

12

DEMANDE DE BREVET D'INVENTION

A1

22 Date de dépôt : 09.03.01.

30 Priorité : 09.03.00 US 09524401.

43 Date de mise à la disposition du public de la
demande : 05.10.01 Bulletin 01/40.

56 Liste des documents cités dans le rapport de
recherche préliminaire : *Ce dernier n'a pas été
établi à la date de publication de la demande.*

60 Références à d'autres documents nationaux
apparentés :

71 Demandeur(s) : DELL PRODUCTS, L.P. — US.

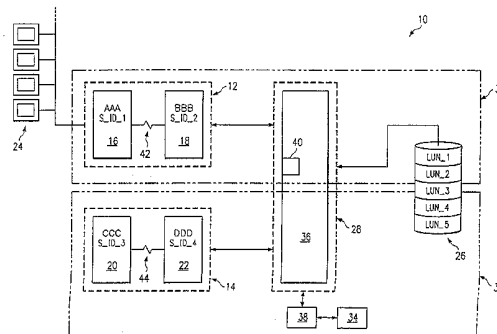
72 Inventeur(s) : TAWIL AHMAD H et NGUYEN NAM V.

73 Titulaire(s) :

74 Mandataire(s) : CABINET BARDEHLE PAGENBERG
ET PARTNER.

54 SYSTEME ET PROCEDE DE GESTION DE RESSOURCES MEMOIRE DANS UN ENVIRONNEMENT
INFORMATIQUE EN GRAPPES.

57 La présente invention concerne un système et un procédé de gestion de ressources mémoire au sein d'un environnement informatique en grappes. Le procédé comprend le maintien d'une réservation sur une ressource mémoire pour un premier noeud (16 ou 18, 20 ou 22) d'un environnement informatique en grappes. Il est possible d'effectuer une déconnexion de traitement tierce pour le premier noeud (16 ou 18, 20 ou 22), et libérer la réservation maintenue pour le premier noeud (16 ou 18, 20 ou 22).



La présente invention concerne de façon générale le domaine des systèmes de mémorisation de données et, en particulier, un système et un procédé de gestion de ressources mémoire dans un environnement informatique en grappes.

Les réseaux locaux de mémorisation (SANs, *Storage Area Networks*) incluent souvent une collection de ressources mémoire de données reliées d'une manière communicante à une pluralité de nœuds tels que des stations de travail et des serveurs. Dans la présente description, les termes "nœud" et "serveur" sont utilisés indifféremment, en s'accordant sur le fait qu'un "serveur" est un type de "nœud".

Au sein d'un réseau SAN, un serveur peut effectuer un accès à une ressource mémoire de données à travers une structure en utilisant le protocole *Fiber Channel* ou Canal de Fibre. Le protocole Canal de Fibre peut agir comme une couche physique classique qui permet le transport de multiples protocoles de couche supérieure, tels que le protocole SCSI (*Small Computer System Interface*, protocole d'interconnexion système pour petits ordinateurs). Dans un environnement SAN, le protocole SCSI peut attribuer des numéros d'unité logique (LUNs, *Logical Unit Numbers*) à la collection de ressources mémoire de données. Les numéros LUN peuvent permettre à un serveur se trouvant au sein d'un réseau SAN d'effectuer un accès à des ressources mémoire de données spécifiques en faisant référence à un LUN SCSI concernant une ressource mémoire de données spécifiques.

Bien qu'un système de mémorisation Canal de Fibre puisse offrir un grand choix de capacité de mémoire, l'implémentation du système peut également avoir un coût très élevé. En résultat, les utilisateurs cherchent souvent à partager la mémoire disponible fournie par le système entre de multiples serveurs. Malheureusement, si un

serveur relié à un réseau SAN donné utilise le système d'exploitation MICROSOFT WINDOWS NT (nom commercial déposé), le serveur peut tenter de s'approprier tout numéro LUN visible par le serveur. Par exemple, si un serveur particulier détecte plusieurs LUNs lorsque le serveur s'amorce, il peut supposer que chaque LUN est à sa disposition. Par conséquent, si de multiples serveurs WINDOWS NT sont attachés à un spoule de mémoire ou à une collection de ressources mémoire de données, chaque serveur peut tenter de prendre la commande de chaque LUN se trouvant dans le spoule de mémoire. Cette situation peut déclencher des conflits lorsque plus d'un serveur tente d'effectuer un accès au même LUN.

Un utilisateur qui cherche à résoudre ce problème peut partitionner ou diviser en zones la mémoire disponible par l'intermédiaire d'un filtrage ou par l'intermédiaire de l'utilisation de pilotes de miniport qui possèdent des capacités de masquage de LUN. En effet, ce partitionnement peut éviter qu'un serveur qui exécute WINDOWS NT puisse obtenir une capacité de mémoire qui ne lui est pas attribuée. Cette approche peut être efficace pour des serveurs autonomes, mais l'approche présente plusieurs inconvénients dans un environnement informatique en grappes.

L'organisation en grappes implique la configuration d'un groupe de serveurs indépendants de sorte qu'ils paraissent dans un réseau comme étant une seule machine. Souvent, les grappes sont gérées comme un seul système, partagent un espace de noms commun, et sont conçues d'une manière spécifique pour tolérer des défaillances de composant et pour supporter d'une manière transparente l'ajout ou le retrait de composants. Malheureusement, du fait qu'une grappe peut avoir deux ou plus de deux serveurs qui apparaissent comme une seule machine, les techniques de partitionnement mentionnées ci-dessus peuvent

s'avérer être une solution inefficace pour éviter des conflits lorsque les deux ou plus de deux serveurs tentent d'effectuer un accès au même LUN.

Le système MICROSOFT CLUSTER SERVER (MSCS) met en oeuvre une technique actuellement disponible pour arbitrer des conflits et gérer la propriété de dispositifs de mémoire dans un environnement informatique en grappes. Un système MSCS peut fonctionner à l'intérieur d'une grappe qui possède deux serveurs, un serveur A, qui peut être en service, et un serveur B. En cours de fonctionnement, le serveur A peut passer un signal de pulsation périodique au serveur B afin que le serveur B sache que le serveur A est "vivant". Si le serveur B ne reçoit pas à temps une pulsation en provenance du serveur A, le serveur B peut chercher à déterminer si le serveur A fonctionne et/ou si le serveur B peut s'approprier un quelconque des LUNs réservés au serveur A. Malheureusement, ce système MSCS peut utiliser des réinitialisations de cible SCSI pendant ce traitement, et les réinitialisations SCSI peuvent créer plusieurs problèmes. Par exemple, une réinitialisation SCSI classique au sein du système MSCS peut amener tous les serveurs se trouvant dans un système à Canal de Fibre donné à interrompre leurs traitements d'entrée/sortie "I/O" en instance. Ces traitements I/O interrompus peuvent éventuellement être achevés mais pas avant la stabilisation du bus. Cette approche d'interruption/attente/relance peut avoir un effet nuisible sur la performance totale du système.

En plus de cet effet potentiel sur les performances, le système MSCS et son utilisation des réinitialisations SCSI peuvent avoir un effet nuisible sur la fiabilité totale du système. En cours de fonctionnement, le système MSCS peut rendre compte d'une seule réinitialisation SCSI à la fois. L'incapacité de rendre compte des réinitialisations SCSI suivantes peut provoquer un com-

portement inattendu et diminuer la fiabilité du système.

La présente invention propose un système et un procédé de gestion de ressources mémoire dans un environnement informatique en grappes qui procurent d'importants avantages par rapport aux techniques antérieurement développées. Le système et le procédé décrits peuvent permettre une gestion de ressources mémoire et un arbitrage de conflits en réduisant le recours aux réinitialisations SCSI.

10 Selon un aspect de la présente invention, un procédé de gestion de ressources mémoire dans un environnement informatique en grappes peut inclure le maintien d'une réservation sur une ressource mémoire pour un premier nœud de l'environnement informatique en grappes. Le
15 nœud peut être, par exemple, un serveur, une station de travail ou tout autre dispositif informatique se trouvant dans la grappe.

 Une déconnexion de traitement tierce du premier nœud peut être effectuée, et la réservation maintenue
20 pour le premier nœud peut être libérée. Dans un mode de réalisation, la déconnexion de traitement tierce peut survenir en réponse à une commande de déconnexion envoyée au nom du premier nœud. La commande de déconnexion de traitement tierce peut être envoyée, par exemple, par un
25 second nœud ou par un commutateur Canal de Fibre. La commande de déconnexion de traitement tierce peut inclure des informations d'identification qui identifient le premier nœud en tant qu'émetteur de la commande de déconnexion même si le premier nœud n'a pas été l'émetteur
30 réel. Les informations d'identification peuvent inclure, par exemple, un nom de la toile mondiale et un identificateur source attribué au premier nœud.

 La gestion de ressources mémoire dans un environnement informatique en grappes peut en plus impliquer la
35 division en zones d'un système de mémorisation à Canal de

Fibre. Une zone peut regrouper un premier nœud et un second nœud et une pluralité de ressources mémoire telles que des disques durs et d'autres dispositifs de mémorisation de données. Dans le système divisé en zones, un second nœud peut se déconnecter tout seul après l'envoi d'une commande de déconnexion de traitement tierce à un premier nœud. Après la déconnexion des deux nœuds, une réinitialisation de liaison de protocole d'initialisation de boucle (LIP, *Loop Initialization Protocol*) peut être amorcée, une notification de changement d'état peut être générée et tout nœud fonctionnel peut se connecter à nouveau.

Selon un autre aspect de la présente invention, un procédé de gestion de ressources mémoire dans un environnement informatique en grappes peut inclure la réception d'une commande de réservation qui tente de réserver une ressource mémoire pour un nœud se trouvant dans la grappe. En réponse à la commande de réservation, une commande de sortie de réserve persistante SCSI ayant une action de service de réserve peut être émise pour réserver la ressource mémoire pour le nœud. Cette réserve persistante peut maintenir une réservation effaçable sur la ressource mémoire. Dans un mode de réalisation, la réservation peut être effacée en envoyant une commande de sortie de réserve persistante SCSI ayant une action de service d'effacement. Les commandes de réserve persistante peuvent permettre à des réservations LUN d'être individuellement libérées par opposition à un effacement de plusieurs réservations LUN à la fois effectué à l'aide d'une réinitialisation SCSI.

Selon un autre aspect de la présente invention, un système informatique servant à gérer des ressources mémoire dans un environnement informatique en grappes peut inclure un premier nœud, un second nœud et un moteur de gestion de ressources sert à déconnecter le premier

nœud en réponse à un échec du second nœud à recevoir à temps un signal de pulsation en provenance du premier nœud. Le système peut également inclure un support lisible par ordinateur mémorisant le moteur de gestion de ressources et une unité centrale de traitement reliée d'une manière communicante au support lisible par ordinateur et qui sert à exécuter le moteur de gestion de ressources.

Dans un mode de réalisation, le système peut également inclure une pluralité de plates-formes informatiques reliées d'une manière communicante au premier nœud. Ces plates-formes informatiques peuvent être, par exemple, un groupe d'ordinateurs personnels organisés en réseau. Le système peut également inclure un commutateur Canal de Fibre relié d'une manière communicante au premier nœud et à une pluralité de ressources mémoire. Le commutateur Canal de Fibre peut, dans certains modes de réalisation, inclure une unité centrale de traitement qui sert à faire fonctionner un moteur de gestion de ressources.

Un système et un procédé incorporant les enseignements de la présente invention peuvent procurer d'importantes améliorations aux solutions classiques de gestion de ressources de grappe. Par exemple, les techniques décrites peuvent être utilisées pour mieux gérer et arbitrer des conflits de ressources mémoire. Comme décrit ci-dessus, une réinitialisation SCSI dans un environnement informatique en grappes peut avoir en résultat le lancement de plusieurs traitements I/O ayant une approche d'interruption/attente/relance, qui peut avoir un effet nuisible sur la performance totale du système. Les enseignements de la présente invention peuvent aider à réduire l'appui des réinitialisations SCSI et les dégradations de performance en résultant.

De plus, les enseignements de la présente inven-

tion peuvent faciliter la résolution des problèmes de fiabilité du système associés aux réinitialisations SCSI dans un environnement informatique en grappes. Un système classique de gestion de ressources de grappe, tel qu'un système MSCS, peut être incapable de rendre compte des réinitialisations SCSI amorcées pendant la perturbation du bus relative à une réinitialisation SCSI précédente. Cette limitation peut provoquer un comportement inattendu et diminuer la fiabilité du système. Du fait que les enseignements de la présente description peuvent faciliter l'annulation d'au moins certaines réinitialisations SCSI, la fiabilité du système peut être améliorée.

D'autres avantages apparaîtront à la lecture de la description qui va suivre, faite en référence aux dessins annexés sur lesquels les références numériques indiquent des caractéristiques analogues.

La figure 1 décrit un schéma des composants d'un réseau local de mémorisation incluant un mode de réalisation d'un moteur de gestion de ressources qui incorpore les enseignements de la présente invention,

La figure 2 représente un ordinogramme relatif à un mode de mise en œuvre d'un procédé de gestion de ressources mémoire dans un environnement informatique en grappes conformément aux enseignements de la présente invention, et

La figure 3 représente un ordinogramme relatif à un autre mode de mise en œuvre d'un procédé de gestion de ressources mémoire dans un environnement informatique en grappes conformément aux enseignements de la présente invention.

La figure 1 représente un schéma fonctionnel général d'un réseau local de mémorisation (SAN), indiqué en général par la référence numérique 10. Le réseau SAN 10

inclut deux systèmes informatiques en grappes, des grappes 12 et 14. Comme décrit, la grappe 12 inclut un nœud 16 et un nœud 18, et la grappe 14 inclut un nœud 20 et un nœud 22. Les nœuds 16, 18, 20 et 22 peuvent être, par exemple, des serveurs, des stations de travail ou d'autres dispositifs informatiques d'un réseau. Comme décrit sur la figure 1, la grappe 12 peut supporter de nombreux dispositifs clients tels que les ordinateurs personnels clients représentés d'une manière schématique par la référence numérique 24.

Le réseau SAN 10 peut également inclure un spoule de mémoire 26, qui peut inclure, par exemple, une pluralité de dispositifs de mémoire physique tels que des unités de disque dur qui sont commandées par un ou plusieurs contrôleurs de mémoire et qui sont reliées à ces derniers. Les dispositifs de mémoire physique du spoule de mémoire 26 peuvent être attribués à des LUNs. Certains dispositifs de mémoire physique peuvent être groupés sous la forme de volumes RAID (*Redundant Array of Inexpensive Disks*, réseau redondant de disques bon marché), à chaque volume étant attribuée une adresse LUN SCSI unique. Un ou plusieurs LUNs peuvent être individuellement attribués à d'autres dispositifs de mémoire physique. Quelle que soit la façon dont les LUNs sont attribués, les LUNs de la figure 1 peuvent mapper la mémoire physique disponible du spoule de mémoire 26 sous la forme d'une pluralité de dispositifs de mémoire logique et permettre à ces dispositifs de mémoire logique d'être identifiés et adressés.

En cours de fonctionnement, les nœuds 16, 18, 20 et 22 peuvent communiquer avec le spoule de mémoire 26 et transférer des données vers le spoule de mémoire 26 et à partir de ce dernier par l'intermédiaire d'une structure 28 en utilisant le protocole Canal de Fibre. Comme décrit sur la figure 1, les nœuds 16 et 18 peuvent être regroupés dans une zone 30 avec LUN_1 et LUN_2. D'une manière

similaire, les nœuds 20 et 22 peuvent être regroupés dans une zone 32 avec LUN_3, LUN_4 et LUN_5. L'utilisation d'une division en zones de commutation afin de créer la zone 30 peut empêcher les nœuds 16 et 18 de voir les nœuds 20 et 22. D'une manière similaire, l'utilisation d'une division en zones de commutation afin de créer la zone 32 peut empêcher les nœuds 20 et 22 de voir les nœuds 16 et 18. En plus de la division en zones, le mode de réalisation de la figure 1 peut utiliser le masquage LUN. Le masquage LUN peut empêcher qu'un nœud ou qu'une grappe spécifique voit certains LUNs. Par exemple, un masquage LUN peut éviter que les nœuds 16 et 18 voient LUN_3, LUN_4 et LUN_5.

Dans le mode de réalisation de la figure 1, les nœuds 16, 18, 20 et 22 peuvent se voir attribuer un nom mondial unique (WWN, *World Wide Name*), qui peut être un identificateur à huit octets. L'IEEE (*Institute of Electrical and Electronic Engineers*, Institut des ingénieurs électroniciens et électriciens) attribue des blocs de WWNs à des fabricants de manière à ce que les fabricants puissent concevoir des dispositifs Canal de Fibre ayant des WWNs uniques. A des fins d'illustration, dans le mode de réalisation de la figure 1, le nœud 16 peut avoir un WWN égal à "AAA", le nœud 18 peut avoir un WWN égal à "BBB", le nœud 20 peut avoir un WWN égal à "CCC" et le nœud 22 peut avoir un WWN égal à "DDD". Ainsi, les nœuds 16, 18, 20 et 22 peuvent être identifiables d'une manière unique par d'autres dispositifs reliés à la structure 28.

Les nœuds 16, 18, 20 et 22 peuvent posséder des informations d'identification en plus de leurs noms WWN respectifs. Par exemple, selon le protocole Canal de Fibre, lorsqu'un nœud tel que le nœud 16 est initialisé et se connecte à la structure 28, le nœud se voit attribuer un identificateur (ID) de canal de fibre. Cet identificateur peut être soumis à un changement à chaque fois qu'un

certain événement d'initialisation survient, par exemple, lorsqu'un autre nœud ou un autre dispositif se connecte à la structure 28. Comme décrit sur la figure 1, la structure 28 a attribué les identificateurs ID de canal de fibre de la manière suivante : le nœud 16 est S_ID_1, le nœud 18 est S_ID_2, le nœud 20 est S_ID_3 et le nœud 22 est S_ID_4.

Dans le mode de réalisation de la figure 1, les divers WWNs et IDs de canal de fibre peuvent être mémorisés sur un support lisible par ordinateur 34, lequel peut être accessible à des dispositifs du réseau SAN 10. Comme représenté sur la figure 1, le réseau SAN 10 peut inclure un dispositif informatique 38 pour établir la structure 28. Un tel dispositif informatique peut inclure une unité centrale de traitement (CPU, *Central Processing Unit*) reliée d'une manière communicante au support lisible par ordinateur 34. Un commutateur 36 peut également posséder au moins un port 40 pour assurer l'interface avec d'autres dispositifs afin de former un réseau à canal de fibre complet.

Dans un mode de réalisation d'un système incorporant les enseignements de la présente description, un dispositif informatique 38 peut être utilisé pour faire fonctionner un moteur de gestion de ressources, qui peut être mémorisé sur le support lisible par ordinateur 34. Le moteur de gestion de ressources peut être utilisé pour assurer plusieurs fonctions. Par exemple, le moteur de gestion de ressources peut être utilisé pour effectuer un accès à une liste maintenue constituée des WWNs et des IDs de canal de fibre des dispositifs du réseau SAN 10. De plus, le moteur de gestion de ressources peut être utilisé pour reconnaître une commande de réinitialisation SCSI envoyée par un nœud et pour convertir la commande sous la forme d'une commande de libération de ressource mémoire. La commande de libération de ressource mémoire

peut être, par exemple, une déconnexion de traitement tierce ou une commande de sortie de réserve persistante SCSI ayant une action d'effacement.

5 Dans une grappe MSCS classique, une commande de réinitialisation SCSI peut être envoyée lorsqu'un nœud analogue au nœud 18 ou 20 ne réussit pas à accuser réception à temps d'une "pulsation" ou *heartbeat* 42 ou 44 provenant d'une seconde grappe respective. Les pulsations 42 et 44 peuvent permettre aux nœuds 18 et 22 respectivement
10 de "voir" si leur grappe partenaire continue de fonctionner.

Par exemple, si le nœud 18 ne peut plus "voir" le nœud 16, le nœud 18 peut tenter d'obtenir de quelconques réservations LUN maintenues pour le nœud 16 libéré. Pour
15 réaliser cette libération, le nœud 18 peut envoyer une commande de réinitialisation SCSI pour amorcer une réinitialisation de bus de bas niveau des bus SCSI associés aux nœuds 16 et 18. Dans certains systèmes, par exemple un système MSCS, le nœud 18 peut attendre une certaine
20 quantité de temps spécifiée avant de tenter de réserver les LUNs qui étaient réservés par le nœud 16. L'attente permet au nœud 16 d'acquiescer à nouveau la commande des LUNs qui lui ont été réservés avant la réinitialisation SCSI. Ainsi, si le nœud 16 est "vivant" malgré l'échec du
25 nœud 18 à recevoir la pulsation 42, le nœud 16 est capable d'établir à nouveau ses réservations de ressource et de permettre ainsi au nœud 18 de savoir qu'il est "vivant".

Malheureusement, comme mentionné ci-dessus, une
30 réinitialisation SCSI dans un environnement informatique en grappes peut avoir un effet nuisible sur la performance totale du système et la fiabilité totale du système. Le système et le moteur de gestion de ressources décrits peuvent aider de plusieurs manières différentes à
35 limiter le recours de l'environnement informatique en

grappes à des réinitialisations SCSI. Des exemples de techniques permettant d'éviter des réinitialisations SCSI peuvent être mieux compris en étudiant les figures 2 et 3.

5 La figure 2 est un ordinogramme d'un mode de réalisation d'un procédé 100 permettant de gérer des ressources mémoire dans un environnement informatique en grappes. Le procédé de la figure 2 peut être implémenté par un moteur de gestion de ressources s'exécutant sur un
10 contrôleur de mémoire attaché à une structure SAN. Dans certains modes de réalisation, le moteur de gestion de ressources peut s'exécuter sur une CPU associée à un commutateur analogue au commutateur 36 de la figure 1. Dans d'autres modes de réalisation, la CPU peut être associée
15 à un dispositif SAN autre que le commutateur. Par exemple, un moteur de gestion de ressources peut s'exécuter sur un ou plusieurs nœuds d'un réseau SAN.

 Pendant le fonctionnement d'un réseau SAN, une commande de connexion de port (*PLOGI, Port Login*) peut
20 être reçue. Comme ceci est connu dans la technique, une commande *PLOGI* est une commande de canal de fibre dans laquelle un nœud se connecte à un dispositif de mémoire attaché à un réseau SAN. Un nœud peut exécuter une commande *PLOGI* après que la structure ait attribué un ID de
25 canal de fibre (*S_ID*) au nœud. Comme ceci est également connu de manière classique, le *S_ID* d'un nœud peut être attribué lorsqu'un nœud exécute une commande de connexion à la structure (*FLOGI, Port Login*).

 A l'étape 102, l'identificateur *S_ID* et le nom
30 *WWN* d'un nœud de grappe peuvent être extraits. L'extraction peut avoir lieu à des instants différents. Par exemple, l'extraction peut avoir lieu lorsqu'un nœud envoie une commande *PLOGI*. Une fois extraits, le *S_ID* et le *WWN* peuvent être mis à jour et peuvent être mémorisés sur un
35 support lisible par ordinateur. Dans certains modes de

réalisation, ce support lisible par ordinateur peut être une partie d'un réseau SAN et peut être accessible à plusieurs dispositifs du réseau SAN.

5 A l'étape 104, une réservation LUN peut être maintenue pour un nœud donné. En effet, le nœud donné peut avoir le droit exclusif d'utiliser le LUN réservé. Comme ceci est mentionné ci-dessus, les nœuds de grappe communiquent souvent les uns avec les autres en utilisant un signal de pulsation. A l'étape 106, un dispositif SAN
10 peut détecter un échec de réception à temps d'un signal de pulsation. Bien que l'échec de réception d'un signal de pulsation puisse uniquement indiquer un échec de liaison de communication entre l'émetteur de pulsation et le récepteur de pulsation, l'échec peut avoir pour résultat,
15 comme représenté à l'étape 108, la détermination indiquant qu'un nœud de grappe est inutilisable.

Dans le mode de réalisation de la figure 2, la détermination indiquant qu'un nœud est inutilisable peut amener un autre nœud à envoyer une réinitialisation SCSI.
20 Comme représenté à l'étape 110, une commande de réinitialisation SCSI peut être envoyée pour libérer des réservations LUN maintenues pour le nœud considéré comme étant inutilisable (le nœud "mort"). A l'étape 112, la commande de réinitialisation SCSI peut être convertie en une dé-
25 connexion de traitement tierce. Cette conversion peut, par exemple, être effectuée par une exécution du moteur de gestion de ressources.

A l'étape 114, une commande de déconnexion concernant le nœud "mort" peut être envoyée par une tierce
30 partie au nom du nœud "mort". Par exemple, un moteur de gestion de ressources peut effectuer un accès à un support lisible par ordinateur mémorisant l'identificateur S_ID et le nom WWN du nœud "mort". Le moteur de gestion de ressources peut utiliser le S_ID et le WWN du nœud
35 "mort" pour déconnecter le nœud "mort". Cette déconnexion

de traitement tierce peut avoir pour résultat la libération des réservations LUN maintenues pour le nœud déconnecté.

Comme représenté à l'étape 116 de la figure 2, d'autres nœuds d'une grappe peuvent également se déconnecter ou être déconnectés et une réinitialisation de liaison de protocole d'initialisation de boucle (LIP, *Loop Initialization Protocol*) peut être amorcée. La réinitialisation de liaison LIP de l'étape 118 peut être suivie de la génération à l'étape 120 d'une notification de changement d'état. Dans le mode de réalisation de la figure 2, la notification de changement d'état peut amener des nœuds de grappe actifs, des nœuds qui ne sont pas morts, à effectuer une connexion à un port et à tenter d'obtenir des réservations LUN. La connexion de nœuds de grappe actifs à un port peut être observée à l'étape 122. Si le nœud "mort" n'était pas mort, il peut être capable d'acquérir à nouveau ses réservations LUN. Si le nœud "mort" était mort, d'autres nœuds de grappe sont maintenant capables de capturer les réservations LUN maintenues par le nœud "mort". En effet, les ressources mémoire maintenues par le nœud mort vont devenir disponibles pour des nœuds "vivants" - donnant en résultat une meilleure utilisation des ressources mémoire - sans réinitialisation SCSI.

Sur la figure 3, on peut observer un autre mode de réalisation d'un procédé 200 de gestion de ressources mémoire dans un environnement informatique en grappes. Le procédé de la figure 3, comme le procédé de la figure 2, peut être implémenté par un moteur de gestion de ressources. Ce moteur peut être placé dans un nombre quelconque d'emplacements. Par exemple, le moteur peut être placé au niveau d'un commutateur, d'un nœud ou d'un contrôleur de mémoire attaché à une structure Canal de Fibre.

Comme représenté à l'étape 202, le procédé 200

peut impliquer la réception d'une commande de réservation LUN SCSI. Une commande de réservation SCSI classique peut être effacée à l'aide d'une réinitialisation SCSI. Comme mentionné ci-dessus, les réinitialisations SCSI peuvent

5 provoquer de nombreux problèmes au sein d'un environnement informatique en grappes. Ainsi, à l'étape 204, la commande de réserve SCSI peut être convertie en une commande de sortie de réserve persistante SCSI ayant une action de service de RESERVE. La conversion à partir d'une

10 réserve SCSI en une réserve persistante SCSI peut être effectuée, par exemple, par une exécution du moteur de gestion de ressources. La commande de sortie de réserve persistante peut maintenir une réservation LUN persistante comme représenté à l'étape 206 pour le nœud mainte-

15 nu, le nœud à l'origine de la commande de réserve SCSI.

A l'étape 208, on peut déterminer que le nœud maintenu est inutilisable. En réponse à cette détermination, une commande de réinitialisation SCSI peut être envoyée. La commande de réinitialisation SCSI de l'étape

20 210 peut être convertie à l'étape 212 en une commande de réserve persistante SCSI ayant une action de service d'EFFACEMENT. En cours d'exécution, la commande de réserve persistante SCSI ayant une action de service d'EFFACEMENT peut libérer les réservations LUN maintenues par

25 la commande de sortie de réserve persistante SCSI initiale. La libération de LUN de l'étape 214 peut effectivement libérer des ressources mémoire maintenues par des nœuds déterminés comme étant inutilisables à l'étape 208. Ceci peut avoir pour résultat une meilleure utilisation

30 de ressources mémoire au sein d'un environnement informatique en grappes, et cette meilleure utilisation peut être réalisée sans utiliser des réinitialisations SCSI.

La présente invention inclut diverses variantes des modes de réalisation décrits ci-dessus. Par exemple,

35 ces modes de réalisation peuvent être implémentés au sein

de réseaux SAN ayant un nombre quelconque de topologies.
Il peut y avoir, par exemple, de nombreux contrôleurs de
mémoire, un moteur de gestion de ressources s'exécutant
sur chaque nœud d'une grappe, ou un moteur de gestion de
5 ressources unique s'exécutant dans chaque zone d'un envi-
ronnement informatique en grappes.

REVENDICATIONS

1. Procédé de gestion de ressources mémoire dans
5 un environnement informatique en grappes, le procédé
étant caractérisé en ce qu'il comporte les étapes consis-
tant à :

- maintenir une réservation sur une ressource mémoire
pour un premier nœud de l'environnement informatique
10 en grappes,
- effectuer une déconnexion de traitement tierce pour le
premier nœud, et
- libérer la réservation maintenue pour le premier nœud.

15 2. Procédé selon la revendication 1, caractérisé
en ce qu'il comporte en outre l'étape consistant à :
- déterminer que le premier nœud ne fonctionne pas.

3. Procédé selon la revendication 1, caractérisé
20 en ce qu'il comporte en outre les étapes consistant à :
- échouer à recevoir un signal de pulsation en prove-
nance du premier nœud, et
- déterminer que le premier nœud ne fonctionne pas en
résultat de l'échec de réception du signal de pulsa-
25 tion.

4. Procédé selon la revendication 1, caractérisé
en ce qu'il comporte en outre l'étape consistant à :
- amorcer la déconnexion de traitement tierce à partir
30 d'un second nœud de l'environnement informatique en
grappes.

5. Procédé selon la revendication 1, caractérisé
en ce qu'il comporte en outre l'étape consistant à :
35 - amorcer la déconnexion de traitement tierce à partir

d'un commutateur associé à l'environnement informatique en grappes.

5 6. Procédé selon la revendication 1, caractérisé
en ce qu'il comporte en outre l'étape consistant à :
- créer une zone à l'intérieur d'un système de mémorisation Canal de Fibre, la zone comportant le premier
nœud de l'environnement informatique en grappes, un
second nœud de l'environnement informatique en grappes,
10 et au moins une ressource mémoire.

7. Procédé selon la revendication 1, caractérisé
en ce que la ressource mémoire comporte une unité de disque dur.

15 8. Procédé selon la revendication 1, caractérisé
en ce qu'il comporte en outre les étapes consistant à :
- mémoriser des informations d'identification concernant
le premier nœud, et
20 - utiliser les informations d'identification pour envoyer une demande de déconnexion de traitement tierce.

9. Procédé selon la revendication 1, caractérisé
en ce qu'il comporte en outre les étapes consistant à :
25 - attribuer un nom mondial et un identificateur source
au premier nœud, et
- utiliser le nom mondial et l'identificateur source
pour envoyer une demande de déconnexion de traitement
tierce au nom du premier nœud.

30 10. Procédé selon la revendication 1, caractérisé
en ce qu'il comporte en outre les étapes consistant à :
- attribuer un nom mondial et un identificateur de
source au premier nœud,
35 - attribuer un second nom mondial et un second identifi-

cateur de source à un second nœud,

- utiliser le nom mondial et l'identificateur de source pour envoyer une demande de déconnexion de traitement tierce pour le premier nœud,
- 5 - déconnecter le premier nœud, et
- déconnecter le second nœud.

11. Procédé selon la revendication 1, caractérisé en ce que le premier nœud est constitué d'un serveur.

10

12. Procédé selon la revendication 1, caractérisé en ce que le premier nœud est constitué d'une station de travail.

15

13. Procédé selon la revendication 1, caractérisé en ce qu'il comporte en outre l'étape consistant à :

- déconnecter un second nœud de l'environnement informatique en grappes.

20

14. Procédé selon la revendication 13, caractérisé en ce qu'il comporte en outre l'étape consistant à :

- générer une notification de changement d'état à l'aide d'un commutateur associé à l'environnement informatique en grappes.

25

15. Procédé selon la revendication 13, caractérisé en ce qu'il comporte en outre les étapes consistant à :

- amorcer une réinitialisation de liaison de protocole d'initialisation de boucle, et
- 30 - générer une notification de changement d'état en réponse à la réinitialisation de liaison de protocole d'initialisation de boucle.

35

16. Procédé selon la revendication 13, caractérisé

sé en ce qu'il comporte en outre l'étape consistant à :

- en réponse à la notification de changement d'état, effectuer une connexion à un port de réponse à l'aide du second nœud.

5

17. Procédé de gestion de ressources mémoire dans un environnement informatique en grappes, le procédé étant caractérisé en ce qu'il comporte les étapes consistant à :

- 10 - recevoir une commande de réservation SCSI qui cherche à réserver une ressource mémoire pour un nœud de l'environnement informatique en grappes, et
- en réponse à la commande de réservation, envoyer une commande de sortie de réserve persistante SCSI ayant
- 15 une action de service de réserve afin de réserver la ressource mémoire pour le nœud.

18. Procédé selon la revendication 17, caractérisé en ce qu'un pilote de miniport reçoit la commande de

20 réservation et envoie la commande de sortie de réserve persistante.

19. Procédé selon la revendication 17, caractérisé en ce qu'il comporte en outre l'étape consistant à :

25 - libérer une réservation maintenue pour le nœud en envoyant une commande de sortie de réserve persistante SCSI ayant une action de service d'effacement.

20. Système informatique, caractérisé en ce qu'il

30 comporte :

- un premier nœud (16 ou 18, 20 ou 22) d'un environnement informatique en grappes,
- un second nœud (16 ou 18, 20 ou 22) de l'environnement informatique en grappes, et
- 35 - un moteur de gestion de ressources pouvant être utili-

sé pour convertir une commande de réinitialisation SCSI en une commande de libération de ressource mémoire.

5 21. Système selon la revendication 20, caractérisé en ce que la commande de libération de ressource comporte une déconnexion de traitement tierce.

10 22. Système selon la revendication 20, caractérisé en ce que la commande de libération de ressource comporte une commande de sortie de réserve persistante SCSI ayant une action d'effacement.

15 23. Système selon la revendication 20, caractérisé en ce qu'il comporte en outre :
- un support lisible par ordinateur (34) mémorisant le moteur de gestion de ressources, et
- une unité centrale de traitement (38) reliée d'une manière communicante au support lisible par ordinateur (34) et pouvant être utilisée pour faire fonctionner
20 le moteur de gestion de ressources.

 24. Système selon la revendication 23, caractérisé en ce qu'il comporte en outre :
25 - une pluralité de plates-formes informatiques reliées d'une manière communicante au premier nœud (16 ou 18, 20 ou 22),
- un commutateur (36) Canal de Fibre relié d'une manière communicante au premier nœud (16 ou 18, 20 ou 22), et
30 - une pluralité de dispositifs de mémoire reliés d'une manière communicante au commutateur (36) Canal de Fibre.

 25. Système selon la revendication 24, caractérisé en ce que le commutateur (36) Canal de Fibre comporte
35

l'unité centrale de traitement.

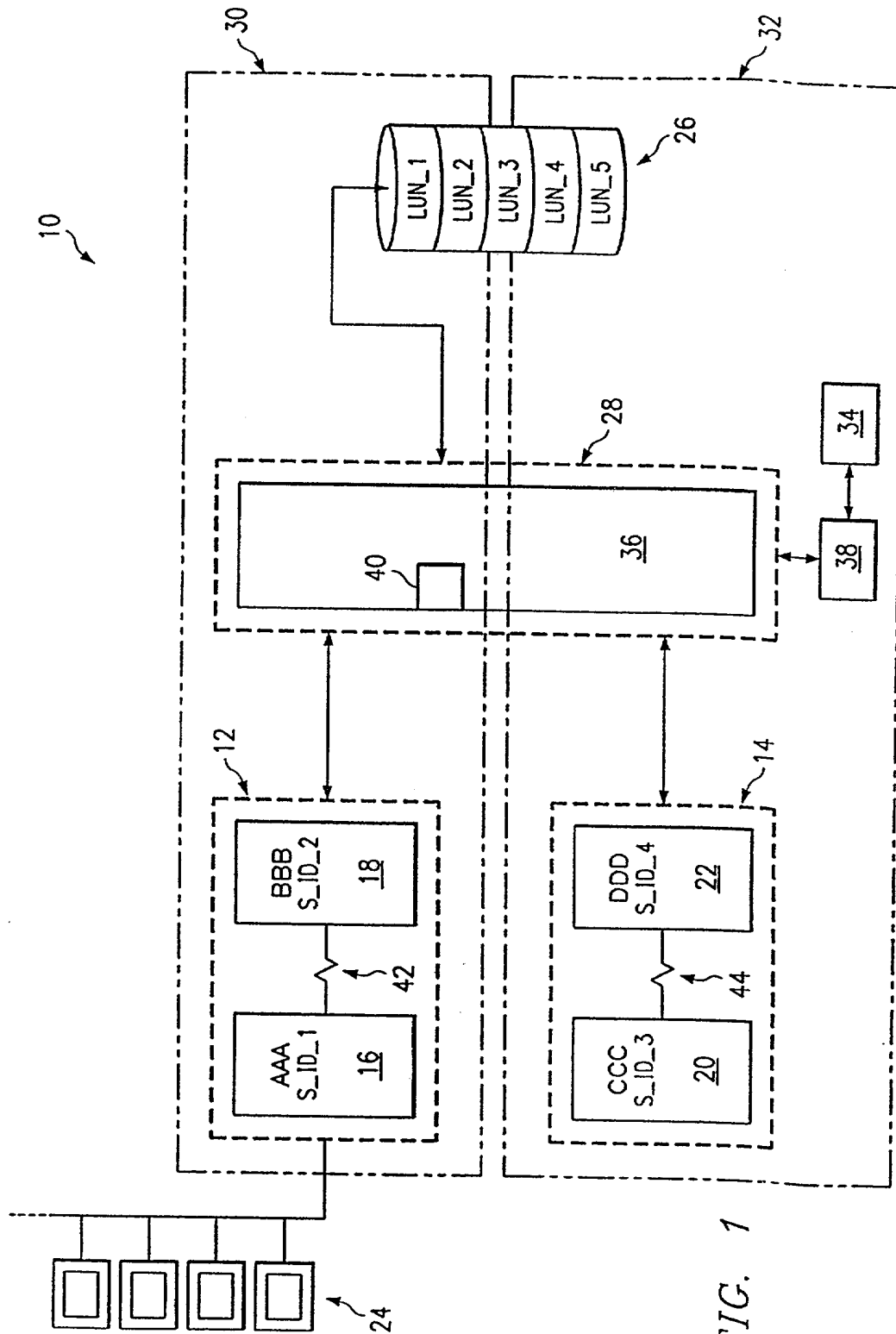


FIG. 1

2/2

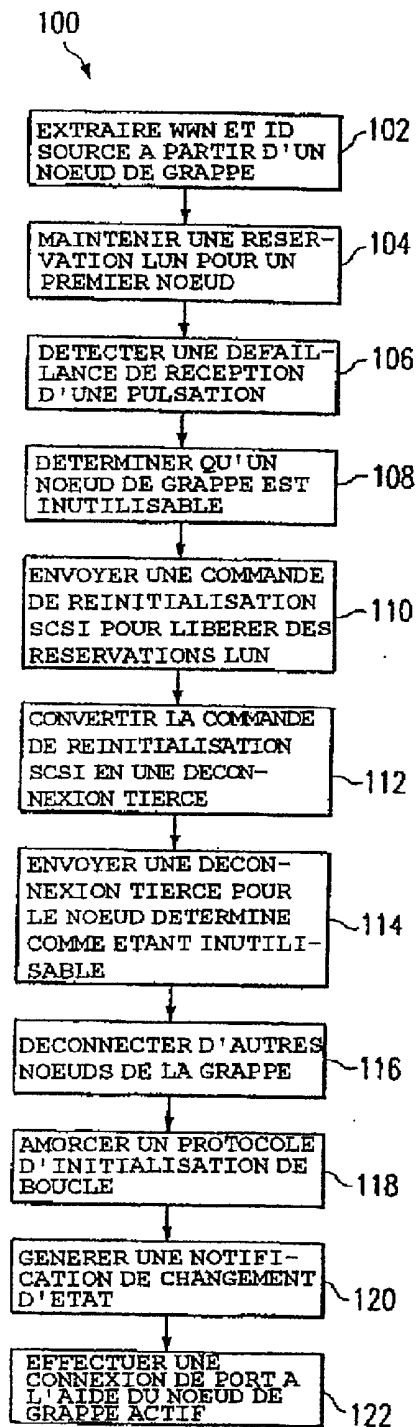


FIG. 2

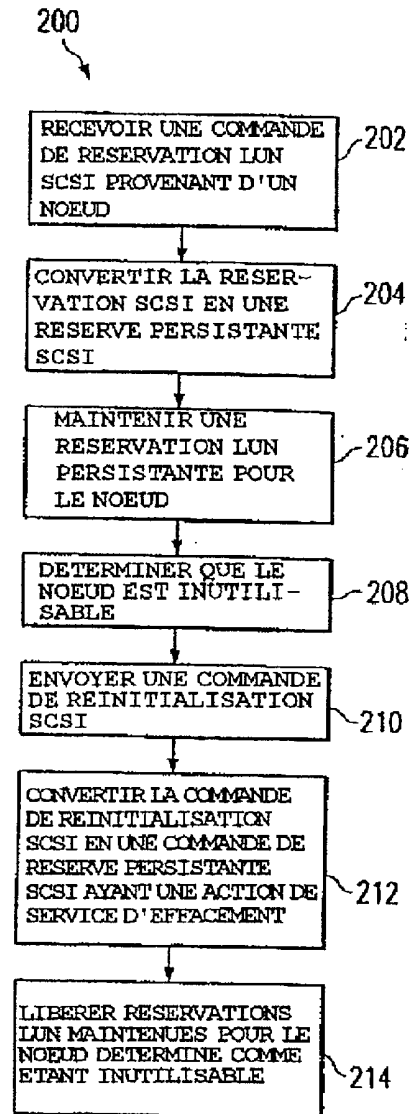


FIG. 3