette de pilotage d'écriture : 32 bits
Ce champ contient la STag Écriture quand le fanion STag Écriture valide est réglé à un. Pour une commande SCSI Écriture ou Bidirectionnelle, la STag Écriture est utilisée pour annoncer la mémoire tampon d'entrée/sortie de l'initiateur qui contient les données sollicitées. Pour une demande de fonction de gestion de tâche avec une fonction Réallocation de tâche, la STag Écriture est utilisée pour annoncer la mémoire tampon d'entrée/sortie de l'initiateur qui contient les données non sollicitées et les données sollicitées non immédiates. Cette STag Écriture est utilisée comme STag Source des données dans l'opération Lire RDMA résultante. Lorsque le fanion STag Écriture valide est à zéro, ce champ DOIT être réglé à zero et ignoré à réception.

Décalage de base d'écriture : 64 bits
Ce champ contient le décalage de base d'écriture associé à la mémoire tampon d'entrée/sortie pour la commande Écriture SCSI quand le fanion STag d'écriture valide est réglé à un. Lorsque le fanion STag d'écriture valide est à zéro, ce champ DOIT être réglé à zéro et ignoré à réception.

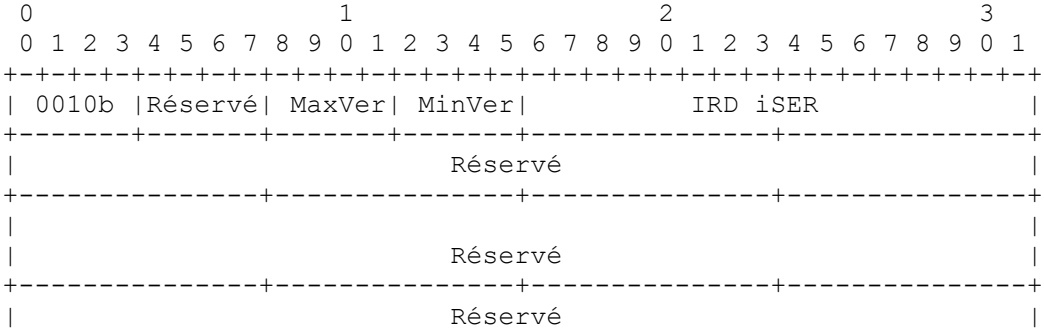
STag Lecture – étiquette de pilotage de lecture : 32 bits
Ce champ contient la STag Lecture quand le fanion STag de lecture valide est réglé à un. La STag Lecture est utilisée pour annoncer la mémoire tampon d'entrée/sortie de lecture de l'initiateur d'une commande SCSI Lecture ou Bidirectionnelle, ou une demande de fonction de gestion de tâche avec la fonction Réallocation de tâche. Cette STag Lecture est utilisée comme STag Collecteur de données dans l'opération Écriture RDMA résultante. Lorsque le fanion STag Lecture valide est à zéro, ce champ DOIT être réglé à zero et ignoré à réception.

Décalage de base de lecture : 64 bits
Ce champ contient le décalage de base de lecture associé à la mémoire tampon d'entrée/sortie pour la commande SCSI Lecture quand le fanion STag Lecture valide est réglé à un. Lorsque le fanion STag Lecture valide est à zéro, ce champ DOIT être réglé à zero et ignoré à réception.

Réservé : Les champs réservés DOIVENT être réglés à zero à l'émission et DOIVENT être ignorés à réception.

9.3 Format d'en-tête iSER pour le message Hello iSER

Un message Hello iSER DOIT seulement contenir l'en-tête iSER, qui DOIT avoir le format décrit à la Figure 4. Si iSERHelloRequired est négocié à "Oui", le message Hello iSER est alors le premier message iSER envoyé sur le flux RCaP depuis la couche iSER chez l'initiateur à la couche iSER à la cible.



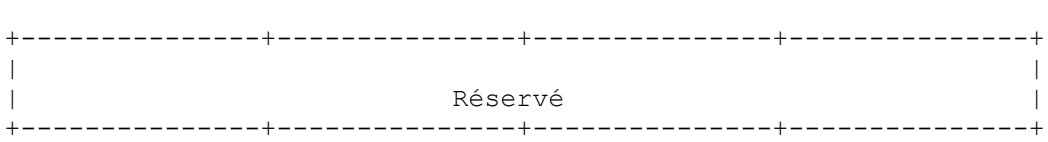


Figure 4 : Format d'en-tête iSER pour message Hello iSER

MaxVer - Version maximum : 4 bits
Ce champ spécifie la version maximum de protocole iSER supportée. Il DOIT être réglé à 10 pour indiquer la version de la spécification décrite dans le présent document.

MinVer - Version minimum : 4 bits
Ce champ spécifie la version minimum de protocole iSER supportée. Il DOIT être réglé à 10 pour indiquer la version de la spécification décrite dans le présent document.

IRD iSER : 16 bits. Ce champ contient la valeur de l'IRD iSER chez l'initiateur.

Réservé : Les champs réservés DOIVENT être réglés à zero à l'émission et DOIVENT être ignorés à réception.

9.4 Format d'en-tête iSER pour message iSER HelloReply

Un message HelloReply iSER DOIT seulement contenir l'en-tête iSER, qui DOIT avoir le format décrit dans la Figure 5. Si iSERHelloRequired est négocié à "Oui", le message HelloReply iSER est alors le premier message iSER envoyé sur le flux RCaP de la couche iSER de la cible à la couche iSER chez l'initiateur.

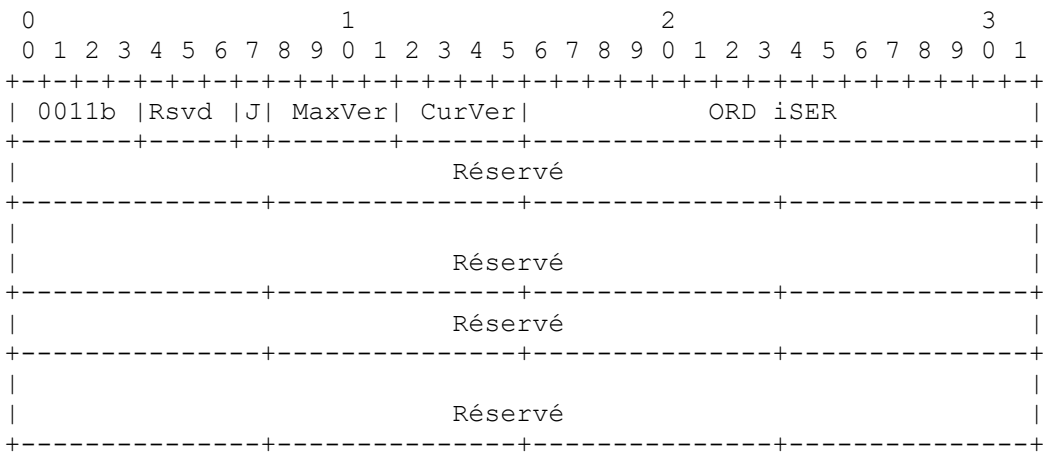


Figure 5 : Format d'en-tête iSER pour message HelloReply iSER

J – Fanion Rejet : 1 bit. Ce fanion indique si la cible rejette cette connexion. Réglé à un, la cible rejette la connexion.

MaxVer – Version maximum : 4 bits
Ce champ spécifie la version maximum de protocole iSER supportée. Il DOIT être réglé à 10 pour indiquer la version de la spécification décrite dans le présent document.

MinVer - Version minimum : 4 bits
Ce champ spécifie la version minimum de protocole iSER supportée. Il DOIT être réglé à 10 pour indiquer la version de la spécification décrite dans le présent document.

ORD iSER : 16 bits. Ce champ contient la valeur de la ORD iSER à la cible.

Réservé : Les champs réservés DOIVENT être réglés à zero à l'émission et DOIVENT être ignorés à réception.

9.5 Opérations de transfert de données SCSI

La couche iSER chez l'initiateur et la couche iSER à la cible traitent chacune les opérations SCSI Écriture, SCSI Lecture, et Bidirectionnelle comme décrit ci-dessous.

9.5.1 Opération Écriture SCSI *(Write)*

La couche iSCSI chez l'initiateur DOIT invoquer la primitive opérationnelle Send_Control pour demander à la couche iSER chez l'initiateur d'envoyer la commande Écriture SCSI. La couche iSER chez l'initiateur DOIT demander à la couche RCaP de transmettre un message Send dont la charge utile consiste en l'en-tête iSER suivi par la PDU Commande SCSI et des données immédiates (si il en est). Le message SendSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP). Si il y a des données sollicitées, la couche iSER DOIT annoncer la STag Écriture et le décalage de base dans l'en-tête iSER du message Send, comme décrit au paragraphe 9.2. À réception du message Send, la couche iSER de la cible DOIT le notifier à la couche iSCSI de la cible en invoquant la primitive opérationnelle Control_Notify qualifiée avec la PDU Commande SCSI. Voir au paragraphe 7.3.1 les détails sur le traitement de la commande Write SCSI.

Pour les données non sollicitées non immédiates, la couche iSCSI chez l'initiateur DOIT invoquer une primitive opérationnelle Send_Control qualifiée avec la PDU SCSI Data-Out. À réception de chaque message Send contenant les données non sollicitées non immédiates, la couche iSER de la cible DOIT le notifier à la couche iSCSI de la cible en invoquant la primitive opérationnelle Control_Notify qualifiée avec la PDU SCSI Data-Out. Voir au paragraphe 7.3.4 les détails sur le traitement de la PDU SCSI Data-Out.

Pour les données sollicitées, quand la couche iSCSI de la cible a une mémoire tampon d'entrée/sortie disponible, elle DOIT invoquer la primitive opérationnelle Get_Data qualifiée avec la PDU R2T. Voir au paragraphe 7.3.6 les détails du traitement de la PDU R2T.

Lorsque le transfert des données associées à cette opération d'écriture SCSI est achevé, la couche iSCSI à la cible DOIT invoquer la primitive opérationnelle Send_Control quand elle est prête à envoyer la PDU Réponse SCSI. À réception d'un message Send contenant la PDU Réponse SCSI, la couche iSER chez l'initiateur DOIT le notifier à la couche iSCSI chez l'initiateur en invoquant la primitive opérationnelle Control_Notify qualifiée avec la PDU Réponse SCSI. Voir au paragraphe 7.3.2 les détails du traitement de la PDU Réponse SCSI.

9.5.2 Opération Lire SCSI *(Read)*

La couche iSCSI chez l'initiateur DOIT invoquer la primitive opérationnelle Send_Control pour demander à la couche iSER chez l'initiateur d'envoyer la commande Lire SCSI. La couche iSER chez l'initiateur DOIT demander à la couche RCaP de transmettre un message Send avec une charge utile de message consistant en l'en-tête iSER suivi par la PDU Commande SCSI. Le message SendSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP). La couche iSER chez l'initiateur DOIT anoncer la STag Lecture et le décalage de base dans l'en-tête iSER du message Send, comme décrit au paragraphe 9.2. À réception du message Send, la couche iSER de la cible DOIT le notifier à la couche iSCSI de la cible en invoquant la primitive opérationnelle Control_Notify qualifiée avec la PDU Commande SCSI. Voir au paragraphe 7.3.1 les détails du traitement de la commande Lecture SCSI.

Lorsque les données SCSI demandées sont disponibles dans la mémoire tampon d'entrée/sortie, la couche iSCSI de la cible DOIT invoquer la primitive opérationnelle Put_Data qualifiée avec la PDU Data-In SCSI. Voir au paragraphe 7.3.5 les détails sur le traitement de la PDU Data-In SCSI.

Lorsque le transfert des données associées à cette opération Lire SCSI est achevé, la couche iSCSI de la cible DOIT invoquer la primitive opérationnelle Send_Control quand elle est prête à envoyer la PDU Réponse SCSI. Le message SendInvSE devrait être utilisé si il est pris en charge par la couche RCaP (par exemple, iWARP). À réception du message Send contenant la PDU Réponse SCSI, la couche iSER chez l'initiateur DOIT le notifier à la couche iSCSI chez l'initiateur en invoquant la primitive opérationnelle Control_Notify qualifiée avec la PDU Réponse SCSI. Voir au paragraphe 7.3.2 les détails du traitement de la PDU Réponse SCSI.

9.5.3 Opération bidirectionnelle

L'initiateur et la cible traitent les portions SCSI Écriture et SCSI Lecture de cette opération bidirectionnelle de la même façon que décrit respectivement aux paragraphes 9.5.1 et 9.5.2.

10. Traitement et récupération d'erreur iSER

RCaP fournit à la couche iSER une livraison fiable dans l'ordre. Donc, les besoins de gestion d'erreur d'une connexion assistée par iSER sont assez différents de ceux d'une connexion iSCSI traditionnelle.

10.1 Traitement d'erreur

Le traitement d'erreur iSER est décrit dans les paragraphes qui suivent, classés selon les sources d'erreurs :
1. générées à la couche transport (par exemple, TCP),
2. générées à la couche RcaP,
3. générées à la couche iSER,
4. générées à la couche iSCSI.

10.1.1 Erreurs dans la couche Transport

Si la couche transport est TCP, les paquets TCP avec des erreurs détectées sont abandonnés en silence par la couche TCP et résultent en une retransmission à la couche TCP. Cela n'a pas d'impact sur la couche iSER. Cependant, une perte de connexion (par exemple, défaillance de liaison) et la terminaison inattendue (par exemple, clôture TCP en douceur ou clôture anormale sans échanges de désétablissement iSCSI) à la couche transport vont aussi causer la terminaison de la connexion iSCSI/iSER.

10.1.1.1 Défaillance dans la couche Transport avant l'activation du mode RCaP

Si la connexion est perdue ou terminée avant que la couche iSCSI invoque la primitive opérationnelle Allocate_Connection_Resources, le processus d'établissement est terminé et aucune autre action n'est requise.

Si la connexion est perdue ou terminée après que la couche iSCSI a invoqué la primitive opérationnelle Allocate_Connection_Resources, la couche iSCSI DOIT alors demander à la couche iSER de désallouer toutes les ressources de connexion en invoquant la primitive opérationnelle Deallocate_Connection_Resources.

10.1.1.2 Défaillance dans la couche Transport après l'activation du mode RCaP

Si la connexion est perdue ou terminée après que la couche iSCSI a invoqué la primitive opérationnelle Enable_Datamover, la couche iSER DOIT notifier à la couche iSCSI la perte de la connexion en invoquant la primitive opérationnelle Connection_Terminate_Notify. Avant d'invoquer la primitive opérationnelle Connection_Terminate_Notify, la couche iSER DOIT effectuer les actions décrites au paragraphe 5.2.3.2.

10.1.2 Erreurs dans la couche RCaP

La couche RCaP n'incorpore pas d'opérations de récupération d'erreur. Si des erreurs sont détectées à la couche RCaP, la couche RCaP va terminer le flux RCaP et la connexion associée.

10.1.2.1 Erreurs détectées dans la couche locale RCaP

Si une erreur se rencontre à la couche RcaP locale, la couche RCaP PEUT envoyer un message Send à l'homologue distant pour rapporter l'erreur si possible. (Pour iWARP, voir dans la [RFC5040] la liste des erreurs où un message Terminer est envoyé.) La couche RCaP est chargée de terminer la connexion. Après que la couche RCaP a notifié à la couche iSER que la connexion est terminée, la couche iSER DOIT le notifier à la couche iSCSI en invoquant la primitive opérationnelle Connection_Terminate_Notify. Avanr d'invoquer la primitive opérationnelle Connection_Terminate_Notify, la couche iSER DOIT effectuer les actions décrites au paragraphe 5.2.3.2.

10.1.2.2 Erreurs détectées dans la couche RCaP chez l'homologue distant

Si une erreur se rencontre à la couche RCaP chez l'homologue distant, la couche RCaP chez l'homologue distant peut envoyer un message Send pour rapporter l'erreur, si possible. Si elle est incapable d'envoyer un message Send, la connexion se termine. Ceci est traité de la même façon qu'une défaillance de la couche transport après l'activation de RDMA, comme décrit au paragraphe 10.1.1.2.

Si une erreur se rencontre à la couche RCaP chez l'homologue distant et si elle est capable d'envoyer un message Send, la couche RCaP chez l'homologue distant est responsable de la terminaison de la connexion. Après que la couche RcaP locale a notifié à la couche iSER que la connexion est terminée, la couche iSER DOIT le notifier à la couche iSCSI en invoquant la primitive opérationnelle Connection_Terminate_Notify. Avant d'invoquer la primitive opérationnelle Connection_Terminate_Notify, la couche iSER DOIT effectuer les actions décrites au paragraphe 5.2.3.2.

10.1.3 Erreurs dans la couche iSER

Le traitement d'erreur dû aux erreurs à la couche iSER est décrit dans les paragraphes qui suivent.

10.1.3.1 Ressources de connexion insuffisantes pour prendre en charge RCaP à l'établissement de la connexion

Après que la couche iSCSI chez l'initiateur a invoqué la primitive opérationnelle Allocate_Connection_Resources durant la phase de négociation d'établissement iSCSI, si la couche iSER chez l'initiateur échoue à allouer les ressources de connexion nécessaires pour prendre en charge RCaP, elle DOIT retourner un état d'échec à la couche iSCSI chez l'initiateur. La couche iSCSI chez l'initiateur DOIT terminer la connexion comme décrit au paragraphe 5.2.3.1.

Après que la couche iSCSI de la cible a invoqué la primitive opérationnelle Allocate_Connection_Resources durant la phase de négociation d'établissement iSCSI, si la couche iSER de la cible échoue à allouer les ressources de connexion nécessaires pour prendre en charge RCaP, elle DOIT retourner un état d'échec à la couche iSCSI de la cible. La couche iSCSI de la cible DOIT envoyer une réponse Établissement avec une classe d'état de 0x03 (Erreur de la cible) et un code d'état de 0x02 (Plus de ressources). Les couches iSCSI chez l'initiateur et la cible DOIVENT terminer la connexion comme décrit au paragraphe 5.2.3.1.

10.1.3.2 Échecs de négociation iSER

Si iSERHelloRequired est négocié à "Oui" et si les paramètres relatifs à RCaP ou iSER déclarés par l'initiateur dans le message Hello iSER sont inacceptables à la couche iSER de la cible, la couche iSER de la cible DOIT établir le fanion Rejet (J) comme décrit au paragraphe 9.4, dans le message HelloReply iSER. Les cas suivants sont ceux où la couche iSER DOIT établir le fanion Rejet à 1 dans le message HelloReply :

* la valeur d'IRD iSER déclarée par l'initiateur est supérieure à 0, et la valeur de ORD iSER déclarée par la cible est 0,
* les versions de protocole iSER prises en charge par l'initiateur et la cible ne se recouvent pas.

Après avoir demandé à la couche RCaP d'envoyer le message HelloReply iSER, le traitement de la situation d'erreur est le même que pour les erreurs de format iSER comme décrit au paragraphe 10.1.3.3.

10.1.3.3 Erreurs de format iSER

Les types d'erreurs suivants dans un en-tête iSER sont considérés comme des erreurs de format :

* contenu illégal de tout champ d'en-tête iSER,
* contenu de champ incohérent dans un en-tête iSER,
* erreur de longueur d'un message iSER Hello ou HelloReply (voir les paragraphes 9.3 et 9.4)

Lorsque une erreur de format est détectée, les événements suivants DOIVENT se produire dans l'ordre spécifié :

1. La couche iSER DOIT demander à la couche RCaP de terminer le flux RCaP. La couche RCaP DOIT terminer la connexion associée.
2. La couche iSER DOIT notifier à la couche iSCSI la terminaison de la connexion en invoquant la primitive opérationnelle Connection_Terminate_Notify. Avant d'invoquer la primitive opérationnelle Connection_Terminate_Notify, la couche iSER DOIT effectuer les actions décrites au paragraphe 5.2.3.2.

10.1.3.4 Erreurs de protocole iSER

Si iSERHelloRequired est négocié à "Oui", alors le premier message iSER envoyé par la couche iSER chez l'initiateur DOIT être le message Hello iSER (voir au paragraphe 9.3). Dans ce cas le premier message iSER envoyé par la couche iSER à la cible DOIT être le message HelloReply iSER (voir au paragraphe 9.4). Manquer à envoyer le message iSER Hello ou HelloReply, comme indiqué par le mauvais Opcode dans l'en-tête iSER, est une erreur de protocole. À l'inverse, si le message Hello iSER est envoyé par la couche iSER chez l'initiateur quand iSERHelloRequired est négocié à "Non", la couche iSER à la cible PEUT traiter cela comme une erreur de protocole ou répondre avec un message HelloReply iSER. Le traitement des erreurs de protocole iSER est le même que celui des erreurs de format iSER comme décrit au paragraphe 10.1.3.3.

Si le côté envoyeur d'une connexion à capacité iSER agit d'une manière non permise par les valeurs de clé opérationnelle négociées ou déclarées d'établissement/texte comme décrit à la Section 6, c'est une erreur de protocole et le côté receveur PEUT traiter cela de la même façon qu'une erreur de format iSER comme décrit au paragraphe 10.1.3.3.

10.1.4 Erreurs dans la couche iSCSI

Le traitement d'erreur dû aux erreurs à la couche iSCSI est décrit dans les paragraphes qui suivent. Pour la récupération d'erreur, voir au paragraphe 10.2.

10.1.4.1 Erreurs de format iSCSI

Lorsque une erreur de format iSCSI est détectée, la couche iSCSI DOIT demander à la couche iSER de terminer le flux RCaP en invoquant la primitive opérationnelle Connection_Terminate. Pour les détails de la terminaison de connexion, voir au paragraphe 5.2.3.1.

10.1.4.2 Erreurs de résumé iSCSI

Dans le mode à assistance iSER, la couche iSCSI ne va voir aucune erreur de résumé parce que les deux clés HeaderDigest et DataDigest sont négociées à "Aucun".

10.1.4.3 Erreurs de séquence iSCSI

Pour l'iSCSI traditionnel, les erreurs de séquence sont causées par des PDU éliminées à cause d'erreurs d'en-tête ou de résumé de données. Comme les résumés ne sont pas utilisés dans le mode à assistance iSER et que la couche RCaP va livrer tous les messages dans leur ordre d'envoi, les erreurs de séquence ne se produisent pas en mode à assistance iSER.

10.1.4.4 Erreurs de protocole iSCSI

Lorsque la couche iSCSI traite certaines erreurs de protocole par l'abandon de la connexion, le traitement d'erreur est le même que pour les erreurs de format iSCSI comme décrit au paragraphe 10.1.4.1.

Lorsque la couche iSCSI utilise la PDU iSCSI Rejet et les codes de réponse pour traiter certaines autres erreurs de protocole, aucun traitement particulier n'est requis à la couche iSER.

10.1.4.5 Erreurs de temporisation et de session SCSI

Ceci est traité à la couche iSCSI, et aucun traitement particuler n'est requis à la couche iSER.

10.1.4.6 Échecs de négociation iSCSI

Pour les échecs de négociation qui surviennent durant la phase Établissement chez l'initiateur après que la couche iSCSI a invoqué la primitive opérationnelle Allocate_Connection_Resources et avant l'invocation de la primitive opérationnelle Enable_Datamover, la couche iSCSI DOIT demander à la couche iSER de désallouer toutes les ressources de connexion en invoquant la primitive opérationnelle Deallocate_Connection_Resources. La couche iSCSI chez l'initiateur DOIT terminer la connexion.

Pour les échecs de négociation durant la phase Établissement à la cible, la couche iSCSI peut utiliser une réponse Établissement avec une classe d'état autre que 0 (succès) pour terminer la phase Établissement. Si la couche iSCSI a invoqué la primitive opérationnelle Allocate_Connection_Resources et n'a pas encore invoqué la primitive opérationnelle Enable_Datamover, la couche iSCSI de la cible DOIT demander à la couche iSER de la cible de désallouer toutes les ressources de connexion en invoquant la primitive opérationnelle Deallocate_Connection_Resources. La couche iSCSI chez l'initiateur et chez la cible DOIVENT terminer la connexion.

Durant la phase d'établissement iSCSI, si la couche iSCSI chez l'initiateur reçoit une réponse Établissement de la cible avec une classe d'état autre que 0 (Succès) après que la couche iSCSI chez l'initiateur a invoqué la primitive opérationnelle Allocate_Connection_Resources, la couche iSCSI DOIT demander à la couche iSER de désallouer toutes les ressources de connexion en invoquant la primitive opérationnelle Deallocate_Connection_Resources. La couche iSCSI DOIT terminer la connexion dans ce cas.

Pour les échecs de négociation durant la phase de pleines caractéristiques, le traitement d'erreur est laissé à la couche iSCSI et aucun traitement particulier n'est requis à la couche iSER.

10.2 Récupération d'erreur

Les exigences de récupération d'erreur de iSCSI/iSER sont les mêmes que celle du iSCSI traditionnel iSCSI. Les trois niveaux de récupération d'erreur définis dans la [RFC7143] sont pris en charge par iSCSI/iSER.

* Pour ErrorRecoveryLevel 0, la récupération de session est traitée par iSCSI et aucun traitement particulier n'est requis à la couche iSER.
* Pour ErrorRecoveryLevel 1, voir au paragraphe 10.2.1 la récupération de PDU.
* Pour ErrorRecoveryLevel 2, voir au paragraphe 10.2.2 la récupération de connexion.

La couche iSCSI peut invoquer la primitive opérationnelle Notice_Key_Values durant l'établissement de connexion pour demander à la couche iSER de prendre note de la valeur du niveau de récupération d'erreur opérationnel, comme décrit aux paragraphes 5.1.1 et 5.1.2.

10.2.1 Récupération de PDU

Comme décrit aux paragraphes 10.1.4.2 et 10.1.4.3, les erreurs de résumé et de séquence ne vont pas se produire dans le mode à assistance iSER. Si la couche RCaP détecte une erreur, elle va clore la connexion iSCSI/iSER, comme décrit au paragraphe 10.1.2. Donc, la récupération de PDU n'est pas utile dans le mode à assistance iSER.

La couche iSCSI chez l'initiateur DEVRAIT désactiver les retransmissions de PDU iSCSI pilotées par des temporisations.

10.2.2 Récupération de connexion

La couche iSCSI chez l'initiateur PEUT réallouer l'allégeance de connexion pour les commandes non immédiates qui sont encore en cours et sont associées à la connexion défaillante en utilisant une demande de fonction de gestion de tâche avec la fonction Réallocation de tâche. Voir les détails au paragraphe 7.3.3.

Lorsque la couche iSCSI chez l'initiateur fait une réallocation de tâche pour une commande Écriture SCSI, elle DOIT qualifier l'invocation de la primitive opérationnelle Send_Control avec DataDescriptorOut, qui définit la mémoire tampon d'entrée/sortie pour les données non sollicitées non immédiates et les données sollicitées. Cela permet à la couche iSCSI de la cible d'utiliser des R2T de récupération pour demander les données envoyées à l'origine comme non sollicitées et solicitées à l'initiateur.

Lorsque la couche iSCSI de la cible accepte une demande de réallocation pour une commande SCSI Lecture, elle DOIT demander à la couche iSER de traiter les Data-In SCSI pour toutes les données non acquittées en invoquant la primitive opérationnelle Put_Data. Voir au paragraphe 7.3.5 le traitement des Data-In SCSI.

Lorsque la couche iSCSI de la cible accepte une demande de réallocation pour une commande Écriture SCSI, elle DOIT demander à la couche iSER de traiter une R2T de récupération pour toutes les données non sollicitées non immédiates et toutes les séquences de données sollicitées qui n'ont pas été reçues en invoquant la primitive opérationnelle Get_Data. Voir au paragraphe 7.3.6 le traitement de Prêt au transfert (R2T).

La couche iSCSI de la cible NE DOIT PAS produire des R2T de récupération sur une connexion iSCSI/iSER pour une tâche pour laquelle l'allégeance de connexion n'a jamais été réallouée. La couche iSER de la cible PEUT rejeter une telle R2T de récupération reçue via l'invocation de la primitive opérationnelle Get_Data de la couche iSCSI de la cible, avec un code d'erreur approprié.

La couche iSER de la cible va traiter les demandes invoquéds par les primitives opérationnelles Put_Data et Get_Data pour une tâche réallouée de la même façon que pour les commandes d'origine.

11 Considérations sur la sécurité

Lorsque iSER est mis en couche par dessus une couche RCaP et fournit les extensions RDMA au protocole iSCSI, les considérations de sécurité de iSER sont les mêmes que celles de la couche RcaP sous-jacente. Pour iWARP, c'est décrit dans les [RFC5040] et [RFC5042], plus les mises à jour à ces deux RFC qui sont contenues dans la [RFC7146].

Comme le protocole iSCSI assisté par iSER est encore fonctionnellement iSCSI du point de vue des considérations pour la sécurité, toutes les exigences de sécurité de iSCSI décrites dans la [RFC7143] s'appliquent. Si iSER est mis en couches par dessus une couche RcaP non fondée sur IP, tous les mécanismes de protocole de sécurité applicables à cette couche RCaP sont aussi applicables à une connexion iSCSI/iSER. Si iSER est mis en couche par dessus un protocole non IP, le mécanisme IPsec spécifié dans la [RFC7143] DOIT être mis en œuvre à tout point où le protocole iSER entre dans le

réseau IP (par exemple, via des passerelles) et le protocole non IP DEVRAIT mettre en œuvre un protocole de sécurité paquet par paquet (d'utilisation facultative) de force égale au mécanisme IPsec spécifié par la [RFC7143].

Afin de protéger les ressources de connexion RCaP de la cible contre de possible attaques en épuisement de ressources, l'allocation de telles ressources pour une nouvelle connexion DOIT être différée jusqu'à ce qu'elle soit raisonnablement certaine que la nouvelle connexion ne fait pas partie d'une attaque en épuisement de ressources (par exemple, jusqu'à la fin de l'étape SecurityNegotiation de l'établissement) ; voir au paragraphe 5.1.2.

Une STag valide expose les ressources de mémoire tampon d'entrée/sortie au réseau pour l'accès via le RCaP. Les mesures de sécurité pour le RCAP et iSER décrites dans les paragraphes précédents peuvent être utilisées pour protéger les données dans une mémoire tampon d'entrée/sortie contre la divulgation ou modification non désirées, et ces mesures sont d'une importance accrue pour les mises en œuvre qui conservent (par exemple, en antémémoire) les STag pour les utiliser dans plusieurs tâches (par exemple, opérations d'entrée/sortie iSCSI) parce que les ressources sont exposées au réseau pendant une plus longue période.

Un moyen complémentaire de contrôler l'exposition des ressources de mémoire tampon d'entrée/sortie est l'invalidation de la STag après l'achèvement de la tâche associée, comme spécifié au paragraphe 1.5.1. L'utilisation de Send avec des messages Invalider (qui cause l'invalidation de STag à distance) est FACULTATIVE, donc la couche iSER NE DOIT PAS s'appuyer sur l'utilisation d'un Send avec Invalider par son homologue distant pour causer l'invalidation locale d'une STag. Si on s'attend à ce qu'une STag soit invalide après l'achèvement d'une tâche, la couche iSER DOIT vérifier la STag et l'invalider si elle est encore valide.

12. Considerations relatives à l'IANA

L'IANA a ajouté les entrées suivantes au registre "iSCSI Login/Text Keys" :
MaxAHSLength, RFC 7145
TaggedBufferForSolicitedDataOnly, RFC 7145
iSERHelloRequired, RFC 7145

L'IANA a mis à jour les entrées suivantes dans le registre "iSCSI Login/Text Keys" pour faire référence à la présente RFC.
InitiatorRecvDataSegmentLength
MaxOutstandingUnexpectedPDUs
RDMAExtensions
TargetRecvDataSegmentLength

L'IANA a aussi changé la référence à la RFC 5046 pour le registre "iSCSI Login/Text Keys" pour se référer à cette RFC.

L'IANA a mis à jour les enregistrements des Opcodes iSER 1 à 3 dans le registre "iSER Opcodes" pour faire référence à la présente RFC. L'IANA a aussi changé la référence à la RFC 5046 pour le registre "iSER Opcodes" pour se référer à la présente RFC.

13. Références

13.1 Références normatives

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

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

[RFC5040] R. Recio et autres, "Spécification d'un protocole d'accès direct à une mémoire distante", octobre 2007. *(P.S.)*

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

[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.)*

[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. (*MàJ par la* RFC6581) *(P.S.)*

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

[RFC7143] M. Chadalapaka et autres, "Protocole (consolidé) d'interface Internet de petit système d'ordinateur (iSCSI)", avril 2014. (*Remplace* RFC3720, 3980, 4850, 5048) (*MàJ* RFC3721) *(P.S.)*

[RFC7146] D. Black, P. Koning, "Sécurisation des protocoles de mémorisation de blocs sur IP : mise à jour des exigences de la RFC3723 pour IPsec v3", avril 2014. (*MàJ* RFC3720, 3723, 3821, 3822, 4018, 4172, 4173, 4174, 5040, 5041, 5042, 5043, 5044, 5045, 5046, 5047, 5048) *(P.S.)*

13.2 Références pour information

[IB] "InfiniBand Architecture Specification Volume 1 Release 1.2", octobre 2004

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

[RFC5047] M. Chadalapaka et autres, "DA : Architecture Datamover pour l'interface système de petit ordinateur à l'Internet (iSCSI)", octobre 2007. *(Information)*

[RFC7144] F. Knight, M. Chadalapaka, "Mise à jour des caractéristiques SCSI d'interface Internet de petit système d'ordinateur (iSCSI)", avril 2014. *(P.S.)*

[SAM5] INCITS Technical Committee T10, "Modèle d'architecture SCSI - 5 (SAM-5)", T10/BSR INCITS 515 rev 04, Committee Draft.

Appendice A. Résumé des changements à la RFC 5046

Tous les changements sont rétro compatibles avec la RFC 5046 sauf pour le n° 8, qui reflète toutes les mises en œuvre connues de iSER, dont chacune a mis en œuvre ce changement, en dépit de son absence dans la RFC 5046. Par suite, une mise en œuvre hypothétique fondée sur la RFC 5046 ne va pas interopérer avec une mise en œuvre fondée sur la présente version de la spécification.

1. Retrait de l'exigence qu'une connexion soit ouverte en mode TCP "normal" et transite au mode zéro copie. Cela permet que la spécification se conforme aux mises en œuvre existantes pour InfiniBand et iWARP. Les changements sont faits à la Section 1, aux paragraphes 3.1.6, 4.2, 5.1, 5.1.1, 5.1.2, 5.1.3, 10.1.3.4, et à la Section 11.

2. Ajout d'une clause au paragraphe 6.2 pour préciser que MaxRecvDataSegmentLength doit être ignorée si elle est déclarée dans la phase d'établissement.

3. Ajout d'une clause au paragraphe 6.2 pour préciser que l'initiateur ne doit pas envoyer plus de données que InitiatorMaxRecvDataSegmentLength quand une demande NOP-Out est envoyée avec une étiquette de tâche d'initiateur valide. Comme InitiatorMaxRecvDataSegmentLength peut être inférieure à TargetMaxRecvDataSegmentLength, retourner les données originales dans la demande NOP-Out dans cette situation peut faire déborder la mémoire tampon de réception sauf si la longueur des données envoyées avec la demande NOP-Out est inférieure à InitiatorMaxRecvDataSegmentLength.

4. Ajout d'une recommandation DEVRAIT négocier pour MaxOutstandingUnexpectedPDUs au paragraphe 6.7.

5. Ajout de la clé MaxAHSLength au paragraphe 6.8 pour mettre une limite à la longueur de AHS. C'est utile quand on établit une mémoire tampon de réception en sachant que la longueur maximum possible de message est dans une PDU qui contient AHS.

6. Ajout de la clé TaggedBufferForSolicitedDataOnly au paragraphe 6.9 pour indiquer comment la région de mémoire va être utilisée. Un initiateur peut traiter différemment les régions de mémoire destinées aux données non sollicitées et aux données sollicitées et peut utiliser des modes d'enregistrement différents. À l'opposé, la RFC 5046 traite la mémoire

occupée par les données comme contiguës (ou virtuellement contiguës, au moyen de mécanisme de dispersion-rassemblement) et une région homogène. L'ajout d'une nouvelle clé permettra que différents modèles de mémoire soient acceptés. Le changement a aussi été fait au paragraphe 7.3.1.

7. Ajout de la clé iSERHelloRequired au paragraphe 6.10 pour permettre à un initiateur d'allouer des ressources de connexion après le processus d'établissement en exigeant l'utilisation des messages iSER Hello avant l'envoi des PDU iSCSI. La valeur par défaut est "Non" car les messages iSER Hello n'ont pas été mis en œuvre et ne sont pas utilisés. Le changement a été fait aux paragraphes 5.1.1, 5.1.2, 5.1.3, 8.2, 9.3, 9.4, 10.1.3.2, et 10.1.3.4.

8. Ajout de deux champs de 64 bits dans l'en-tête iSER au paragraphe 9.2 pour le décalage de base en lecture et le décalage de base en écriture pour s'accommoder d'un décalage de base non zéro. Cela permet qu'une mise en œuvre telle que OFED (*Open Fabrics Enterprise Distribution*) soit utilisée dans les deux environnements InfiniBand et iWARP. Les changements ont été faits aux définitions de décalage de base, annonce, et mémoire tampon étiquetée. Le changement a aussi été fait aux paragraphes 1.5.1, 1.6, 1.7, 7.3.1, 7.3.3, 7.3.5, 7.3.6, 9.1, 9.3, 9.4, 9.5.1, et 9.5.2. Ce changement n'est pas rétro compatible avec la RFC 5046, mais il fait partie de toutes les mises en œuvre connues de iSER au moment du développement du présent document.

9. Suppression du comportement spécifique de iWARP. Les changements ont été faits dans les définitions d'opération RDMA et le type de message Send. Des précisions ont été ajoutées au paragraphe 1.5.2 sur l'utilisation de SendSE et SendInvSE. Ces précisions reflètent la suppression des exigences de la RFC 5046 sur l'utilisation de ces messages, car les mises en œuvre n'ont pas suivi la RFC 5046 dans ce domaine. Les changements qui affectent Send avec Invalider ont été faits aux paragraphes 1.5.1, 1.6, 1.7, 4.1, et 7.3.2. Les changements qui affectent Terminer ont été faits aux paragraphes 10.1.2.1 et 10.1.2.2. Des changements ont été faits à l'Appendice B pour supprimer les en-têtes iWARP.

10. Suppression des descriptions de déni de service pour l'initiateur au paragraphe 5.1.1 car elles ne sont applicables qu'à la cible.

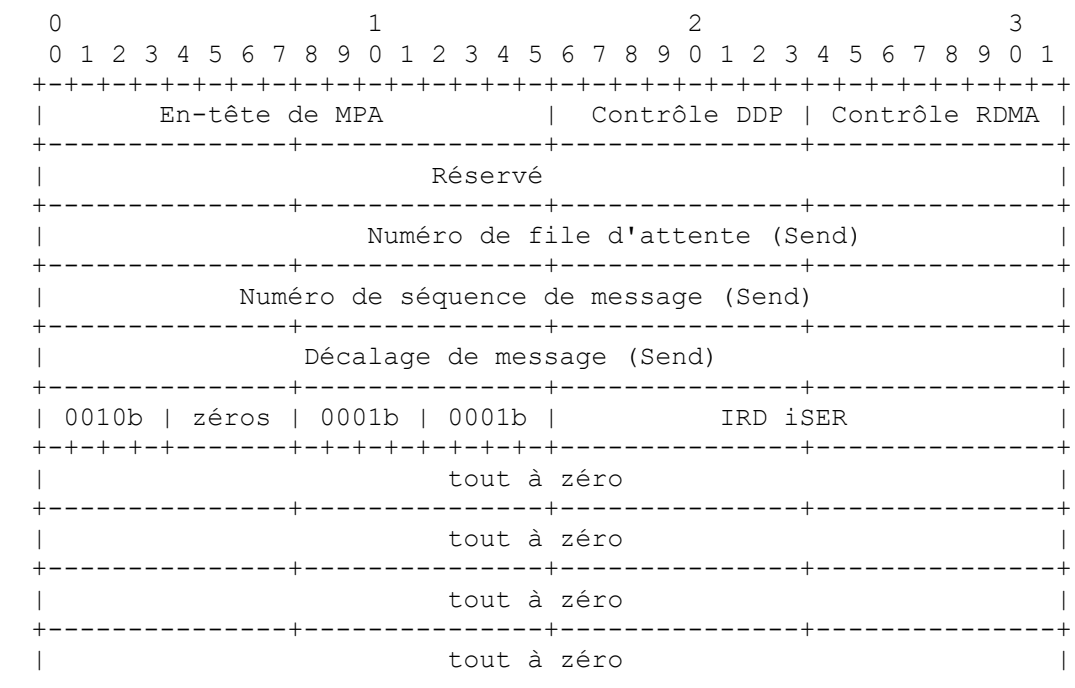
11. Précision au paragraphe 1.5.1 que l'invalidation de STag est de la responsabilité de l'initiateur pour des raisons de sécurité, et l'initiateur ne peut pas s'appuyer sur l'utilisation par la cible d'une version d'invalidation de Send. Ajout de texte à la Section 11 sur l'invalidation de STag.

Appendice B. Format de message pour iSER

Cet appendice est seulement pour information et NE FAIT PAS partie de la norme.

B.1 Format de message iWARP pour message Hello ISER

La figure suivante décrit un message Hello iSER encapsulé dans un message SendSE iWARP.



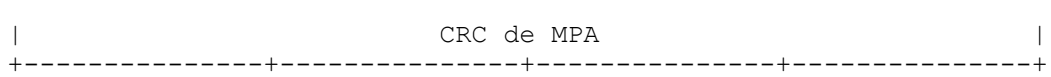


Figure 6 : Message SendSE contenant un message Hello iSER

B.2 Format de message iWARP pour message iSER HelloReply

La figure suivante décrit un message HelloReply iSER encapsulé dans un message SendSE iWARP. Le fanion Rejet (J) est réglé à zéro.

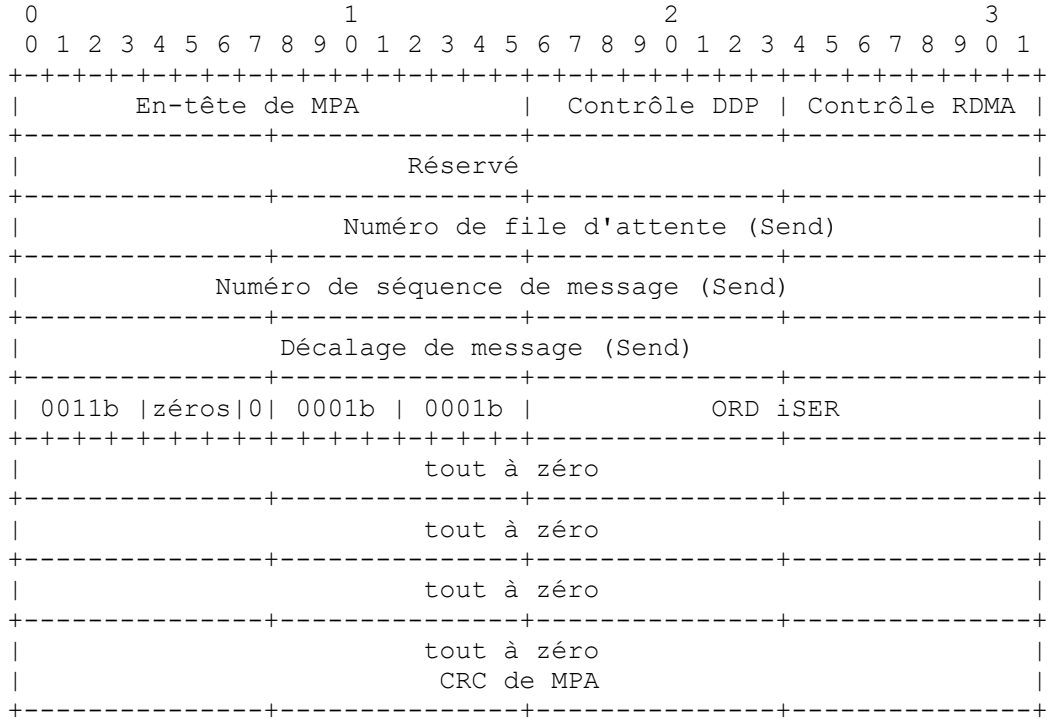


Figure 7 : Message SendSE contenant un message HelloReply iSER

B.3 Format d'en-tête iSER pour PDU Commande de lecture SCSI

La figure suivante décrit une PDU Commande de lecture SCSI incorporée dans un message iSER. Pour cet exemple particulier, dans l'en-tête iSER, le fanion STag Écriture valide est réglé à zéro, le fanion STag Lecture valide est réglé à un, le champ STag Écriture est réglé tout à zéro, le champ Décalage de base d'écriture est réglé tout à zéro, le champ STag Lecture contient une STag Lecture valide, et le champ Décalage de base de lecture contient un décalage de base valide pour la mémoire tampon étiquetée de lecture.

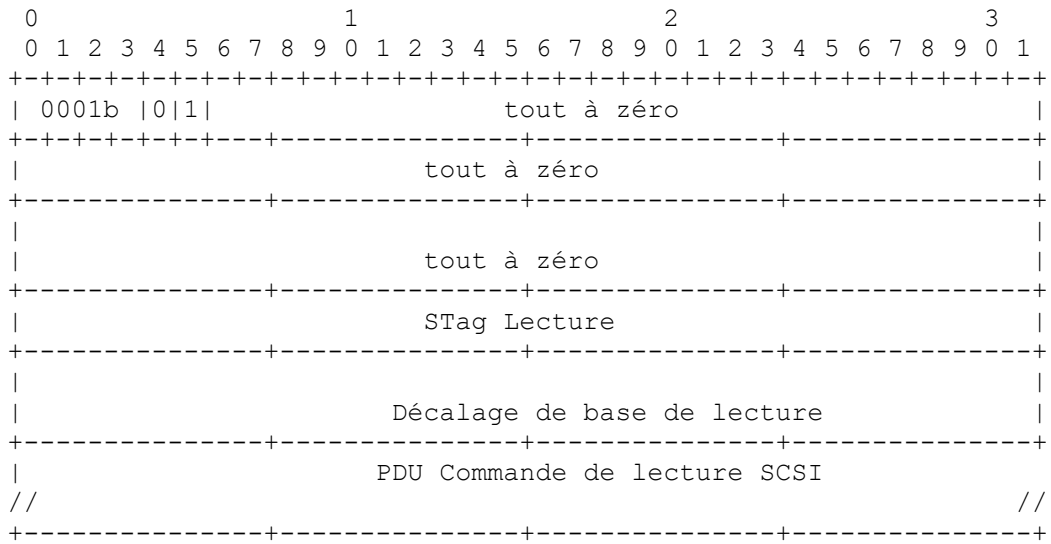


Figure 8 : Format d'en-tête iSER pour PDU Commande de lecture SCSI

B.4 Format d'en-tête iSER pour PDU Commande d'écriture SCSI

La figure qui suit décrit une PDU Commande d'écriture SCSI incorporée dans un message iSER. Pour cet exemple particulier, dans l'en-tête iSER, le fanion STag Écriture valide est réglé à un, le fanion STag Lecture valide est réglé à zéro, le champ STag Écriture contient une STag Écriture valide, le champ décalage de base d'écriture contient un décalage de base valide pour la mémoire tampon d'écriture étiquetée, le champ STag Lecture est réglé tout à zéro car il n'est pas utilisé, et le champ décalage de base de lecture est réglé tout à zéro.

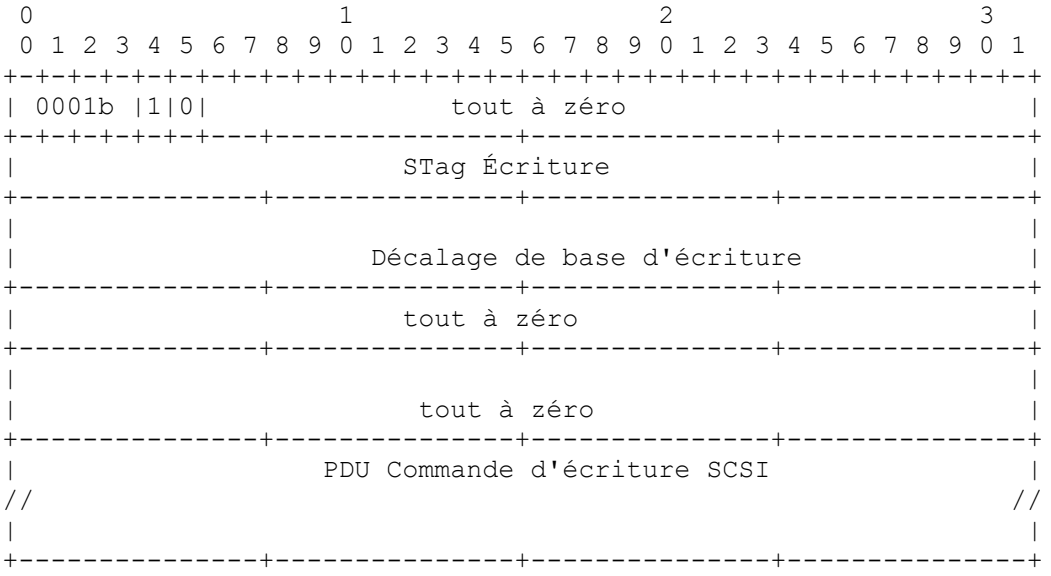


Figure 9 : Format d'en-tête iSER pour PDU commande d'écriture SCSI

B.5 Format d'en-tête iSER pour PDU Réponse SCSI

La figure suivante décrit une PDU Réponse SCSI incorporée dans un message iSER :

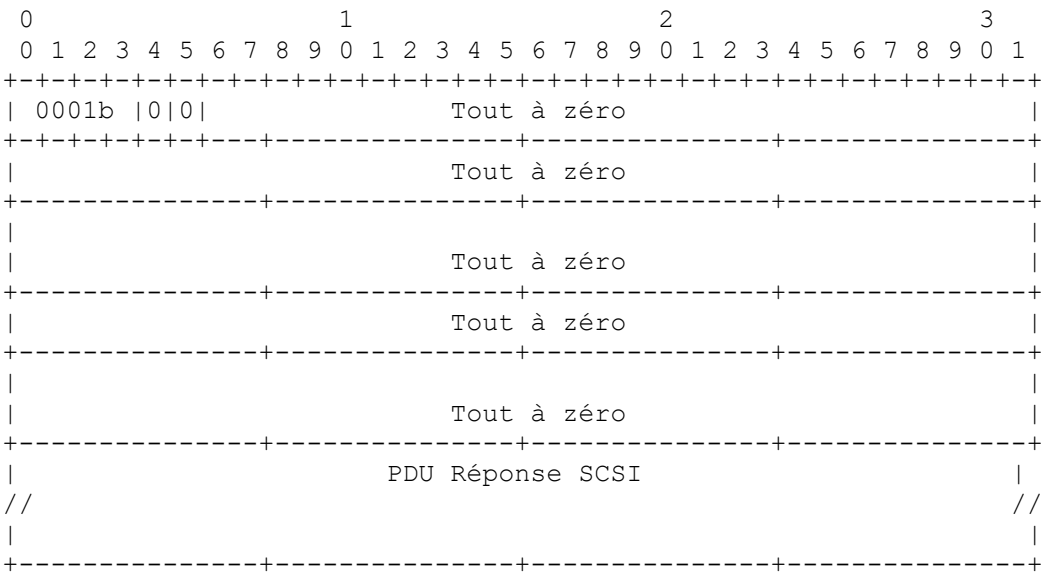


Figure 10: Format d'en-tête iSER pour PDU Réponse SCSI

Appendice C. Discussion de l'architecture iSER sur InfiniBand

Cette section explique comment un réseau InfiniBand (avec des passerelles) serait structuré. Elle est seulement pour information et n'est destinée qu'à donner une vue de la façon dont iSER est utilisé dans un environnement InfiniBand.

C.1 Côté hôte de connexions iSCSI et iSER en InfiniBand

La Figure 11 définit la topologie dans laquelle iSCSI et iSER seront capables de fonctionner sur un réseau InfiniBand.

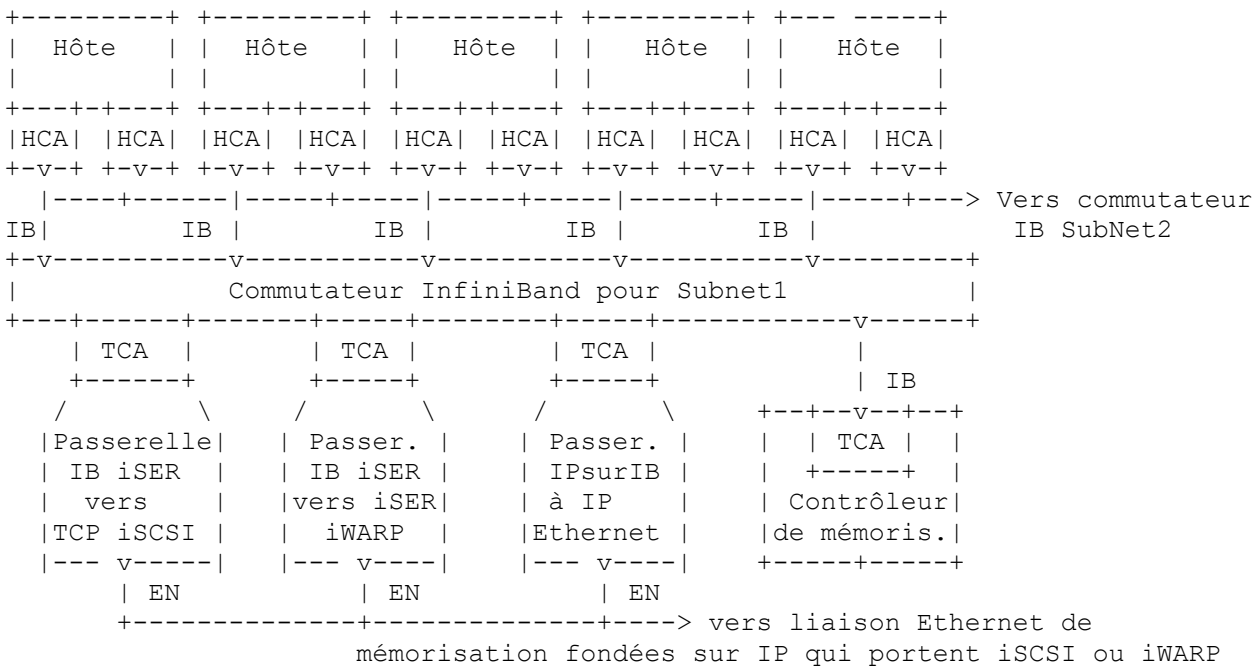


Figure 11 : iSCSI et iSER sur IB

Dans la Figure 11, les systèmes hôtes sont connectés via les adaptateurs de canal hôte InfiniBand (HCA, *Host Channel Adapter*) aux liens InfiniBand. Avec l'utilisation de commutateurs IB, les liaisons InfiniBand connectent le HCA aux adaptateurs de canal cible InfiniBand (TCA, *Target Channel Adapter*) situés dans les passerelles ou contrôleurs de mémorisation. Une passerelle IB-IP à capacité iSER convertit les messages iSER encapsulés dans les protocoles IB soit en iSCSI standard, soit en messages iSER pour iWARP. Une passerelle de la [RFC4391] convertit le protocole InfiniBand [RFC4391] en protocole IP, et dans le cas de iSCSI, permet que iSCSI fonctionne sur un réseau IB entre les hôtes et la passerelle [RFC4391].

C.2 Environnement réseau mixte du côté mémorisation de iSCSI et iSER

La Figure 12 montre un contrôleur de mémorisation qui a trois groupes portails différents : un qui ne prend en charge que iSCSI (TPG-4), un qui prend en charge iSER/iWARP ou iSCSI (TPG-2), et un qui prend en charge iSER/IB (TPG-1). Ici, TPG, "Target Portal Group" signifie groupe portail cible.

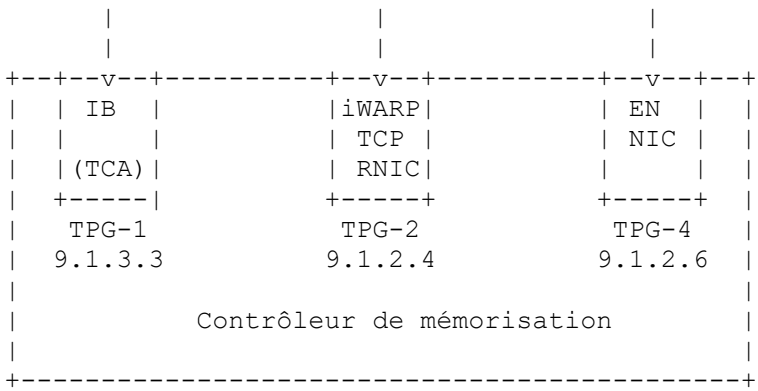


Figure 12 : Contrôleur de mémorisation avec connexions TCP, iWARP, et IB

Le processus normal d'annonce de groupe portail iSCSI (via le protocole de localisation de service (SLP, *Service Location Protocol*), le service de nom de mémorisation Internet (iSNS, *Internet Storage Name Service*), ou SendTargets) est disponible pour un contrôleur de mémorisation.

C.3 Processus de découverte pour un hôte InfiniBand

Un système d'hôte InfiniBand peut rassembler les adresses IP de groupe portail à partir des processus de découverte de SLP, iSNS, ou SendTargets en utilisant TCP/IP via la [RFC4391]. Après l'obtention d'une ou plusieurs adresses IP de portail distant, l'initiateur utilise les mécanismes IP standard pour résoudre l'adresse IP en une interfacve sortante locale et l'adresse de matériel de destination (MAC Ethernet ou identifiant mondial InfiniBand (GID) de la cible ou d'une passerelle menant à la cible). Si l'interface résoute est une interface réseau de la [RFC4391], le portail cible peut alors être atteint par un tissu InfiniBand. Dans ce cas, l'initiateur peut établir une session iSCSI/TCP ou iSCSI/iSER avec la cible sur cette interface InfiniBand, en utilisant l'adresse de matériel (GID InfiniBand) obtenue par le processus standard du protocole de résolution d'adresse (ARP, *Address Resolution Protocol*).

Si plus d'une adresse IP est obtenue par le processus de découverte, l'initiateur devrait choisir une adresse IP de cible qui soit sur le même sous-réseau IP que l'initiateur, si il en existe. Cela va éviter une surcharge potentielle de passage par une passerelle alors qu'existe un chemin direct.

De plus, un utilisateur peut configurer des entrées manuelle de chemin IP statique , si un chemin particulier vers la cible est préféré.

C.4 Spécifications de connexion IBTA

Cela sort du domaine d'application du présent document, mais on s'attend à ce que la InfiniBand Trade Association (IBTA) définisse :

* L'identifiant de service iSER
* Un moyen pour permettre à un hôte d'établir une connexion avec un nœud d'extrémité InfiniBand homologue, et que cet homologue indique quand ce nœud d'extrémité prend en charge iSER, de sorte que l'hôte soit capable de revenir à iSCSI/TCP sur la [RFC4391].
* Un moyen de permettre à l'hôte d'établir des connexions avec des connexions IB iSER sur des contrôleurs de mémorisation ou des passerelles IB connectées sur iSER de préférence à des passerelles/pont connectés sur IPoIB ou des connexions à des contrôleurs de mémorisation cibles qui acceptent aussi iSCSI via la [RFC4391].
* Un moyen de combiner l'identifiant de service IB pour iSER et le numéro d'accès IP de telle façon que l'hôte IB puisse utiliser les processus normaux de connexion IB, et de s'assurer donc que l'homologue cible iSER peut bien se connecter au numéro d'accès IP requis.

Appendice D. Remerciements

Les auteurs remercient les personnes suivantes qui ont identifié les problèmes de mise en œuvre et/ou suggéré des solutions aux problèmes traités dans le présent document : Robert Russell, Arne Redlich, David Black, Mallikarjun Chadalapaka, Tom Talpey, Felix Marti, Robert Sharp, Caitlin Bestler, Hemal Shah, Spencer Dawkins, Pete Resnick, Ted Lemon, Pete McCann, et Steve Kent. Merci aussi aux auteurs de la spécification iSER d'origine [RFC5046], incluant Michael Ko, Mallikarjun Chadalapaka, John Hufferd, Uri Elzur, Hemal Shah, et Patricia Thaler. Le présent document a bénéficié de toutes leurs contributions.

Adresse des auteurs

Michael Ko
mél : mkosjc@gmail.com

Alexander Nezhinsky
Mellanox Technologis
13 Zarchin St.
Raanana 43662
Israel
téléphone : +972-74-712-9000
mél : alexandern@mellanox.com , nezhinsky@gmail.com