Guide des notions fondamentales de Sun Cluster 3.1 10/03

Sun Microsystems, Inc.
4150 Network Circle
Santa Clara, CA 95054
U.S.A.

Référence : 817–4260–10
Novembre 2003, Version A

Copyright 2003 Sun Microsystems, Inc. 4150 Network Circle, Santa Clara, CA 95054 U.S.A. Tous droits réservés.

031217@7518

Table des matières

Préface

Le *Guide des notions fondamentales de Sun™ Cluster 3.1* contient des informations théoriques et référentielles relatives au système SunPlex™. Le système SunPlex comprend tous les composants matériels et logiciels de la solution cluster de Sun.

Ce document s'adresse aux administrateurs système expérimentés maîtrisant l'utilisation du logiciel Sun Cluster. Ne l'utilisez pas comme un guide de planification ou de pré-vente. Vous devez déjà avoir déterminé vos besoins système et acheté l'équipement et les logiciels appropriés avant de lire ce document.

Pour comprendre les notions fondamentales décrites dans cet ouvrage, il est nécessaire de connaître l'environnement d'exploitation Solaris™ et de maîtriser le logiciel de gestion de volumes utilisé avec le système SunPlex.

Conventions typographiques

Le tableau suivant présente les modifications typographiques utilisées dans ce manuel.

TABLEAU P–1 Conventions typographiques

Type de caractère ou symbole	Signification	Exemple
`AaBbCc123`	Noms de commandes, fichiers, répertoires et messages système s'affichant à l'écran.	Modifiez votre fichier `.login`. Utilisez `ls -a` pour afficher la liste de tous les fichiers. `nom_machine% vous avez reçu du courrier.`

TABLEAU P–1 Conventions typographiques *(Suite)*

Type de caractère ou symbole	Signification	Exemple
AaBbCc123	Ce que vous entrez, par opposition à l'affichage à l'écran	nom_machine% **su** Password:
AaBbCc123	Paramètre substituable de ligne de commande à remplacer par un nom ou une valeur	Pour supprimer un fichier, entrez **rm** *nom_fichier* .
AaBbCc123	Titres de manuels, termes nouveaux ou mis en évidence.	Reportez-vous au chapitre 6 du *Guide de l'utilisateur*. Ces options sont appelées options de *classe*. Vous devez être *superutilisateur* pour effectuer cette action.

Invites du Shell dans les exemples de commandes

Le tableau suivant présente les invites système et les invites de superutilisateur par défaut des shells C, Bourne et Korn.

TABLEAU P–2 Invites de shell

Shell	Invite
Invite en shell C	nom_machine%
Invite du superutilisateur en shell C	nom_machine#
Invite en shell Bourne et Korn	$
Invite de superutilisateur en shell Bourne et Korn	#

Documentation connexe

Objet	Titre	Référence
Installation	*Guide d'installation du logiciel Sun Cluster 3.1 10/03*	817-4254-10
Matériel	*Sun Cluster 3.x Hardware Administration Manual*	817-0168
	La collection relative à l'administration matérielle de Sun Cluster 3.x à l'adresse `http://docs.sun.com/db/coll/1024.1/`	
Services de données	*Sun Cluster 3.1 Data Service Planning and Administration Guide*	817–3305
	Sun Cluster 3.1 Data Services 10/03 Collection à l'adresse `http://docs.sun.com/db/coll/573.11`	
API/ Développement	*Guide des développeurs pour les services de données Sun Cluster 3.1 10/03*	817-4266-10
Administration	*Guide d'administration système de Sun Cluster 3.1 10/03*	817-4248-10
Messages d'erreur	*Sun Cluster 3.1 10/03 Error Messages Guide*	817–0521
Pages de manuel	*Sun Cluster 3.1 10/03 Reference Manual*	817–0522
Notes de version	*Notes de version de Sun Cluster 3.1 10/03*	817-4521-10
	Sun Cluster 3.x Release Notes Supplement	816-3381

Accès à la documentation Sun en ligne

Le site Web docs.sun.com[SM] vous permet d'accéder à la documentation technique Sun en ligne. Vous pouvez le parcourir ou y rechercher un titre de manuel ou un sujet particulier. L'URL de ce site est `http://docs.sun.com`.

Commande de documents Sun

Sun Microsystems offre une sélection de documentation produit imprimée. Pour obtenir une liste de ces documents et savoir comment les commander, consultez la rubrique “Acheter la documentation imprimée” sur le site http://docs.sun.com.

Accès à l’aide

Si vous rencontrez des problèmes lors de l’installation ou de l’utilisation du système SunPlex, contactez votre fournisseur de services et fournissez-lui les informations suivantes :

- votre nom et votre adresse email (le cas échéant) ;
- le nom, l’adresse et le numéro de téléphone de votre société ;
- les numéros de modèle et de série de vos systèmes ;
- le numéro de version de l’environnement d’exploitation (par exemple Solaris 9) ;
- le numéro de version du logiciel Sun Cluster (par exemple, Sun Cluster 3.1).

Pour obtenir des informations sur chaque noeud de votre système, exécutez les commandes suivantes :

Commande	Fonction
`prtconf -v`	Indique la taille de la mémoire système et affiche des informations sur les périphériques.
`psrinfo -v`	Affiche des informations sur les processeurs.
`showrev –p`	Répertorie les patchs installés.
`prtdiag -v`	Affiche des informations de diagnostic sur le système.
`scinstall -pv`	Affiche les informations de version et de package du logiciel Sun Cluster.
`scstat`	Fournit un aperçu ponctuel de l’état du cluster.
`scconf –p`	Présente les informations de configuration du cluster.
`scrgadm -p`	Affiche les informations des ressources installées, des groupes de ressources et des types de ressources.

Gardez également à disposition le contenu du fichier `/var/adm/messages` .

CHAPITRE 1

Introduction et présentation

Le système SunPlex est une solution intégrant matériel et logiciel Sun Cluster, destinée à la création de services évolutifs et hautement disponibles.

Le Guide des notions fondamentales de Sun Cluster 3.1 10/03 présente les notions fondamentales nécessaires aux principaux utilisateurs de la documentation SunPlex. Ces utilisateurs comprennent :

- les fournisseurs de services assurant l'installation et la maintenance du matériel de cluster ;
- les administrateurs système chargés de l'installation, de la configuration et de l'administration du logiciel Sun Cluster ;
- les programmeurs développant des services évolutifs et de basculement pour des applications non incluses avec le logiciel Sun Cluster.

Cet ouvrage complète la documentation SunPlex, destinée à fournir une vue d'ensemble du système SunPlex.

Ce chapitre

- Propose une introduction et une présentation détaillée du système SunPlex.
- Décrit les divers points de vue des utilisateurs de SunPlex.
- Identifie les notions-clés à assimiler avant de travailler sur le système SunPlex.
- Répertorie les notions-clés de la documentation SunPlex contenant les procédures et informations connexes.
- Répertorie les tâches liées au cluster de la documentation contenant les procédures d'exécution de ces tâches.

Introduction au système SunPlex

Le système SunPlex étend l'environnement d'exploitation Solaris à un système d'exploitation de cluster. Un cluster, ou plex, est un ensemble de noeuds en configuration dispersée fournissant au client une vue unique des services et applications réseau, notamment des bases de données, des services Web et des services de fichiers.

Chaque noeud du cluster est un serveur autonome fonctionnant avec ses propres processus. Ceux-ci communiquent entre eux pour former ce qui ressemble (pour un client réseau) à un système unique fournissant les applications, les ressources système et les données aux utilisateurs.

Un cluster présente plusieurs avantages par rapport aux systèmes à serveur unique traditionnels. Il prend notamment en charge les services évolutifs et de basculement, se prête à une évolution modulaire et son coût est relativement peu élevé par rapport aux systèmes de tolérance aux pannes traditionnels.

Le système SunPlex a pour objectif de :

- réduire ou éliminer les temps d'arrêt système liés aux pannes logicielles ou matérielles ;
- assurer la disponibilité des données et applications aux utilisateurs finaux, même en présence d'une panne entraînant normalement l'arrêt d'un système à serveur unique ;
- augmenter le rendement des applications en permettant aux services de se mettre en corrélation avec d'autres processeurs en ajoutant des noeuds au cluster ;
- accroître la disponibilité du système en vous permettant de réaliser des opérations de maintenance sans arrêter la totalité du cluster.

Haute disponibilité et tolérance de pannes

SunPlex est un système *à haute disponibilité* (HA), c'est-à-dire qu'il offre un accès quasi continu aux données et applications.

Les systèmes *à tolérance de pannes* offrent quant à eux un accès permanent aux données et applications, mais leur coût est plus élevé du fait de l'utilisation de matériel spécialisé. En outre, les systèmes à tolérance de pannes ne prennent généralement pas en compte les pannes logicielles.

Le système SunPlex offre une haute disponibilité, grâce à une combinaison de matériel et de logiciels. Des interconnexions de cluster redondantes, des dispositifs de stockage et des réseaux publics protègent des points de panne uniques. Le logiciel de cluster

contrôle en permanence l'état des noeuds membres et empêche les noeuds défectueux d'agir dans le cluster afin d'éviter les altérations de données. Le cluster contrôle aussi les services et les ressources système dépendantes de ceux-ci, et bascule ou redémarre les services en cas de panne.

Reportez-vous à la rubrique « Questions récurrentes concernant la haute disponibilité » à la page 89 pour prendre connaissance des questions et réponses relatives à la haute disponibilité.

Basculement et évolutivité dans le système SunPlex

Le système SunPlex vous permet d'implémenter soit des services de *basculement*, soit des services *évolutifs*. En général, un service de basculement ne procure qu'une haute disponibilité (redondance), tandis qu'un service évolutif offre une haute disponibilité ainsi que des performances optimisées. Un même cluster peut à la fois prendre en charge des services de basculement et des services évolutifs.

Services de basculement

Le basculement désigne le processus par lequel le cluster déplace automatiquement un service d'un noeud principal défaillant vers un noeud secondaire désigné. Grâce au basculement, le logiciel Sun Cluster fournit une haute disponibilité.

Lorsqu'un basculement intervient, les clients risquent de subir une brève interruption de service et de devoir se reconnecter après le basculement. Toutefois, ils ne savent pas quel serveur physique fournit le service.

Services évolutifs

Tandis que le basculement est lié à la redondance, l'évolutivité apporte un temps de réponse et un rendement constants, indépendamment de la charge. Un service évolutif exploite les noeuds multiples d'un cluster pour exécuter simultanément une application, fournissant ainsi des performances optimisées. Dans une configuration évolutive, chaque noeud du cluster peut fournir des données et traiter les requêtes du client.

Reportez-vous à la rubrique « Services de données » à la page 57 pour obtenir des informations complémentaires sur les services de basculement et les services évolutifs.

Trois points de vue concernant le système SunPlex

Cette rubrique décrit trois points de vue concernant le système SunPlex, ainsi que les notions-clés et la documentation se rapportant à chaque point de vue. Ces points de vue sont ceux :

- du personnel chargé de l'installation et des prestations matérielles ;
- des administrateurs système ;
- des programmeurs.

Point de vue des prestataires de services matériels

Pour les prestataires de services matériels, le système SunPlex s'apparente à un ensemble de composants matériels disponibles en magasin comprenant des serveurs, des réseaux et des dispositifs de stockage. Ces composants sont tous reliés les uns aux autres à l'aide d'un câble, de sorte que chacun possède une sauvegarde et qu'il n'existe aucun point de panne unique.

Notions-clés : matériel

Les prestataires de services matériels doivent maîtriser les notions suivantes :

- configurations et câblage du matériel de cluster ;
- installation et maintenance (ajout, suppression, remplacement) :
  - des composants de l'interface réseau (adaptateurs, jonctions, câbles) ;
  - des cartes d'interface de disques ;
  - des tableaux de disques ;
  - des lecteurs de disques ;
  - de la console administrative et du périphérique d'accès à la console ;
- paramétrage de la console administrative et du périphérique d'accès à la console.

Références théoriques matérielles proposées

Les rubriques suivantes contiennent le matériel correspondant aux notions-clés précédentes.

- « Noeuds de cluster » à la page 20

Documentation SunPlex applicable

Le document SunPlex suivant comporte les procédures et informations relatives à la partie matérielle :

- *Sun Cluster 3.1 Hardware Collection*

Point de vue de l'administrateur système

Pour l'administrateur système, le système SunPlex s'apparente à un ensemble de serveurs (noeuds) reliés les uns aux autres par des câbles et partageant des dispositifs de stockage. L'administrateur du système voit :

- Un logiciel de cluster spécialisé intégré au logiciel Solaris pour contrôler la connectivité entre les noeuds de cluster.
- Un logiciel spécialisé contrôlant l'état des programmes d'application utilisateur exécutés sur les noeuds de cluster.
- Un logiciel de gestion des volumes définissant et administrant les disques.
- Un logiciel de cluster spécialisé permettant à tous les noeuds d'accéder à tous les dispositifs de stockage, même à ceux n'étant pas directement connectés aux disques.
- Un logiciel de cluster spécialisé, permettant aux fichiers d'apparaître sur chaque noeud comme s'ils étaient localement reliés à ce noeud.

Notions-clés : administration système

Les administrateurs système doivent maîtriser les notions et processus suivants :

- l'interaction entre les composants matériels et logiciels ;
- la procédure générale d'installation et de configuration du cluster, notamment :
  - l'installation de l'environnement d'exploitation Solaris ;
  - l'installation et la configuration du logiciel Sun Cluster ;

    - l'installation et la configuration d'un gestionnaire de volumes ;
    - l'installation et la configuration des logiciels prêts à être clusterisés ;
    - l'installation et la configuration du logiciel de service de données Sun Cluster ;
- les procédures d'administration de cluster pour l'ajout, la suppression, le remplacement et la maintenance des composants matériels et logiciels de cluster ;
- les modifications de configuration pour l'amélioration des performances.

Références théoriques proposées aux administrateurs système

Les rubriques suivantes contiennent le matériel correspondant aux notions-clés précédentes.

- « Interfaces d'administration » à la page 34
- « Structure à haute disponibilité » à la page 36
- « Périphériques globaux » à la page 39
- « Groupes de périphériques de disques » à la page 40
- « Espace de noms global » à la page 44
- « Systèmes de fichiers de cluster » à la page 45
- « Quorum et périphériques de quorum » à la page 52
- « Gestionnaires de volumes » à la page 57
- « Services de données » à la page 57
- « Ressources, groupes de ressources et types de ressources » à la page 70
- « Adaptateurs de réseau public et multi-acheminement sur réseau IP » à la page 82
- Chapitre 4

Documentation SunPlex applicable : administrateur système

Les documents SunPlex suivants comportent les procédures et informations relatives à l'administration système :

- *Guide d'installation du logiciel Sun Cluster 3.1*
- *Guide d'administration système de Sun Cluster 3.1*
- *Sun Cluster 3.1 Error Messages Guide*
- *Notes de version de Sun Cluster 3.1*
- *Sun Cluster 3.1 Release Notes Supplement*

Point de vue du programmeur

Le système SunPlex fournit des *services de données* à des applications telles qu'Oracle, NFS, DNS, le serveur Web Sun™ ONE, le serveur Web Apache et le serveur d'annuaires Sun™ ONE. Les services de données sont créés lors de la configuration d'applications disponibles en magasin afin qu'elles fonctionnent sous le contrôle du

logiciel Sun Cluster. Le logiciel Sun Cluster offre des fichiers de configuration et des méthodes de gestion démarrant, arrêtant et contrôlant les applications. Si vous avez besoin de créer un nouveau service évolutif ou de basculement, vous pouvez utiliser l'API SunPlex et l'API DSET (Data Service Enabling Technologies) pour développer les fichiers de configuration et les méthodes de gestion permettant à son application de fonctionner comme un service de données sur le cluster.

Notions-clés : programmeur

Les programmeurs doivent maîtriser les notions suivantes :

- Les caractéristiques de leur application afin de déterminer si elle doit s'exécuter en tant que service évolutif ou de basculement.
- L'API Sun Cluster, l'API DSET de et le service de données « générique ». Les programmeurs doivent choisir l'outil le mieux adapté à l'écriture de programmes ou de scripts pour configurer leur application pour l'environnement de cluster.

Références théoriques proposées au programmeur

Les rubriques suivantes contiennent le matériel correspondant aux notions-clés précédentes.

- « Services de données » à la page 57
- « Ressources, groupes de ressources et types de ressources » à la page 70
- Chapitre 4

Documentation SunPlex applicable : programmeur

Les documents SunPlex suivants comportent les procédures et informations relatives à la programmation d'applications :

- *Sun Cluster 3.1 Data Services Developer's Guide*
- *Sun Cluster 3.1 Data Services Installation and Configuration Guide*

Tâche système SunPlex

Toutes les tâches liées au système SunPlex requièrent certaines connaissances théoriques. Le tableau suivant présente une vue d'ensemble détaillée des tâches et de la documentation illustrant les étapes d'exécution de ces tâches. Les rubriques du présent ouvrage décrivent le lien entre les notions fondamentales et ces tâches.

TABLEAU 1–1 Liste des tâches : correspondance entre la documentation et les tâches utilisateur

Pour exécuter cette tâche...	Utilisez cette documentation...
Installation du matériel de cluster	*Sun Cluster 3.1 Hardware Collection*
Installation du logiciel Solaris sur le cluster	*Guide d'installation du logiciel Sun Cluster 3.1*
Installation du logiciel Sun™ Management Center	*Guide d'installation du logiciel Sun Cluster 3.1*
Installation et configuration du logiciel Sun Cluster	*Guide d'installation du logiciel Sun Cluster 3.1*
Installation et configuration du logiciel de gestion de volumes	*Guide d'installation du logiciel Sun Cluster 3.1* Votre documentation relative à la gestion de volumes
Installation et configuration des services de données Sun Cluster	*Sun Cluster 3.1 Data Services Installation and Configuration Guide*
Maintenance du matériel de cluster	*Sun Cluster 3.1 Hardware Collection*
Administration du logiciel Sun Cluster	*Guide d'administration système de Sun Cluster 3.1*
Administration du logiciel de gestion de volumes	*Guide d'administration système de Sun Cluster 3.1* ainsi que votre documentation relative à la gestion de volumes
Administration du logiciel d'application	La documentation de l'application
Identification des problèmes et suggestions d'actions aux utilisateurs	*Sun Cluster 3.1 Error Messages Guide*
Création d'un nouveau service de données	*Sun Cluster 3.1 Data Services Developer's Guide*

CHAPITRE 2

Notions-clés : fournisseurs de services matériel

Ce chapitre décrit les notions-clés liées aux composants matériels d'une configuration système SunPlex. Il aborde les sujets suivants :

Les composants matériels du système SunPlex

Ces informations s'adressent en priorité aux fournisseurs de services matériel. Ces notions peuvent aider les fournisseurs de services à mieux comprendre les relations entre les divers composants matériels avant d'installer, de configurer ou d'entretenir le matériel de cluster. Elles peuvent aussi s'avérer utiles pour les administrateurs système qui y trouveront des informations de référence pour l'installation, la configuration et l'administration du logiciel de cluster.

Un cluster comprend plusieurs composants matériels, notamment :

- des noeuds de cluster avec des disques locaux (non partagés) ;
- un stockage multihôte (disques partagés entre les noeuds) ;
- des supports amovibles (bandes et CD) ;

- une interconnexion de cluster ;
- des interfaces de réseau public ;
- des systèmes client ;
- une console administrative ;
- des périphériques d'accès à la console.

Le système SunPlex vous permet de combiner ces composants dans plusieurs types de configurations, décrits dans la rubrique « Exemples de topologies Sun Cluster » à la page 28.

La figure suivante illustre un exemple de configuration de cluster.

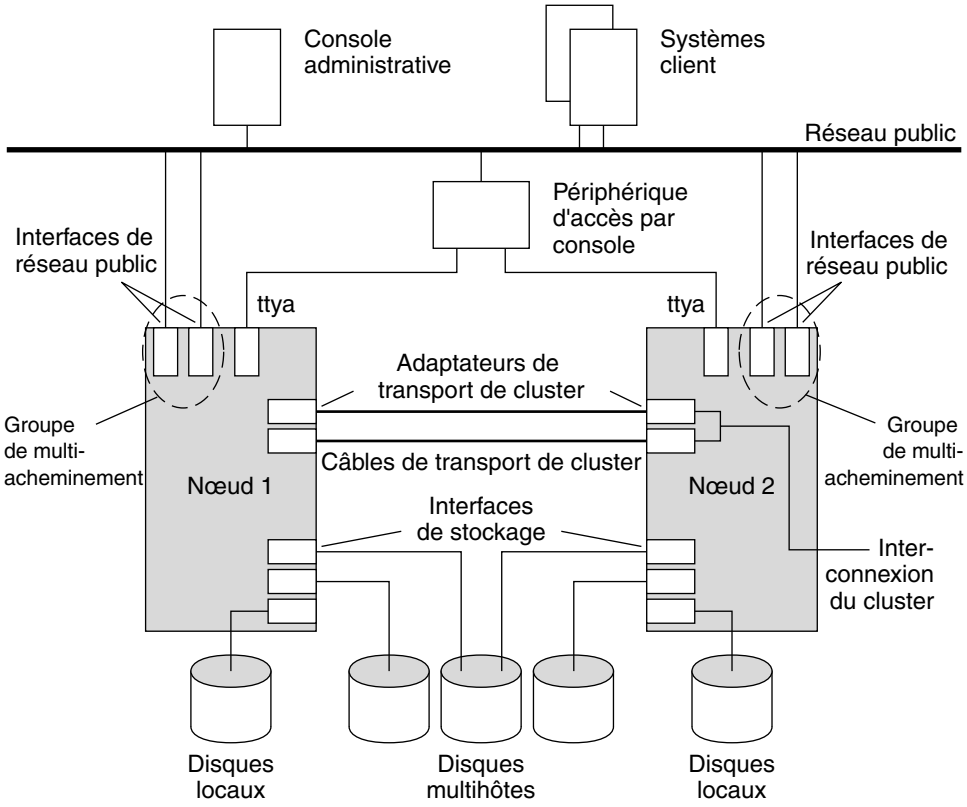


FIGURE 2–1 Exemple de configuration de cluster à deux noeuds

Noeuds de cluster

Un noeud de cluster est une machine exécutant à la fois l'environnement d'exploitation Solaris et le logiciel Sun Cluster, cette machine est soit un membre courant du cluster (*membre du cluster*), soit un membre potentiel. Le logiciel Sun

Cluster permet d'avoir des clusters de deux à huit noeuds. Reportez-vous à la rubrique « Exemples de topologies Sun Cluster » à la page 28 pour obtenir de plus amples informations sur les configurations de noeuds prises en charge.

Les noeuds de cluster sont généralement rattachés à un ou plusieurs disques multihôtes. Ceux qui ne le sont pas utilisent le système de fichiers de cluster pour accéder aux disques multihôtes. Par exemple, dans une configuration de services évolutifs les noeuds peuvent servir les requêtes sans être directement rattachés aux disques multihôtes.

En outre, les noeuds des configurations de bases de données parallèles partagent un accès simultané à tous les disques. Reportez-vous à la rubrique « Disques multihôtes » à la page 23 et au Chapitre 3 pour de plus amples informations sur les configurations des bases de données parallèles.

Tous les noeuds du cluster sont regroupés sous un nom commun, « le nom du cluster », utilisé pour accéder au cluster et le gérer.

Les adaptateurs de réseau public relient les noeuds aux réseaux publics, permettant ainsi au client d'accéder au cluster.

Les membres du cluster communiquent avec les autres noeuds à travers un ou plusieurs réseaux physiquement indépendants. L'ensemble de ces réseaux est appelé *interconnexion de cluster*.

Chaque noeud du cluster sait lorsqu'un autre noeud rejoint ou quitte le cluster. De même, il sait quelles ressources fonctionnent localement et quelles ressources fonctionnent sur les autres noeuds du cluster.

Les noeuds d'un même cluster doivent avoir des capacités de traitement, de mémoire et d'E/S similaires afin que les basculements éventuels s'effectuent sans dégradation significative des performances. En raison de cette possibilité de basculement, chaque noeud doit posséder suffisamment de capacité excédentaire pour prendre la charge des noeuds dont il constitue la sauvegarde ou le noeud secondaire.

Chaque noeud initialise son propre système de fichiers racine (/).

Composants logiciels pour les membres matériels du Cluster

Pour qu'ils puissent fonctionner comme des membres de cluster, les logiciels suivants doivent être installés :

- environnement d'exploitation Solaris ;
- logiciel Sun Cluster ;
- application de service de données ;
- gestion de volumes (Solaris Volume Manager™ ou VERITAS Volume Manager).

Une configuration utilisant un RAID (réseau redondant de disques indépendants) matériel constitue une exception. Cette configuration ne requiert pas forcément de logiciel de gestion de volumes tel que Solaris Volume Manager ou VERITAS Volume Manager.

Reportez-vous au *Guide d'installation du logiciel Sun Cluster 3.1* pour obtenir des informations sur la procédure d'installation de l'environnement d'exploitation Solaris, de Sun Cluster et des logiciels de gestion de volumes.

Reportez-vous au document *Sun Cluster 3.1 Data Services Installation and Configuration Guide* pour obtenir des informations sur la procédure d'installation et de configuration des services de données.

Reportez-vous au Chapitre 3 pour obtenir des informations théoriques sur les composants logiciels précédemment décrits.

La figure suivante fournit une vue d'ensemble des composants logiciels fonctionnant de concert pour créer l'environnement du logiciel Sun Cluster.

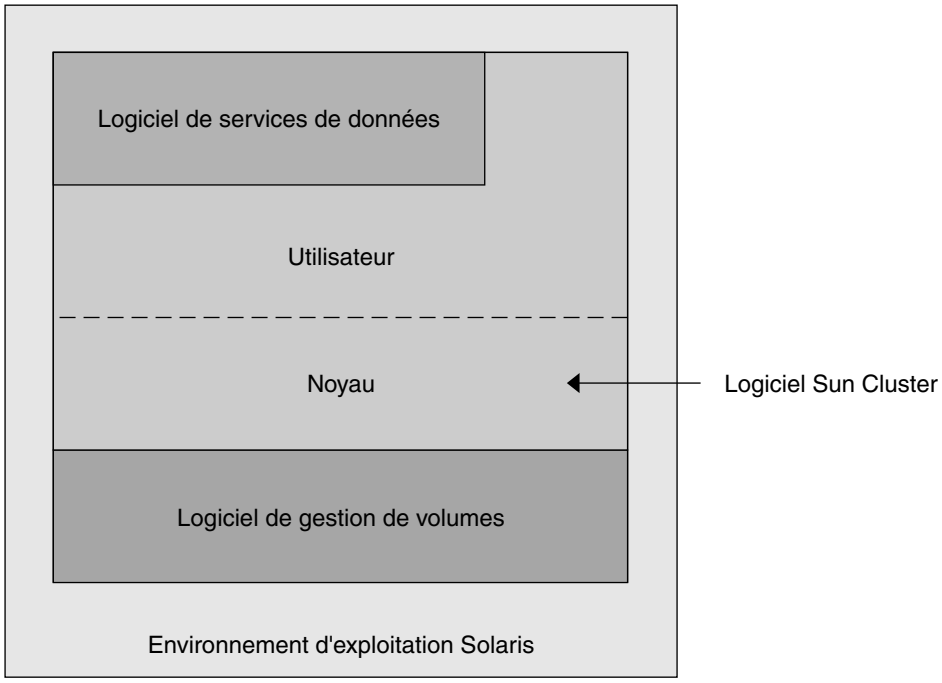


FIGURE 2–2 Relation synthétique entre les composants du logiciel Sun Cluster

Reportez-vous au Chapitre 4 pour prendre connaissance des questions et réponses relatives aux membres du cluster.

Disques multihôtes

Les disques pouvant être connectés à plus d'un noeud à la fois sont appelés disques multihôtes. Dans l'environnement Sun Cluster, le stockage multihôte rend les disques hautement disponibles. Sun Cluster requiert un système de stockage multihôte pour que les clusters à deux noeuds puissent établir un quorum. Les clusters de plus de trois noeuds n'ont pas besoin de stockage multihôte.

Les disques multihôtes possèdent les caractéristiques suivantes :

- Ils peuvent tolérer des pannes sur un seul noeud.
- Ils stockent les données d'applications et peuvent aussi stocker les fichiers binaires et les fichiers de configuration d'applications.
- Ils protègent des pannes de noeud. Si les requêtes client accèdent aux données via un noeud qui échoue, elles sont basculées vers un autre noeud ayant une connexion directe aux mêmes disques.
- Ils sont accessibles soit globalement, à travers un noeud principal « contrôlant » les disques, soit simultanément et directement, à travers des chemins locaux. Serveur Oracle Parallel est la seule application utilisant actuellement l'accès simultané direct .

Un gestionnaire de volumes fournit aux configurations miroirs ou RAID-5 une redondance de données des disques multihôtes. À l'heure actuelle, Sun Cluster prend en charge les gestionnaires de volumes Solaris Volume Manager™ et VERITAS Volume Manager, ainsi que le contrôleur matériel RDAC RAID-5 sur plusieurs plates-formes RAID matérielles.

La combinaison de disques multihôtes avec l'écriture miroir et l'entrelacement (striping) protège à la fois contre les pannes de noeuds et les pannes de disques individuels.

Reportez-vous au Chapitre 4 pour prendre connaissance des questions et réponses relatives au stockage multihôte.

SCSI multi-initiateur

Cette rubrique s'applique uniquement aux périphériques de stockage SCSI et non au stockage fibre Channel utilisé pour les disques multihôtes.

Sur un serveur autonome, le noeud contrôle les activités du bus SCSI par le biais du circuit de l'adaptateur d'hôte SCSI connectant ce serveur à un bus SCSI particulier. Ce circuit est appelé *initiateur SCSI*, il initie toutes les activités de ce bus SCSI. L'adresse SCSI par défaut des adaptateurs d'hôtes SCSI dans les systèmes Sun est 7.

Dans les configurations en cluster, le stockage est partagé entre plusieurs noeuds du serveur à l'aide de disques multihôtes. Lorsque le stockage de cluster se compose de périphériques SCSI en mode asymétrique ou différentiel, la configuration est appelée SCSI multi-initiateur. Comme son nom l'indique, plusieurs initiateurs SCSI figurent sur le bus SCSI.

Les spécifications SCSI requièrent une adresse SCSI unique pour chaque périphérique du bus SCSI . (L'adaptateur d'hôte est également un dispositif du bus SCSI.) La configuration matérielle par défaut dans un environnement multi-initiateur entraîne un conflit, car tous les adaptateurs d'hôtes sont définis par défaut sur 7.

Pour résoudre ce conflit, sur chaque bus SCSI, laissez l'adresse d'un des adaptateurs d'hôtes sur 7, et définissez tous les autres sur des adresses SCSI non utilisées. Afin de bien répartir ces adresses SCSI « non utilisées », choisissez des adresses non utilisées à l'heure actuelle et non utilisables dans le futur. Par exemple, l'ajout de capacité de stockage par l'installation de nouveaux lecteurs à des emplacements vides crée des adresses non utilisables dans le futur. Dans la plupart des configurations, l'adresse SCSI disponible pour un second adaptateur d'hôte est 6.

Vous pouvez modifier les adresses SCSI sélectionnées pour ces adaptateurs d'hôtes en définissant la propriété `scsi-initiator-id` de la PROM Open Boot (OBP). Elle peut être définie globalement pour un noeud ou individuellement pour chaque adaptateur d'hôte. La procédure de définition d'une propriété `scsi-initiator-id` unique pour chaque adaptateur d'hôte SCSI est décrite dans le chapitre concernant chaque boîtier de disque de la *Sun Cluster 3.1 Hardware Collection*.

Disques locaux

Les disques locaux ne sont connectés qu'à un seul noeud. Ils ne sont donc pas protégés contre les pannes de noeuds (ils ne sont pas hautement disponibles). Toutefois, tous les disques, y compris les disques locaux, sont inclus dans l'espace de noms global et configurés comme des *périphériques globaux*. Ainsi, les disques eux-mêmes sont visibles depuis tous les noeuds de cluster.

Vous pouvez rendre les systèmes de fichiers des disques locaux accessibles aux autres noeuds en les mettant sous un point de montage global. Si un des noeuds sur lequel est monté l'un de ces systèmes de fichiers globaux échoue, tous les noeuds perdent l'accès à ce système de fichiers. L'utilisation d'un gestionnaire de volumes vous permet de mettre ces disques en miroir, évitant ainsi qu'une panne ne rende ces systèmes de fichiers inaccessibles ; les gestionnaires de volumes ne protègent cependant pas des pannes de noeuds.

Reportez-vous à la rubrique « Périphériques globaux » à la page 39 pour de plus amples informations sur les périphériques globaux.

Supports amovibles

Les supports amovibles tels que les lecteurs de bandes et lecteurs de CD sont pris en charge par le cluster. En général, vous installez, configurez et entretenez ces périphériques de la même manière que dans un environnement non clusterisé. Ils sont configurés comme des périphériques globaux dans Sun Cluster, chacun est donc accessible à tous les noeuds du cluster. Reportez-vous à la *Sun Cluster 3.1 Hardware Collection* pour de plus amples informations sur l'installation et la configuration des supports amovibles.

Reportez-vous à la rubrique « Périphériques globaux » à la page 39 pour de plus amples informations sur les périphériques globaux.

Interconnexion de cluster

L'*interconnexion de cluster* est la configuration physique des périphériques utilisés pour le transfert de communications privées du cluster et de celles des services de données entre les noeuds du cluster. L'interconnexion étant fortement sollicitée pour les communications privées du cluster, les performances peuvent être amoindries.

Seuls les noeuds du cluster peuvent être y connectés. Le modèle de sécurité Sun Cluster suppose que seuls les noeuds du cluster ont un accès physique à l'interconnexion.

Tous les noeuds doivent être connectés par l'interconnexion de cluster via au moins deux réseaux ou chemins d'accès redondants physiquement indépendants, afin d'éviter un point de panne unique. Il peut y avoir plusieurs réseaux physiquement indépendants (de deux à six) entre deux noeuds, quels qu'ils soient. L'interconnexion de cluster se compose de trois éléments matériels : les adaptateurs, les jonctions et les câbles.

La liste présentée ci-dessous décrit chacun de ces éléments matériels.

- Adaptateurs : cartes d'interface réseau résidant dans chaque noeud du cluster. Leur nom se compose d'un nom de périphérique immédiatement suivi d'un numéro d'unité physique, par exemple, qfe2. Certains adaptateurs n'ont qu'une connexion physique au réseau, tandis que d'autres, comme la carte qfe, en possèdent plusieurs. En outre, certains contiennent à la fois des interfaces réseau et des interfaces de stockage.

  Un adaptateur réseau à plusieurs interfaces peut entraîner un point de panne unique s'il tombe en panne. Pour assurer une disponibilité maximale, planifiez votre cluster de sorte que l'unique chemin d'accès entre deux noeuds ne dépende pas d'un seul adaptateur réseau.
- Jonctions : commutateurs résidant à l'extérieur des noeuds du cluster. Ils remplissent des fonctions d'intercommunication et de commutation vous permettant de connecter plus de deux noeuds entre eux. Dans un cluster à deux noeuds, vous n'avez pas besoin de jonctions car les noeuds peuvent être

directement connectés l'un à l'autre via des câbles physiques redondants connectés à des adaptateurs redondants sur chaque noeud. Les configurations à plus de deux noeuds requièrent généralement des jonctions.

- Câbles : connections physiques reliant soit deux adaptateurs réseau, soit un adaptateur et une jonction.

Reportez-vous au Chapitre 4 pour prendre connaissance des questions et réponses relatives à l'interconnexion de cluster.

Interfaces de réseau public

Les clients se connectent au cluster via des interfaces de réseau public. Chaque carte d'adaptateur réseau peut être connectée à un ou plusieurs réseaux publics, selon si elle possède plusieurs interfaces matérielles ou non. Vous pouvez définir des noeuds pour qu'ils prennent en charge plusieurs cartes d'interfaces de réseau public configurées de façon à ce qu'elles soient toutes actives, et puissent ainsi se servir mutuellement de sauvegarde en cas de basculement. En cas d'échec d'un des adaptateurs, le logiciel de multi-acheminement sur réseau IP intervient pour basculer de l'interface défectueuse vers un autre adaptateur du groupe.

Aucune remarque d'ordre matériel particulière ne s'applique au clustering pour les interfaces de réseau public.

Reportez-vous au Chapitre 4 pour prendre connaissance des questions et réponses relatives aux réseaux publics.

Systèmes client

Les systèmes client comprennent des stations de travail et autres serveurs accédant au cluster via le réseau public. Les programmes client utilisent des données ou autres services fournis par les applications serveur exécutées sur le cluster.

Les systèmes client ne sont pas hautement disponibles. Par contre, les données et applications du cluster le sont.

Reportez-vous au Chapitre 4 pour prendre connaissance des questions et réponses relatives aux systèmes client.

Périphériques d'accès par console

Vous devez disposer d'un accès par console à l'ensemble des noeuds du cluster. Pour obtenir cet accès, utilisez le concentrateur de terminaux fourni avec le matériel de cluster, le SSP (System Service Processor) sur les serveurs Sun Enterprise E10000™, le contrôleur système sur les serveurs Sun Fire™, ou un autre dispositif ayant accès à `ttya` sur chaque noeud.

Un seul concentrateur de terminaux pris en charge est disponible auprès de Sun et son utilisation n'est pas obligatoire. Il permet d'accéder à `/dev/console` sur chaque noeud, via un réseau TCP/IP. Vous disposez ainsi d'un accès par console pour chaque noeud à partir d'une station de travail distante située n'importe où sur le réseau.

Le SSP (System Service Processor) fournit un accès par console au Serveur Sun Enterprise E10000. Le SSP est une machine sur un réseau Ethernet configurée pour pendre en charge le Serveur Sun Enterprise E10000. C'est la console administrative du Serveur Sun Enterprise E10000. En utilisant les fonctions de la console de réseau Sun Enterprise E10000, toutes les stations de travail du réseau peuvent ouvrir une session de console hôte.

Il existe d'autres méthodes d'accès par console, notamment par d'autres concentrateurs de terminaux, l'accès port série `tip(1)` à partir d'un autre noeud, et des terminaux non intelligents. Vous pouvez utiliser les claviers et moniteurs Sun™ ou d'autres périphériques port série si votre fournisseur de services matériels les prend en charge.

Console administrative

Vous pouvez utiliser une station de travail UltraSPARC ™ dédiée, appelée *console administrative* pour administrer le cluster actif. Habituellement vous utilisez et exécutez des outils d'administration, tels que le Cluster Control Panel (CCP) et le module Sun Cluster pour le produit Sun Management Center™, sur la console administrative. L'option `cconsole` du CCP vous permet de vous connecter à plus d'un noeud en même temps. Pour de plus amples informations sur l'utilisation du CCP, reportez-vous au *Guide d'administration système de Sun Cluster 3.1*.

La console administrative n'est pas un noeud du cluster. Elle permet d'accéder à distance aux noeuds du cluster, soit via le réseau public, soit éventuellement à travers un concentrateur de terminaux basé sur le réseau. Si votre cluster est une plate-forme Sun Enterprise E10000, vous devez avoir la capacité de vous connecter au SSP à partir de la console administrative puis de vous connecter à l'aide de la commande `netcon(1M)`.

Les noeuds sont généralement configurés sans moniteur. Vous accédez ensuite à la console des noeuds à travers une session `telnet` ouverte à partir de la console administrative. Celle-ci est connectée à un concentrateur de terminaux à partir duquel vous accédez au port série du noeud (avec un Serveur Sun Enterprise E10000, vous vous connectez à partir du SSP). Reportez-vous à la rubrique « Périphériques d'accès par console » à la page 26 pour de plus amples informations.

Avec Sun Cluster, il n'est pas nécessaire d'utiliser une console administrative dédiée, bien qu'elle présente les avantages suivants :

- Elle permet une gestion centralisée des clusters en regroupant les outils de gestion et de console sur la même machine.

- Elle peut permettre une résolution plus rapide des problèmes par votre fournisseur de services matériels.

Reportez-vous au Chapitre 4 pour prendre connaissance des questions et réponses relatives à la console administrative.

Exemples de topologies Sun Cluster

Une topologie est le plan de connexion reliant les noeuds aux plates-formes de stockage du cluster. Sun Cluster prend en charge toutes les topologies respectant les directives indiquées ci-dessous.

- Sun Cluster peut prendre en charge un maximum de huit noeuds par cluster, indépendamment des configurations de stockage implémentées.
- Un périphérique de stockage partagé peut être connecté à autant de noeuds qu'il peut en prendre en charge.
- Les périphériques de stockage partagés n'ont pas besoin d'être connectés à tous les noeuds du cluster. Toutefois, ils doivent être connectés à au moins deux noeuds.

Sun Cluster ne requiert pas de topologies spécifiques pour la configuration d'un cluster. Les topologies décrites ci-dessous ont pour but de fournir le vocabulaire nécessaire à une discussion relative à un plan de connexion de cluster, et correspondent à des plans de connexions classiques :

- paires clusterisées ;
- paire+N ;
- N+1 (étoile) ;
- N*N (évolutive).

Les rubriques suivantes présentent des exemples schématiques de chaque topologie.

Topologie de paires clusterisées

Une topologie de paires clusterisées consiste en une ou plusieurs paires de noeuds opérant sous une structure d'administration de cluster unique. Dans cette configuration, les basculements ne se produisent qu'entre deux éléments d'une paire. Toutefois, tous les noeuds sont connectés par l'interconnexion de cluster et fonctionnent sous le contrôle du logiciel Sun Cluster. Cette topologie peut être utilisée pour exécuter une application de base de données parallèle sur une paire, et une application évolutive ou de basculement sur une autre paire.

En utilisant le système de fichiers de cluster, vous pourriez aussi avoir une configuration à deux paires où plus de deux noeuds exécutent un service évolutif ou une base de données parallèle, et ce, même si tous les noeuds ne sont pas directement connectés aux disques contenant les données d'application.

La figure présentée ci-après illustre une configuration de paires clusterisées.

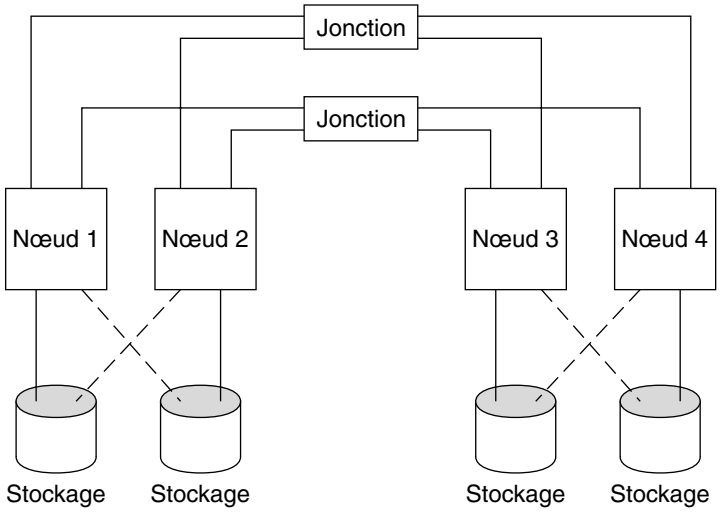


FIGURE 2–3 Topologie de paires clusterisées

Topologie Paire+N

La topologie paire+N comprend une paire de noeuds directement connectés au système de stockage partagé et un ensemble de noeuds utilisant l'interconnexion de cluster pour accéder au système de stockage partagé ; ils n'ont eux-mêmes pas de connexion directe.

La figure présentée ci-après illustre une topologie paire+N où deux des quatre noeuds (le noeud 3 et le noeud 4) utilisent l'interconnexion de cluster pour accéder au système de stockage. Cette configuration peut être étendue et inclure des noeuds supplémentaires n'ayant pas d'accès direct au système de stockage partagé.

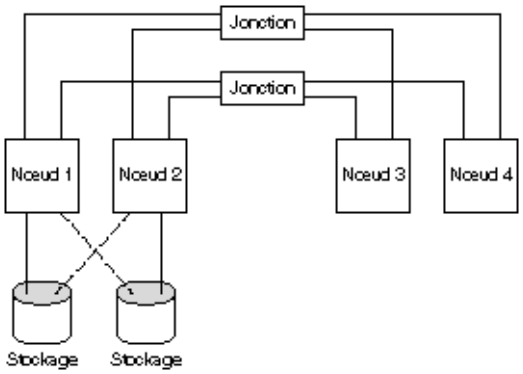


FIGURE 2–4 Topologie paire+N

Topologie N+1 (étoile)

Une topologie N+1 comprend un certain nombre de noeuds principaux et un noeud secondaire. Vous n'avez pas besoin de configurer les noeuds principaux et le noeud secondaire de la même manière. Les noeuds principaux sont activement impliqués dans la fourniture de services d'application. Le noeud secondaire n'a pas besoin d'être inactif en attendant la panne d'un noeud principal.

Le noeud secondaire est le seul noeud de la configuration à être physiquement connecté à tout le système de stockage multihôte.

Si une panne survient sur un noeud principal, Sun Cluster bascule les ressources vers le noeud secondaire, où elles continuent de fonctionner jusqu'à être rebasculées (automatiquement ou manuellement) vers le noeud principal.

Le noeud secondaire doit toujours avoir une capacité d'unité centrale suffisamment importante pour assumer la charge supplémentaire d'un noeud principal défectueux.

La figure présentée ci-après illustre une configuration N+1.

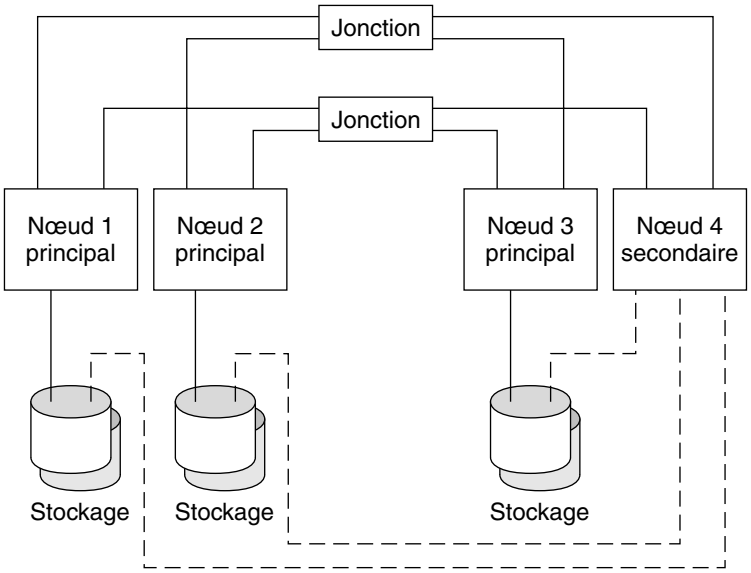


FIGURE 2–5 Topologie N+1

Topologie N*N (évolutive)

Une topologie N*N permet à tous les périphériques de stockage partagés du cluster de se connecter à tous les noeuds du cluster. Elle permet aux applications à haute disponibilité de basculer d'un noeud à l'autre sans dégradation du service. Lorsqu'un basculement a lieu, le nouveau noeud peut accéder au périphérique de stockage via un chemin d'accès local au lieu d'utiliser l'interconnexion privée.

La figure présentée ci-après illustre une configuration N*N.

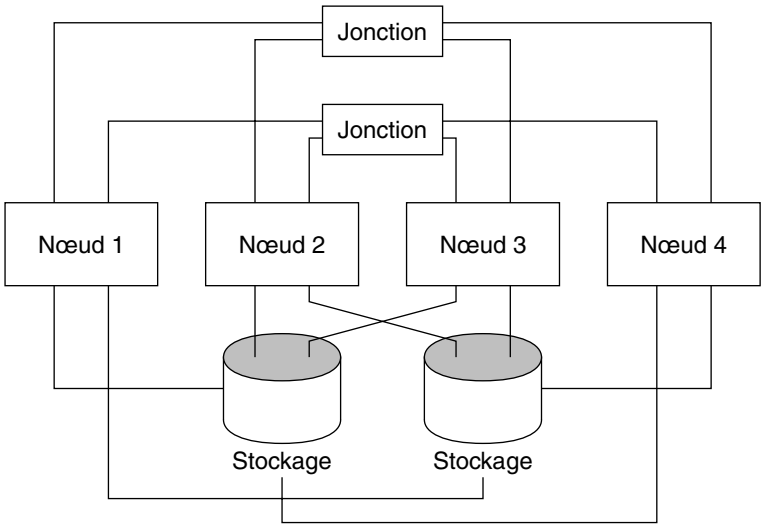


FIGURE 2–6 Topologie N*N

CHAPITRE 3

Notions-clés : administration et développement d'applications

Ce chapitre décrit les notions-clés relatives aux composants logiciels d'un système SunPlex. Il aborde les sujets suivants :

Administration du cluster et développement d'applications

Ces informations sont plus particulièrement destinées aux administrateurs système et aux développeurs d'applications utilisant l'API et le SDK de SunPlex. Les administrateurs système y trouveront des informations de fond pour l'installation, la configuration et l'administration du logiciel de cluster. Les développeurs d'applications y trouveront les informations nécessaires à la compréhension de l'environnement de cluster dans lequel ils travailleront.

La figure suivante présente une vue d'ensemble des correspondances entre les notions fondamentales d'administration du cluster et son architecture.

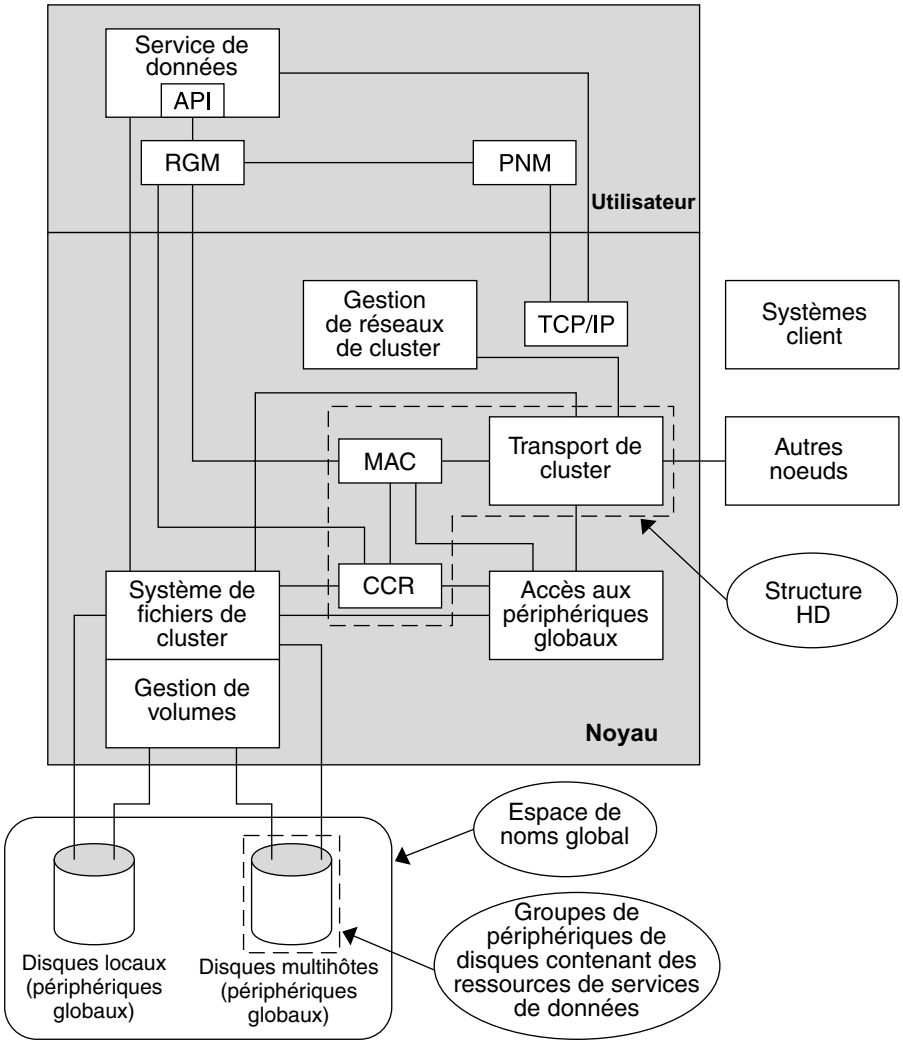


FIGURE 3–1 Architecture du logiciel Sun Cluster

Interfaces d'administration

Vous avez le choix entre plusieurs interfaces utilisateur pour installer, configurer et administrer le système SunPlex. Les tâches d'administration système peuvent être effectuées à travers l'interface utilisateur graphique de SunPlex Manager ou à travers

l'interface de ligne de commande. L'interface de ligne de commande comporte des utilitaires, tels que `scinstall` et `scsetup` permettant de simplifier les tâches d'installation et de configuration sélectionnées. Le système SunPlex contient également un module fonctionnant dans le cadre de Sun Management Center fournissant une interface utilisateur graphique pour effectuer certaines tâches du cluster. Reportez-vous au chapitre d'introduction du *Guide d'administration système de Sun Cluster 3.1* pour obtenir une description complète des interfaces d'administration.

Heure du cluster

L'heure entre tous les noeuds du cluster doit être synchronisée. La synchronisation éventuelle des noeuds du cluster avec une heure extérieure n'a aucune incidence sur son fonctionnement. Le système SunPlex utilise le protocole NTP (Network Time Protocol) pour synchroniser les horloges entre les noeuds.

En général, si l'horloge système est modifiée d'une fraction de seconde, cela ne pose aucun problème. Toutefois, si vous exécutez `date(1)`, `rdate(1M)` ou `xntpdate(1M)` (interactivement ou à l'intérieur de scripts `cron`) sur un cluster en activité, vous pouvez forcer la modification de l'heure de bien plus d'une fraction de seconde pour synchroniser l'horloge système avec la source de synchronisation. Cette modification forcée peut alors entraîner des problèmes avec l'horodateur des modifications de fichiers ou troubler le service NTP.

Lorsque vous installez l'environnement d'exploitation Solaris sur chaque noeud du cluster, vous avez la possibilité de modifier la date et l'heure par défaut du noeud. Vous pouvez en général accepter les paramètres par défaut.

Une des étapes du processus d'installation de Sun Cluster à l'aide de `scinstall(1M)` consiste à configurer NTP pour le cluster. Sun Cluster fournit un fichier de modèles, `ntp.cluster` (consultez `/etc/inet/ntp.cluster` sur un noeud installé) établissant une relation d'homologues entre tous les noeuds du cluster, un noeud étant désigné comme le « préféré ». Les noeuds sont identifiés par leurs noms d'hôtes privés et la synchronisation de l'heure a lieu à travers l'interconnexion de cluster. Vous trouverez les instructions concernant la configuration du cluster pour NTP dans le *Guide d'installation du logiciel Sun Cluster 3.1*.

Vous pouvez aussi installer un ou plusieurs serveurs NTP à l'extérieur du cluster et modifier le fichier `ntp.conf` pour qu'il reflète cette configuration.

Lors d'un fonctionnement normal, vous ne devriez jamais avoir à modifier l'heure du cluster. Toutefois, si celle était mal définie lors de l'installation de l'environnement d'exploitation Solaris et que vous souhaitez la modifier, vous trouverez des directives dans le *Guide d'administration système de Sun Cluster 3.1*.

Structure à haute disponibilité

Le système SunPlex rend tous les composants du « chemin d'accès » reliant les utilisateurs aux données hautement disponibles, y compris les interfaces réseaux, les applications elles-mêmes, le système de fichiers et les disques multihôtes. Un composant du cluster est dit hautement disponible s'il est capable de survivre à toute panne (matérielle ou logicielle) unique du système.

Le tableau suivant présente les types de pannes de SunPlex (matérielles et logicielles) et les types de récupération intégrés à la structure à haute disponibilité.

TABLEAU 3–1 Niveaux de détection et de récupération des pannes de SunPlex

Composant du cluster en panne	Récupération logicielle	Récupération matérielle
Service de données	API HD, structure HD	N/A
Adaptateur de réseau public	multi-acheminement sur réseau IP	Plusieurs cartes d'adaptateur de réseau public
Systèmes de fichiers de cluster	Répliques principales et secondaires	Disques multihôtes
Disque multihôte mis en miroir	Gestion de volumes (Solaris Volume Manager et VERITAS Volume Manager)	RAID-5 matériel (par exemple, Sun StorEdge™ A3x00)
Périphérique global	Répliques principales et secondaires	Plusieurs chemins d'accès au périphérique, jonctions de transport de cluster
Réseau privé	Logiciel de transport HD	Plusieurs réseaux matériels privés indépendants
Noeud	Moniteur d'appartenance au cluster, pilote failfast	Plusieurs noeuds

La structure à haute disponibilité de Sun Cluster détecte rapidement l'échec d'un noeud et crée un nouveau serveur équivalent pour les ressources de la structure sur un autre noeud. Les ressources de la structure ne sont toutes indisponibles à aucun moment. Celles n'ayant pas été affectées par l'échec d'un noeud sont totalement disponibles durant la récupération. En outre, celles du noeud défectueux redeviennent disponibles dès que la récupération est terminée. Une ressource de la structure récupérée n'a pas à attendre la récupération de toutes les autres.

La plupart des ressources de la structure à haute disponibilité sont récupérées de façon transparente pour les applications (services de données) les utilisant. La sémantique d'accès aux ressources de la structure est totalement préservée après l'échec d'un noeud. Les applications ne s'aperçoivent simplement pas que le serveur de la ressource a été déplacé vers un autre noeud. L'échec d'un seul noeud est

totalement transparent pour les programmes sur les autres noeuds utilisant les fichiers, les périphériques et les volumes de disques rattachés à ce noeud, tant qu'il existe un autre chemin d'accès matériel aux disques depuis un autre noeud. On peut par exemple utiliser des disques multihôtes ayant des ports connectés à plusieurs noeuds.

Moniteur d'appartenance au cluster (MAC)

Le moniteur d'appartenance au cluster (MAC) est un ensemble d'agents répartis (un agent par membre du cluster). Ces agents échangent des messages sur l'interconnexion du cluster pour :

- appliquer une vue cohérente de l'appartenance sur tous les noeuds (quorum) ;
- piloter une reconfiguration synchronisée en réponse aux modifications d'appartenance, à l'aide de rappels enregistrés ;
- gérer le partitionnement du cluster (split brain , amnésie) ;
- assurer la connectivité totale entre tous les membres du cluster.

Contrairement à certaines versions antérieures du logiciel Sun Cluster, le MAC s'exécute entièrement dans le noyau.

Appartenance au cluster

La fonction principale du MAC est d'établir des accords au niveau du cluster entre tous les noeuds participant à l'activité du cluster à un moment donné. Cette contrainte s'appelle l'*appartenance au cluster*.

Pour déterminer l'appartenance au cluster et finalement assurer l'intégrité des données, le moniteur d'appartenance au cluster :

- Comptabilise les modifications d'appartenance au cluster, telles qu'un noeud rejoignant ou quittant le cluster.
- Veille à ce qu'un « mauvais » noeud quitte le cluster.
- Veille à ce qu'un « mauvais » noeud reste en dehors du cluster tant qu'il n'a pas été réparé.
- Empêche le cluster de se partitionner en sous-ensembles de noeuds.

Reportez-vous à la rubrique « Quorum et périphériques de quorum » à la page 52 pour obtenir de plus amples informations sur la manière dont le cluster se protège lui-même des partitions en plusieurs clusters.

Reconfiguration du moniteur d'appartenance au cluster

Pour assurer que les données ne s'altèrent pas, tous les noeuds doivent arriver à un accord cohérent sur l'appartenance au cluster. Si nécessaire, le moniteur d'appartenance au cluster coordonne la reconfiguration des services (applications) du cluster en réponse à une panne.

Le MAC reçoit des informations sur la connectivité aux autres noeuds depuis la couche de transport du cluster. Il utilise l'interconnexion du cluster pour échanger des informations d'état au cours d'une reconfiguration.

Après avoir détecté une modification d'appartenance, le MAC réalise une configuration synchronisée du cluster, au cours de laquelle les ressources peuvent être redistribuées en fonction de la nouvelle appartenance au cluster.

Mécanisme failfast

Si le MAC détecte un problème critique sur un noeud, il fait appel à la structure du cluster pour arrêter le noeud de force (panique) et le supprimer de l'appartenance au cluster. Le mécanisme par lequel ce processus intervient est appelé *failfast*. Il provoque l'arrêt d'un noeud de deux manières.

- Si un noeud quitte le cluster puis tente de démarrer un nouveau cluster sans avoir de quorum, il est « séparé » pour être empêché d'accéder aux disques partagés. Reportez-vous à la rubrique « Séparation en cas d'échec » à la page 55 pour de plus amples informations sur l'utilisation du mécanisme failfast.
- Si un ou plusieurs démons spécifiques au cluster meurent (`clexecd`, `rpc.pmfd`, `rgmd`, ou `rpc.ed`), la panne est détectée par le MAC et le noeud panique. Lorsque la mort d'un démon du cluster entraîne la panique d'un noeud, un message similaire à celui-ci s'affiche sur la console pour ce noeud :

```
panic[cpu0]/thread=40e60: Failfast: Aborting because "pmfd" died 35 seconds ago.
409b8 cl_runtime:__0FZsc_syslog_msg_log_no_argsPviTCPCcTB+48 (70f900, 30, 70df54, 407acc, 0)
%l0-7: 1006c80 000000a 000000a 10093bc 406d3c80 7110340 0000000 4001 fbf0
```

Après la panique, le noeud peut soit se réinitialiser et tenter de rejoindre le cluster, soit rester sur l'invite de la PROM OpenBoot™ (OBP). L'action retenue est déterminée par la définition du paramètre `auto-boot?` de l'OBP.

Cluster Configuration Repository (CCR)

Le Cluster Configuration Repository (CCR) est une base de données privée sur le l'ensemble du cluster pour le stockage d'informations relatives à la configuration et à l'état du cluster. C'est une base de données distribuée. Chaque noeud maintient une copie complète de la base de données. Le CCR garantit que tous les noeuds ont une vue cohérente du « monde » du cluster. Pour éviter l'altération des données, chaque noeud a besoin de connaître l'état en cours des ressources du cluster.

Le CCR utilise un algorithme de validation à deux phases pour les mises à jour : si une mise à jour ne se déroule pas correctement sur tous les membres du cluster, elle est réexécutée. Le CCR utilise l'interconnexion du cluster pour appliquer les mises à jour distribuées.

Attention – le CCR est constitué de fichiers texte, mais vous ne devez jamais les modifier manuellement. Chaque fichier contient une somme de contrôle enregistrée pour assurer la cohérence entre les noeuds. Une mise à jour manuelle de ces fichiers peut provoquer l'arrêt d'un noeud ou de l'ensemble du cluster.

Le CCR s'appuie sur le moniteur d'appartenance pour garantir qu'un cluster ne fonctionne que si un quorum a été atteint. Il est chargé de vérifier la cohérence des données au sein du cluster, d'effectuer des récupérations lorsque nécessaire et de faciliter les mises à jour des données.

Périphériques globaux

Le système SunPlex utilise des *périphériques globaux* pour fournir un accès hautement disponible dans tout le cluster à tous ses périphériques, à partir de n'importe quel noeud, indépendamment de l'emplacement auquel est physiquement rattaché le périphérique. En général, si un noeud tombe en panne au moment où il fournit un accès à un périphérique global, Sun Cluster trouve automatiquement un autre chemin vers lequel il redirige l'accès. Les périphériques globaux de SunPlex comprennent les disques, les CD et les bandes. Cependant, les disques sont les seuls périphériques globaux multiports pris en charge. Cela signifie qu'actuellement, les CD et bandes ne sont pas des périphériques hautement disponibles. Les disques locaux installés sur chaque serveur ne disposent pas non plus d'accès multiples et ne sont donc pas hautement disponibles.

Le cluster assigne automatiquement un ID unique à chaque disque, CD et bande du cluster. Cela permet un accès consistant à chaque périphérique à partir de n'importe quel noeud du cluster. L'espace de noms du périphérique global se trouve dans le répertoire `/dev/global`. Reportez-vous à la rubrique « Espace de noms global » à la page 44 pour de plus amples informations.

Les périphériques globaux à accès multiples fournissent plus d'un chemin d'accès au périphérique. Les disques multihôtes sont quant à eux rendus hautement disponibles, car ils font partie d'un groupe de périphériques de disques hébergés par plus d'un noeud.

ID de périphérique (IDP)

Sun Cluster gère les périphériques globaux à travers une structure appelée pseudo pilote d'ID de périphériques (IDP). Ce pilote est utilisé pour assigner automatiquement un ID unique à chaque périphérique du cluster, notamment aux disques multihôtes, aux lecteurs de bandes et aux CD.

Le pseudo pilote d'ID de périphériques (IDP) fait partie intégrante de la fonction d'accès aux périphériques globaux du cluster. Il examine tous les noeuds du cluster et élabore une liste de périphériques de disques uniques, assignant à chacun un numéro majeur et mineur unique, consistant sur tous les noeuds du cluster. L'accès aux périphériques globaux se fait à l'aide de l'ID de périphérique unique assigné par le pilote d'IDP, plutôt que par l'ID de périphérique Solaris traditionnelle, par exemple `c0t0d0` pour un disque.

Cette approche assure que toute application accédant aux disques (telle qu'un gestionnaire de volumes ou des applications utilisant des périphériques bruts) utilise un chemin d'accès consistant au sein du cluster. Cette cohérence est particulièrement importante pour les disques multihôtes, car les numéros majeurs et mineurs de chaque périphérique peuvent varier d'un noeud à l'autre, modifiant ainsi les conventions d'attribution de nom des périphériques Solaris. Le noeud 1 peut par exemple considérer un disque multihôte comme `c1t2d0`, et le noeud 2 peut considérer ce même disque complètement différemment, comme `c3t2d0` . Le pilote d'IDP assigne un nom global, tel que d10, que les noeuds utilisent. Ainsi, chaque noeud a une représentation consistante des disques multihôtes.

La mise à jour et l'administration des ID de périphériques se fait à l'aide des commandes `scdidadm(1M)` et `scgdevs(1M)`. Pour de plus amples informations, reportez-vous aux pages de manuel correspondantes.

Groupes de périphériques de disques

Dans le système SunPlex, tous les disques multihôtes doivent être sous le contrôle de Sun Cluster. Vous créez d'abord les groupes de disques du gestionnaire de volumes (ensembles de disques Solaris Volume Manager ou groupes de disques VERITAS Volume Manager) sur les disques multihôtes. Vous enregistrez ensuite les groupes de disques du gestionnaire de volumes comme *groupes de périphériques de disques*. Un groupe de périphériques de disques est un type de périphérique global. En outre, Sun Cluster crée automatiquement un groupe de périphériques de disques bruts pour chaque disque et bande du cluster. Toutefois, ces groupes de périphériques du cluster restent à l'état hors ligne tant que vous ne les utilisez pas comme périphériques globaux.

L'enregistrement indique au système SunPlex pour chaque noeud les détails de chemin d'accès aux groupes de disques du gestionnaire de volumes. Ceux-ci deviennent alors globalement accessibles au sein du cluster. Si plus d'un noeud peut

écrire sur (contrôler) un groupe de périphériques de disques, les données stockées sur ce groupe deviennent hautement disponibles. Le groupe de périphériques de disques à haute disponibilité peut être utilisé pour héberger les systèmes de fichiers de cluster.

Remarque – les groupes de périphériques de disques sont indépendants des groupes de ressources. Un noeud peut contrôler un groupe de ressources (représentant un groupe de processus de services de données) tandis qu'un autre peut contrôler le(s) groupe(s) de disques auxquels accèdent les services de données. Toutefois, la meilleure façon de procéder est de garder sur le même noeud le groupe de périphériques stockant des données d'application particulières et le groupe de ressources contenant les ressources de l'application (le démon de l'application). Reportez-vous au chapitre de présentation du document *Sun Cluster 3.1 Data Services Installation and Configuration Guide* pour de plus amples informations sur l'association de groupes de périphériques de disques et de groupes de ressources.

Dans un groupe de périphériques de disques, le groupe de disques du gestionnaire de volumes devient « global » parce qu'il fournit un support multivoie aux disques sous-jacents. Chaque noeud du cluster physiquement relié aux disques multihôtes fournit un chemin d'accès au groupe de périphériques de disques.

Basculement du groupe de périphériques de disques

Un boîtier de disque étant connecté à plus d'un noeud, tous les groupes de périphériques de ce boîtier sont accessibles via un autre chemin en cas d'échec du noeud contrôlant le groupe de périphériques. Cette panne n'affecte pas l'accès au groupe de périphériques sauf pendant le laps de temps nécessaire à la récupération et aux contrôles de cohérence. Durant ce laps de temps, toutes les requêtes sont bloquées (de manière transparente pour l'application) jusqu'à ce que le système rende disponible le groupe de périphériques.

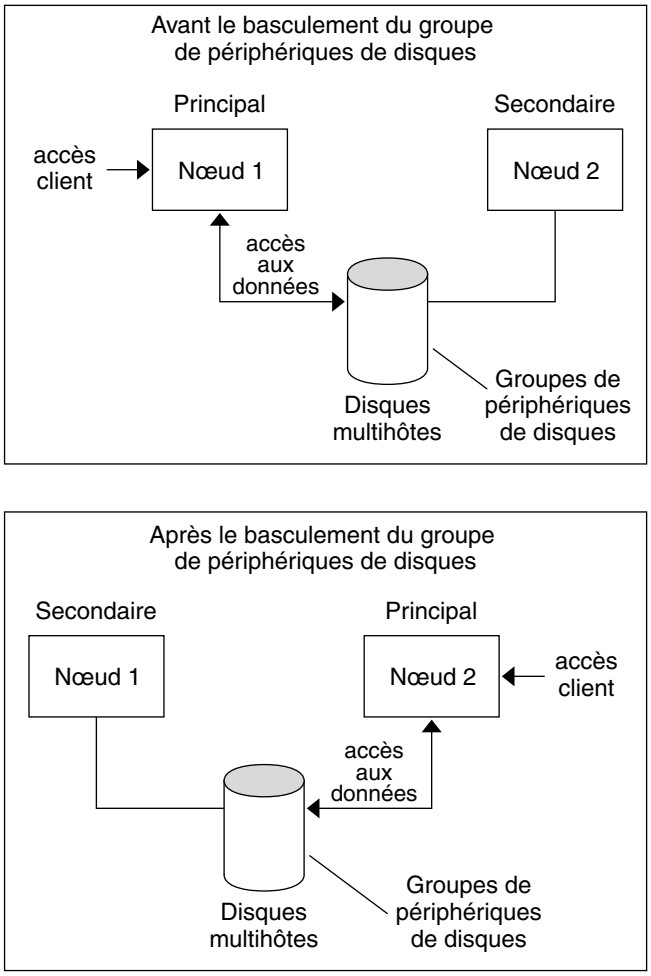


FIGURE 3–2 Basculement du groupe de périphériques de disques

Groupes de périphériques de disques à accès multiples

Cette rubrique décrit les propriétés des groupes de périphériques de disques vous permettant d'équilibrer performances et disponibilité dans une configuration de disques à accès multiples. Le logiciel Sun Cluster propose deux propriétés pour la configuration de disques à accès multiples : `prédilection` et `nombre_noeuds_secondaires` . La propriété `prédilection` permet de déterminer l'ordre dans lequel les noeuds tentent de prendre le contrôle en cas de basculement. La propriété `nombre_noeuds_secondaires` permet de définir le nombre souhaité de noeuds secondaires d'un groupe de périphériques.

Un système à haute disponibilité est considéré comme défectueux lorsque le noeud principal tombe en panne et qu'aucun noeud secondaire n'est susceptible de prendre le relais. Si un basculement survient alors que la propriété `prédilection` est définie sur `true`, un noeud secondaire est sélectionné en fonction de l'ordre de priorité de la liste de noeuds. La liste de noeuds définit l'ordre dans lequel les noeuds vont tenter d'endosser le rôle de noeud principal ou de passer de noeud de réserve à noeud secondaire. La préférence d'un service de périphérique peut être modifiée de manière dynamique à l'aide de l'utilitaire `scsetup`(1M). La préférence associée aux fournisseurs de services dépendants, par exemple à un système de fichiers global, sera celle du service de périphérique.

Les noeuds secondaires sont contrôlés par le noeud principal au cours du fonctionnement normal. Dans une configuration de disques à accès multiples, le contrôle de chaque noeud secondaire entraîne une dégradation des performances et une surconsommation de mémoire. La prise en charge de noeuds de réserve a été implémentée pour minimiser ces conséquences fâcheuses. Par défaut, votre groupe de périphériques de disques aura un noeud principal et un noeud secondaire. Les autres noeuds disponibles apparaîtront en ligne à l'état de noeud de réserve. En cas de basculement, le noeud secondaire devient principal et le noeud prioritaire sur la liste de noeuds devient secondaire.

Le nombre souhaité de noeuds secondaires peut être égal à n'importe quel nombre entier compris entre un et le nombre de noeuds fournisseurs non principaux opérationnels dans le groupe de périphériques.

Remarque – si vous utilisez Solaris Volume Manager, vous devez créer le groupe de périphériques de disques avant de pouvoir modifier la valeur par défaut de la propriété `nombre_noeuds_secondaires`.

Le nombre souhaité de noeuds secondaires par défaut pour les services de périphériques est un. Le nombre réel de noeuds secondaires maintenu par la structure de réplication correspond au nombre souhaité, à moins que le nombre de noeuds non principaux opérationnels soit inférieur à ce nombre. Si vous ajoutez ou supprimez des noeuds de votre configuration, vous devrez modifier la propriété `nombre_noeuds_secondaires` et vérifier deux fois la liste de noeuds. La maintenance de la liste de noeuds et du nombre souhaité de noeuds secondaires empêche les conflits entre le nombre de noeuds secondaires configurés et le nombre réel autorisé par la structure. L'utilisation conjointe de la commande `scconf`(1M) pour les groupes de périphériques de disques VxVM ou de la commande `metaset` (1M) pour les groupes de périphériques de Solaris Volume Manager et des paramètres des propriétés `prédilection` et `nombre_noeuds_secondaires` permet de gérer l'ajout et la suppression de noeuds de votre configuration. Reportez-vous à la rubrique " Administering Global Devices and Cluster File Systems" in *Sun Cluster 3.1 System Administration Guide* pour de plus amples informations sur la procédure de modification des propriétés des groupes de périphériques de disques.

Espace de noms global

Le mécanisme du logiciel Sun Cluster permettant d'activer les périphériques globaux est appelé *espace de noms global*. Il inclut la hiérarchie `/dev/global/` ainsi que les espaces de noms du gestionnaire de volumes. L'espace de noms global reflète les disques multihôtes et les disques locaux (et tout autre périphérique du cluster, tel que CD et bandes) et fournit aux disques multihôtes plusieurs chemins de basculement. Chaque noeud physiquement connecté aux disques multihôtes fournit un chemin d'accès aux périphériques de stockage à n'importe quel noeud du cluster.

Habituellement, les espaces de noms du gestionnaire de volumes se trouvent dans les répertoires `/dev/md/`*ensemble_disques*`/dsk` (et `rdsk`), pour Solaris Volume Manager ; et dans les répertoires `/dev/vx/dsk/` *groupe_disques* et `/dev/vx/rdsk/` *groupe_disques* pour VxVM. Ces espaces de noms correspondent à des répertoires pour chaque ensemble de disques Solaris Volume Manager et chaque groupe de disques VxVM respectivement importés dans le cluster. Chacun de ces répertoires abrite un noeud de périphérique pour chaque métapériphérique ou volume de cet ensemble ou groupe de disques.

Dans le système SunPlex, chaque noeud de périphérique de l'espace de noms du gestionnaire de volumes local est remplacé par un lien symbolique vers un noeud de périphérique du système de fichiers `/global/.devices/node@` *ID_noeud*, où *ID_noeud* est un nombre entier représentant les noeuds du cluster. Les périphériques du gestionnaire de volumes continuent d'être représentés comme des liens symboliques par Sun Cluster, même à leur emplacement standard. L'espace de noms global et l'espace de noms du gestionnaire de volumes standard sont tous deux disponibles à partir de n'importe quel noeud du cluster.

L'espace de noms global présente les avantages suivants :

- Chaque noeud demeure pratiquement indépendant et le modèle d'administration du périphérique est légèrement modifié.
- Les périphériques peuvent devenir globaux de manière sélective.
- Les générateurs de liens de tiers demeurent opérationnels.
- À partir d'un nom de périphérique local donné, on peut facilement établir une correspondance pour obtenir son nom global.

Exemple d'espaces de noms locaux et globaux

Le tableau suivant montre les correspondances entre les espaces de noms locaux et globaux d'un disque multihôte, `c0t0d0s0`.

TABLEAU 3–2 Correspondances entre les espaces de noms locaux et globaux

Composant/Chemin	Espace de noms du noeud local	Espace de noms global
Nom logique Solaris	`/dev/dsk/c0t0d0s0`	`/global/.devices/node@`*ID_noeud* `/dev/dsk/c0t0d0s0`
Nom de l'ID de périphérique	`/dev/did/dsk/d0s0`	`/global/.devices/node@`*ID_noeud* `/dev/did/dsk/d0s0`
Solaris Volume Manager	`/dev/md/`*ensemble_disques*`/dsk/d0`	`/global/.devices/node@`*ID_noeud*`/dev/md/`*ensemble_disques*`/dsk/d0`
VERITAS Volume Manager	`/dev/vx/dsk/`*groupe_disques*`/v0`	`/global/.devices/node@`*ID_noeud*`/dev/vx/dsk/`*groupe_disque*`/v0`

L'espace de noms global est automatiquement généré au moment de l'installation et mis à jour à chaque réinitialisation de configuration. Vous pouvez aussi le générer en exécutant la commande `scgdevs(1M)`.

Systèmes de fichiers de cluster

Un système de fichiers de cluster est un proxy entre le noyau d'un noeud, son système de fichiers sous-jacent et le gestionnaire de volumes fonctionnant sur un noeud ayant une connexion physique au(x) disque(s).

Les systèmes de fichiers de cluster dépendent des périphériques globaux (disques, bandes, CD) ayant des connections physiques à un ou plusieurs noeuds. Les périphériques globaux sont accessibles à partir de n'importe quel noeud du cluster à travers le même nom de fichier (par exemple, `/dev/global/`), que le noeud ait ou non une connexion physique au périphérique de stockage. Un périphérique global peut être utilisé comme tout autre périphérique, c'est-à-dire qu'il est possible d'y créer un système de fichiers à l'aide des commandes `newfs` et/ou `mkfs`.

Le système de fichiers peut être monté globalement à l'aide de la commande `mount -g` ou localement en utilisant la commande `mount`.

Les programmes peuvent accéder au fichier d'un système de fichiers de cluster à partir de n'importe quel noeud et à travers le même nom de fichier (par exemple, `/global/foo`).

Un système de fichiers de cluster est monté sur tous les membres du cluster. Il est impossible de le monter sur un sous-ensemble des membres du cluster.

Un système de fichiers de cluster n'est pas un type de système de fichiers à part. Autrement dit, les clients voient le système de fichiers sous-jacent (par exemple, UFS).

Utilisation des systèmes de fichiers de cluster

Dans le système SunPlex, tous les disques multihôtes sont placés dans des groupes de périphériques de disques, tels que des ensembles de disques Solaris Volume Manager, des groupes de disques VxVM ou des disques individuels n'étant pas sous le contrôle d'un gestionnaire de volumes du logiciel.

Pour qu'un système de fichiers de cluster soit hautement disponible, le périphérique de stockage de disques sous-jacent doit être connecté à plus d'un noeud. Ainsi, un système de fichiers local (système de fichiers stocké sur le disque local d'un noeud) élaboré à l'intérieur d'un système de fichiers de cluster n'est pas hautement disponible.

Tout comme les systèmes de fichiers classiques, les systèmes de fichiers de cluster peuvent être montés de deux manières :

- **Manuellement** : utilisez la commande `mount` avec les options de montage `-g` ou `-o global` pour monter le système de fichiers du cluster à partir de la ligne de commande, par exemple :

  ```
  # mount -g /dev/global/dsk/d0s0 /global/oracle/data
  ```

- **Automatiquement** : créez une entrée dans le fichier `/etc/vfstab` avec l'option de montage `global` pour monter le système de fichiers du cluster au moment de l'initialisation. Créez ensuite un point de montage sous le répertoire `/global` sur tous les noeuds. Le répertoire `/global` est un emplacement recommandé, mais non obligatoire. Voici une ligne d'échantillon d'un système de fichiers de cluster provenant d'un fichier `/etc/vfstab` :

```
/dev/md/oracle/dsk/d1 /dev/md/oracle/rdsk/d1 /global/oracle/data ufs 2 yes global,logging
```

Remarque – le logiciel Sun Cluster n'impose pas de politique d'attribution de noms particulière pour les systèmes de fichiers de cluster, mais vous pouvez faciliter l'administration en créant un point de montage pour tous les systèmes de fichiers du cluster sous le même répertoire, par exemple `/global/` *groupe_périphériques_disques*. Reportez-vous au *Guide d'installation du logiciel Sun Cluster 3.1* et au *Guide d'administration système de Sun Cluster 3.1* pour de plus amples informations.

Caractéristiques d'un système de fichiers de cluster

Le système de fichiers de cluster possède les caractéristiques suivantes :

- Les emplacements d'accès aux fichiers sont transparents. Un processus peut ouvrir un fichier situé n'importe où dans le système et les processus sur tous les noeuds peuvent utiliser le même nom de chemin d'accès pour localiser un fichier.

Remarque – lorsqu'un système de fichiers de cluster lit des fichiers, il ne met pas à jour l'horaire d'accès sur ces fichiers.

- Des protocoles de cohérence sont utilisés pour préserver la sémantique d'accès aux fichiers UNIX même lorsqu'on accède au fichier simultanément à partir de plusieurs noeuds.
- La mise en mémoire cache extensive est utilisée avec des mouvements d'entrée/sortie de masse sans copie pour déplacer les données des fichiers de manière efficace.
- Le système de fichiers d'un cluster fournit des fonctionnalités de verrouillage de fichiers informatives à haute disponibilité par le biais des interfaces `fcntl(2)`. Les applications tournant sur plusieurs noeuds peuvent synchroniser l'accès aux données en utilisant ces fonctionnalités sur un fichier du système de fichiers de cluster. Les verrous de fichiers sont immédiatement récupérés à partir de noeuds quittant le cluster ou d'applications échouant au verrouillage.
- L'accès aux données est assuré, même en cas de pannes. Les applications ne sont pas affectées par les pannes tant qu'un chemin d'accès aux disques demeure opérationnel. Cette garantie est aussi valable pour l'accès aux disques bruts et pour toutes les opérations du système de fichiers.
- Les systèmes de fichiers du cluster sont indépendants du système de fichiers sous-jacent et du logiciel de gestion de volumes. Les systèmes de fichiers de cluster rendent globaux tous les systèmes de fichiers sur les disques pris en charge.

Type de ressource `HAStoragePlus`

Le type de ressource `HAStoragePlus` est destiné à rendre hautement disponibles les configurations de systèmes de fichiers non globaux tels qu'UFS et VxFS. Il permet d'intégrer votre système de fichiers local à l'environnement Sun Cluster et de le rendre hautement disponible. `HAStoragePlus` fournit des fonctionnalités supplémentaires pour les systèmes de fichiers telles que les contrôles, les montages et les démontages forcés permettant à Sun Cluster de basculer sur des systèmes de fichiers locaux. Pour permettre le basculement, le système de fichiers local doit résider sur des groupes de disques globaux dont les bascules correspondantes sont activées.

Reportez-vous aux chapitres concernant les services de données individuels du document *Data Services Installation and Configuration Guide* ou à la rubrique "Enabling Highly Available Local File Systems" du chapitre 14 "Administering Data Services Resources" pour de plus amples informations sur l'utilisation du type de ressources `HAStoragePlus`.

`HAStoragePlus` peut aussi être utilisé pour synchroniser le démarrage des ressources et des groupes de périphériques de disques dont dépendent les ressources. Pour de plus amples informations, reportez-vous à la rubrique « Ressources, groupes de ressources et types de ressources » à la page 70.

L'option de montage Syncdir

L'option de montage `syncdir` peut être utilisée pour les systèmes de fichiers de cluster dont le système de fichiers sous-jacent est UFS. Toutefois, on constate une amélioration significative des performances lorsque `syncdir` n'est pas spécifié. Si vous spécifiez `syncdir`, les écritures sont garanties comme étant conformes à POSIX. Dans le cas contraire, vous aurez le même comportement qu'avec les systèmes de fichiers NFS. Par exemple, dans certains cas, sans `syncdir`, vous ne découvrirez que l'espace disponible est insuffisant qu'au moment où vous fermez un fichier. Avec `syncdir` (et le système POSIX), l'insuffisance d'espace disponible est découverte au moment de l'opération d'écriture. Il est rare que des problèmes se présentent si vous ne spécifiez pas `syncdir`, nous vous recommandons donc de ne pas le spécifier pour bénéficier des gains de performance.

VxFS n'intègre pas d'option de montage équivalente à l'option `syncdir` du système de fichiers UNIX. VxFS se comporte comme le système de fichiers UNIX lorsque `syncdir` n'est pas spécifiée.

Reportez-vous à la rubrique « Questions récurrentes concernant les systèmes de fichiers » à la page 90 pour consulter les questions fréquentes concernant les périphériques globaux et les systèmes de fichiers de cluster.

Contrôle de chemins de disques

La version actuelle du logiciel Sun Cluster prend en charge le contrôle de chemins de disques (CCD). Cette rubrique donne des informations théoriques sur le CCD, le démon CCD et les outils d'administration utilisés pour contrôler les chemins d'accès aux disques. Reportez-vous au document *Sun Cluster 3.1 9/03 System Administration Guide* pour de plus amples informations sur les procédures de contrôle, de désactivation du contrôle et de vérification du statut des chemins d'accès aux disques.

Remarque – le CCD n'est pas pris en charge par les noeuds dont la version est antérieure à Logiciel Sun Cluster 3.1 5/03. N'utilisez pas les commandes de CCD au cours d'une mise à niveau progressive. Après la mise à niveau de tous les noeuds, ces derniers doivent être en ligne pour pouvoir utiliser les commandes de CCD.

Présentation générale

Le CCD améliore la fiabilité globale des basculements et commutations en contrôlant la disponibilité du chemin d'accès au disque secondaire. La commande `scdpm` permet de vérifier la disponibilité du chemin d'accès au disque utilisé par une ressource avant sa commutation. Les options intégrées à la commande `scdpm` vous permettent de contrôler les chemins d'accès aux disques d'un seul noeud ou de tous les noeuds du cluster. Reportez-vous à la page de manuel `scdpm`(1M) pour de plus amples informations sur les options de ligne de commande.

Les composants CCD sont installés à partir du package `SUNWscu`. Il est installé au cours de la procédure d'installation standard de Sun Cluster. Reportez-vous à la page de manuel `scinstall` (1M) pour de plus amples informations sur l'interface d'installation. Le tableau suivant décrit l'emplacement d'installation par défaut des composants CCD :

Emplacement	Composant
Démon	`/usr/cluster/lib/sc/scdpmd`
Interface de ligne de commande	`/usr/cluster/bin/scdpm`
Bibliothèques partagées	`/user/cluster/lib/libscdpm.so`
Fichier de statut du démon (créé au moment de l'exécution)	`/var/run/cluster/scdpm.status`

Un démon CCD à multifile tourne sur chaque noeud. Le démon CCD (`scdpmd`) est lancé par un script `rc.d` lorsqu'un noeud s'initialise. Si un problème survient, il est géré par `pmfd` et relancé automatiquement. La liste présentée ci-dessous décrit le fonctionnement de `scdpmd` au moment du démarrage initial.

Remarque – au démarrage, le statut de chaque chemin d'accès au disque est initialisé sur `UNKNOWN`.

---

1. Le démon CCD rassemble les informations relatives aux chemins d'accès aux disques et aux noms des noeuds à partir du fichier d'état précédent ou à partir de la base de données du CCR. Reportez-vous à la rubrique « Cluster Configuration Repository (CCR) » à la page 38 pour de plus amples informations sur le CCR. Une fois le démon CCD lancé, vous pouvez le forcer à lire la liste des disques contrôlés à partir d'un nom de fichier spécifié.
2. Le démon CCD initialise l'interface de communication pour répondre aux requêtes de composants extérieurs au démon, tels que l'interface de ligne de commande.
3. Le démon CCD pingue, toutes les dix minutes, l'état de chaque chemin d'accès aux disques inclus dans la liste contrôlée à l'aide des commandes `scsi_inquiry`. Chaque entrée est verrouillée pour empêcher l'interface de communication d'accéder au contenu d'une entrée en cours de modification.
4. Le démon CCD notifie la structure d'évènement Sun Cluster et enregistre le nouveau statut du chemin à travers le mécanisme `syslogd` (1M) d'UNIX.

Remarque – toutes les erreurs liées au démon sont rapportées par `pmfd` (1M). Toutes les fonctions de l'API renvoient la valeur `0` en cas de succès et `-1` en cas d'échec.

Le démon CCD contrôle la disponibilité du chemin logique rendu visible par des pilotes multivoies tels que MPxIO, HDLM et PowerPath. Les chemins d'accès physiques individuels gérés par ces pilotes ne sont pas contrôlés parce que le pilote multivoie masque les pannes individuelles du démon CCD.

Contrôle de chemins de disques

Cette rubrique présente deux méthodes de contrôle des chemins d'accès aux disques dans le cluster. La première consiste à utiliser la commande `scdpm`. Cette commande permet de contrôler, de désactiver le contrôle ou d'afficher le statut des chemins d'accès aux disques dans le cluster. Elle permet aussi d'imprimer la liste des disques défectueux et de contrôler les chemins d'accès aux disques à partir d'un fichier.

La seconde méthode consiste à utiliser l'interface utilisateur graphique de SunPlex Manager. SunPlex Manager propose une vue topologique des chemins d'accès aux disques contrôlés dans le cluster. Cette vue est mise à jour toutes les 10 minutes afin de donner des informations sur le nombre de pings ayant échoué. Les informations fournies par l'interface utilisateur graphique de SunPlex doivent être utilisées en conjonction avec la commande `scdpm`(1M) pour administrer les chemins d'accès aux disques. Reportez-vous à la rubrique " Administering Sun Cluster With the Graphical User Interfaces" in *Sun Cluster 3.1 9/03 System Administration Guide* pour de plus amples informations sur SunPlex Manager.

Utilisation de la commande `scdpm` pour contrôler les chemins d'accès aux disques

La commande `scdpm`(1M) intègre des commandes d'administration CCD vous permettant d'effectuer les tâches suivantes :

- contrôler un nouveau chemin de disque ;
- désactiver le contrôle d'un chemin de disque ;
- relire les données de configuration à partir de la base de données du CCR ;
- lire les disques pour les contrôler ou désactiver le contrôle depuis un fichier spécifié ;
- rapporter le statut d'un chemin ou de tous les chemins d'accès aux disques dans le cluster ;
- imprimer tous les chemins d'accès aux disques accessibles depuis un noeud.

Pour effectuer des tâches d'administration CCD sur le cluster, exécutez la commande `scdpm`(1M) avec l'argument chemin de disque à partir de n'importe quel noeud actif. L'argument chemin de disque est toujours constitué d'un nom de noeud et d'un nom de disque. Le nom de noeud n'est pas obligatoire et il est défini par défaut sur `all` si aucun nom n'est spécifié. Le tableau présenté ci-après décrit les conventions d'attribution de noms applicables aux chemins d'accès aux disques.

Remarque – l'utilisation du nom du chemin de disque global est fortement conseillée, car il est consistant dans tout le cluster. Le nom du chemin de disque UNIX ne l'est pas. Il peut varier d'un noeud du cluster à l'autre. Il peut par exemple être `c1t0d0` sur un noeud et `c2t0d0` sur un autre. Si vous utilisez les noms des chemins de disques UNIX, exécutez la commande `scdidadm -L` afin d'établir une correspondance entre le nom du chemin UNIX et celui du chemin global avant d'utiliser les commandes CCD. Reportez-vous à la page de manuel `scdidadm`(1M).

TABLEAU 3–3 Exemples de noms de chemins de disques

Type de nom	Exemple de nom de chemin de disque	Description
Chemin de disque global	`schost-1:/dev/did/dsk/d1`	Chemin de disque `d1` sur le noeud `schost-1`
	`all:d1`	Chemin de disque `d1` sur tous les noeuds du cluster
Chemin de disque UNIX	`schost-1:/dev/rdsk/c0t0d0s0`	Chemin de disque `c0t0d0s0` sur le noeud `schost-1`
	`schost-1:all`	Tous les chemins de disques sur le noeud `schost-1`
Tous les chemins de disques	`all:all`	Tous les chemins de disques sur tous les noeuds du cluster

Utilisation de SunPlex Manager pour contrôler les chemins d'accès aux disques

SunPlex Manager vous permet de réaliser les tâches d'administration CCD de base suivantes :

- contrôler un chemin de disque ;
- désactiver le contrôle d'un chemin de disque ;
- afficher le statut de tous les chemins de disques du cluster.

Consultez l'aide en ligne de SunPlex Manager pour de plus amples informations sur la procédure d'administration des chemins de disques avec SunPlex Manager.

Quorum et périphériques de quorum

Les noeuds partageant des données et des ressources, il est important que le cluster ne soit jamais divisé en plusieurs partitions actives en même temps. Le moniteur d'appartenance au cluster (MAC) garantit qu'un seul cluster est en fonctionnement, même si l'interconnexion de cluster est partitionnée.

Des partitions de cluster peuvent provoquer deux types de problèmes : le split brain et l'amnésie. Un split brain se produit lorsque l'interconnexion entre les noeuds du cluster est perdue et que celui-ci se partitionne en sous-clusters, tous persuadés d'être une partition unique. Ce phénomène est dû à des problèmes de communication entre les noeuds du cluster. Une amnésie survient lorsque le cluster redémarre après un arrêt avec des données antérieures au moment de l'arrêt. Cette situation peut apparaître si plusieurs versions des données sont stockées sur le disque et qu'une nouvelle représentation du cluster est lancée alors que la dernière version n'est pas disponible.

Le split brain et l'amnésie peuvent être évités en donnant un vote à chaque noeud et en attribuant une majorité de votes à un autre cluster opérationnel. Une partition dotée d'une majorité de votes possède un *quorum* et est autorisée à fonctionner. Ce mécanisme de votes majoritaires fonctionne bien tant qu'il y a plus de deux noeuds dans le cluster. Dans un cluster à deux noeuds, la majorité est deux. Si ce cluster est partitionné, un vote externe est nécessaire aux partitions pour obtenir un quorum. Ce vote externe est alors fourni par un *périphérique de quorum*. Un périphérique de quorum peut être n'importe quel disque partagé entre les deux noeuds. Les disques utilisés comme périphériques de quorum peuvent contenir des données utilisateur.

Le Tableau 3–4 décrit comment le logiciel Sun Cluster utilise le quorum pour éviter le split brain et l'amnésie.

TABLEAU 3–4 Quorum du cluster et problèmes de split-brain et d'amnésie

Type de partition	Solution du quorum
Split brain	Autorise uniquement la partition (sous-cluster) possédant une majorité de votes à fonctionner sur le cluster (où au plus une partition peut exister avec une telle majorité) ; lorsqu'un noeud a perdu la course au quorum, il panique.
Amnésie	Garantit, lors de l'initialisation, que le cluster possède au moins un noeud faisant partie des derniers membres du cluster (possédant donc les données de configuration les plus à jour).

L'algorithme de quorum fonctionne de manière dynamique : comme ce sont les événements du cluster qui déclenchent son calcul, les résultats peuvent varier au cours de la durée de vie du cluster.

Nombre de votes de quorum

Les noeuds du cluster et les périphériques de quorum votent tous pour former un quorum. Par défaut les noeuds du cluster acquièrent un nombre de vote de quorum égal à un lorsqu'ils sont initialisés et deviennent membres du cluster. Les noeuds peuvent aussi avoir un nombre de vote égal à zéro, par exemple, lorsque le noeud est en cours d'installation ou qu'il a été placé à l'état de maintenance par un administrateur.

Les périphériques de quorum acquièrent des votes de quorum en fonction du nombre de noeuds connectés au périphérique. Lorsqu'un périphérique de quorum est défini, il acquiert un maximum de votes égal à *N*-1 où *N* est le nombre de noeuds connectés au périphérique de quorum. Par exemple, un périphérique de quorum connecté à deux noeuds avec un nombre de votes égal à zéro possède un nombre de quorums égal à un (deux moins un).

Vous pouvez configurer les périphériques de quorum au cours de l'installation du cluster ou ultérieurement à l'aide des procédures décrites dans le *Guide d'administration système de Sun Cluster 3.1*.

Remarque – un périphérique de quorum ne participe au compte de votes que si au moins un des noeuds auquel il est rattaché est un membre du cluster. Ainsi, durant l'initialisation du cluster, un périphérique de quorum ne participe au compte que si au moins un des noeuds auquel il est rattaché s'initialise et était un membre du dernier cluster initialisé lorsqu'il a été arrêté.

Configurations du quorum

Les configurations du quorum dépendent du nombre de noeuds du cluster.

- **Clusters à deux noeuds :** deux votes de quorum sont nécessaires pour former un cluster à deux noeuds. Ces deux votes peuvent venir des deux noeuds du cluster ou juste d'un seul noeud et d'un périphérique de quorum. Néanmoins, un périphérique de quorum doit être configuré dans un cluster à noeuds pour assurer qu'un seul noeud puisse continuer si l'autre tombe en panne.
- **Clusters à plus de deux noeuds** : vous devez spécifier un périphérique de quorum entre chaque paire de noeuds partageant l'accès à un boîtier de stockage de disques. Prenez par exemple un cluster à trois noeuds similaire à celui présenté dans la Figure 3–3. Sur cette figure, le noeudA et le noeudB partagent l'accès au même boîtier de disques, et le noeudB et le noeudC partagent l'accès à un autre boîtier de disques. Il y a un total de cinq votes de quorum, trois venant du noeud et deux des périphériques de quorum partagés entre les noeuds. Un cluster a besoin de la majorité des votes de quorum pour être formé.

  La spécification d'un périphérique de quorum entre chaque paire de noeuds partageant un accès à un boîtier de stockage de disques n'est pas requise par Sun Cluster ni obligatoire. Toutefois, elle peut apporter les votes de quorum requis

lorsqu'une configuration N+1 dégénère en cluster à deux noeuds et que le noeud ayant accès aux deux boîtiers de disques tombe également en panne. Si vous avez configuré des périphériques de quorum entre toutes les paires, le noeud restant peut encore fonctionner comme cluster.

Reportez-vous à la Figure 3–3 pour consulter des exemples de ces configurations.

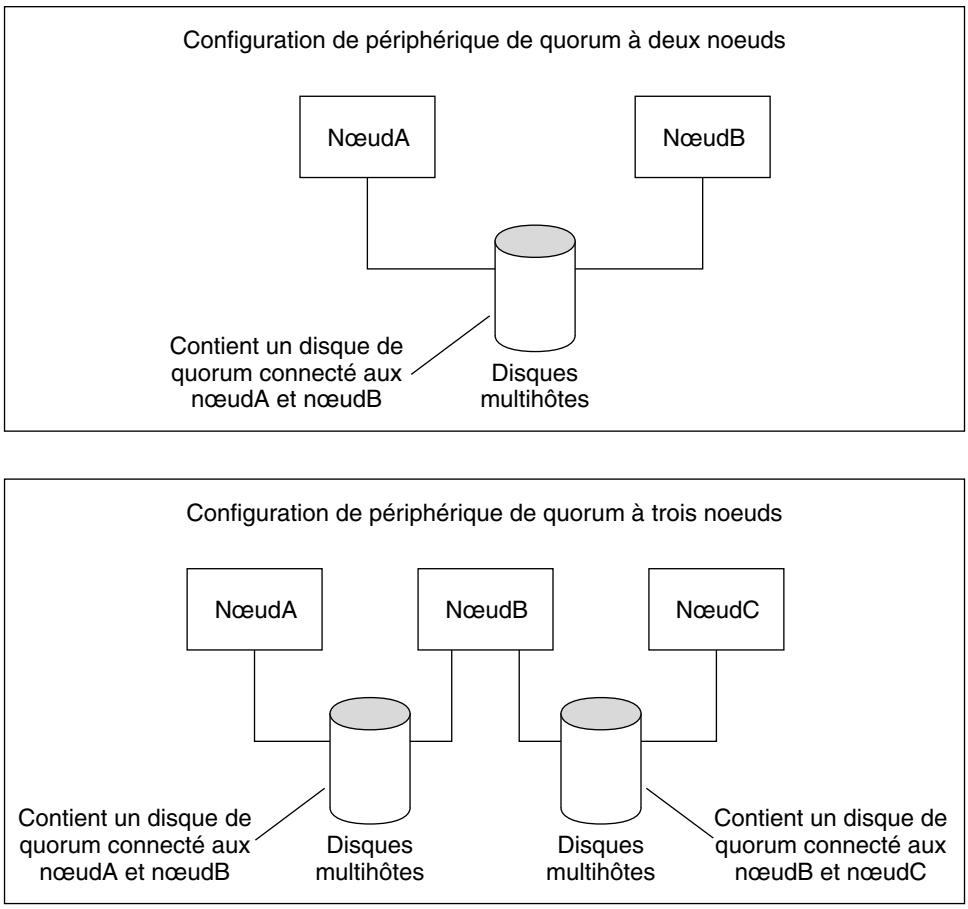


FIGURE 3–3 Exemples de configurations de périphériques de quorum

Directives s'appliquant au quorum

Pour installer des périphériques de quorum, suivez les directives suivantes :

- Établissez un périphérique de quorum entre tous les noeuds rattachés au même boîtier de stockage de disques. Ajoutez un disque comme périphérique de quorum dans le boîtier partagé pour assurer, en cas d'échec d'un noeud, que les autres

peuvent maintenir le quorum et contrôler les groupes de périphériques de disques du boîtier partagé.

- Vous devez connecter le périphérique de quorum à au moins deux noeuds.
- Un périphérique de quorum peut être n'importe quel disque SCSI-2 ou SCSI-3 utilisé comme périphérique de quorum à deux accès. Les disques connectés à plus de deux noeuds doivent prendre en charge la réservation de groupe persistante (PGR) de SCSI-3, qu'ils soient ou non utilisés comme périphériques de quorum. Reportez-vous au chapitre concernant la planification dans le *Guide d'installation du logiciel Sun Cluster 3.1* pour de plus amples informations.
- Un disque contenant des données utilisateur peut être utilisé comme périphérique de quorum.

Séparation en cas d'échec

Une panne entraînant la partition du cluster (appelée *split brain*) est un des problèmes majeurs que peut rencontrer un cluster. Lorsque ce phénomène se produit, tous les noeuds ne peuvent pas communiquer, ainsi, des noeuds individuels ou des sous-ensembles de noeuds risquent de tenter de former des clusters individuels ou des sous-ensembles. Chaque sous-ensemble ou partition peut croire qu'elle est la seule à être propriétaire des disques multihôtes et à en posséder l'accès. Si plusieurs noeuds tentent d'écrire sur les disques, les données peuvent être altérées.

La séparation en cas d'échec limite l'accès des noeuds aux disques multihôtes en les empêchant physiquement d'accéder aux disques. Lorsqu'un noeud quitte le cluster (parce qu'il a échoué ou a été partitionné), la séparation en cas d'échec assure qu'il ne peut plus accéder aux disques. Seuls les membres actuels des noeuds ont accès aux disques, garantissant ainsi l'intégrité des données.

Les services des périphériques de disques intègrent des capacités de basculement pour les services utilisant les disques multihôtes. Lorsqu'un membre du cluster fonctionnant comme noeud principal (propriétaire) du groupe de périphériques de disques échoue ou devient inaccessible, un nouveau noeud principal est choisi pour prendre le relais et accéder au groupe de périphériques de disques avec un minimum d'interruption. Au cours de ce processus, l'ancien noeud principal doit renoncer à l'accès aux périphériques avant que le nouveau noeud principal puisse être démarré. Toutefois, lorsqu'un membre se détache du cluster et n'est plus joignable, le cluster ne peut pas informer ce noeud de libérer les périphériques dont il était le noeud principal. Il faut donc trouver un moyen de permettre aux membres survivants d'accéder aux périphériques globaux et d'en prendre le contrôle à la place des membres défectueux.

Ce moyen, le système SunPlex le trouve dans la fonction de séparation en cas d'échec du mode de réservation des disques SCSI. Grâce au système de réservation SCSI, les noeuds défectueux sont « isolés » à l'extérieur des disques multihôtes, pour les empêcher d'accéder à ces disques.

Le système de réservation de disques SCSI-2 prend en charge un type de réservation pouvant soit donner accès à tous les noeuds reliés aux disques (lorsqu'aucune réservation n'est en place), soit restreindre l'accès à un seul noeud (détenant la réservation).

Lorsqu'un membre détecte qu'un autre noeud ne communique plus sur l'interconnexion du cluster, il lance une procédure de séparation en cas d'échec pour empêcher le noeud d'accéder aux disques partagés. En présence d'une séparation en cas d'échec, il est normal que le noeud séparé panique et que des messages indiquant un « conflit de réservation » apparaissent sur sa console.

Le conflit de réservation a lieu parce qu'après qu'un noeud a été détecté comme n'appartenant plus au cluster, une réservation SCSI est placée sur l'ensemble des disques partagés entre ce noeud et les autres noeuds. Le noeud séparé n'est peut-être pas conscient de son état et s'il essaie d'accéder à un des disques partagés, il détecte la réservation et panique.

Mécanisme failfast pour séparation en cas d'échec

Le mécanisme par lequel le cluster assure qu'un noeud défectueux ne se réinitialise pas et ne commence pas à écrire sur le stockage partagé est appelé *failfast* .

Les noeuds membres du cluster activent en permanence un ioctl spécifique, MHIOCENFAILFAST, pour les disques auxquels ils ont accès, notamment les disques de quorum. Cet ioctl est une directive adressée au pilote de disques, il donne au noeud la capacité de paniquer au cas où il ne pourrait accéder à un disque parce que ce dernier a été réservé par d'autres noeuds.

Avec MHIOCENFAILFAST, le pilote contrôle le retour d'erreur de toutes les opérations de lecture et d'écriture qu'un noeud réalise sur le disque pour le code d'erreur `Reservation_Conflict`. L'ioctl, en arrière-plan, lance périodiquement une opération de test sur le disque pour détecter la présence de `Reservation_Conflict`. Les chemins de flux de contrôle de premier plan et d'arrière plan paniquent tous deux si le code `Reservation_Conflict` est renvoyé.

Pour les disques SCSI-2, les réservations ne sont pas persistantes ; elles ne survivent pas aux réinitialisations des noeuds. Pour les disques SCSI-3 dotés de la fonction PGR (réservation de groupe persistante), les informations de réservation sont stockées sur le disque et persistent après réinitialisation des noeuds. Le mécanisme failfast fonctionne de la même manière avec des disques SCSI-2 ou SCSI-3.

Si un noeud perd la connectivité aux autres noeuds du cluster et qu'il ne fait pas partie d'une partition pouvant atteindre un quorum, il est supprimé de force du cluster par un autre noeud. Un autre noeud faisant partie de la partition et pouvant atteindre un quorum place les réservations sur les disques partagés, et lorsque le noeud ne possédant pas de quorum tente d'accéder aux disques partagés, il reçoit un conflit de réservation et panique du fait de la présence du mécanisme failfast.

Après la panique, le noeud peut soit se réinitialiser et tenter de rejoindre le cluster, soit rester sur l'invite de la PROM OpenBoot (OBP). L'action retenue est déterminée par la définition du paramètre `auto-boot?` de l'OBP.

Gestionnaires de volumes

Le système SunPlex utilise un logiciel de gestion de volumes pour accroître la disponibilité des données à l'aide de miroirs et de disques de remplacement à chaud et pour gérer les pannes et remplacements des disques.

Il ne possède pas de gestionnaire de volumes propre, mais s'appuie sur les gestionnaires suivants :

- Solaris Volume Manager ;
- VERITAS Volume Manager.

Un gestionnaire de volumes au sein d'un cluster assure :

- la gestion des basculements des noeuds défectueux ;
- un support multivoie à partir de différents noeuds ;
- un accès distant transparent aux groupes de périphériques.

Lorsque les objets de la gestion de volumes arrivent sous le contrôle du cluster, ils deviennent des groupes de périphériques de disques. Pour de plus amples informations sur les gestionnaires de volumes, reportez-vous à la documentation de votre logiciel de gestion de volumes.

Remarque – lorsque vous planifiez vos ensembles de disques ou groupes de disques, il est important de comprendre comment leurs groupes de périphériques de disques associés sont liés aux ressources d'application (données) au sein du cluster. Reportez-vous au Guide d'installation du logiciel Sun Cluster 3.1 et au Sun Cluster 3.1 Data Services Installation and Configuration Guide pour de plus amples informations sur ces sujets.

Services de données

Le terme *service de données* décrit une application tiers telle qu'Oracle ou Sun ONE Web Server, configurée pour fonctionner sur un cluster plutôt que sur un seul serveur. Un service de données se compose d'une application, de fichiers de configuration Sun Cluster spécialisés et de méthodes de gestion Sun Cluster contrôlant les actions suivantes de l'application :

- le démarrage ;
- l'arrêt ;
- le contrôle et la prise de mesures correctives.

La Figure 3–4 compare une application fonctionnant sur un seul serveur d'applications (modèle de serveur unique) à la même application fonctionnant sur un cluster (modèle de serveur clusterisé). Notez que du point de vue de l'utilisateur, il n'y a aucune différence entre les deux configurations si ce n'est que l'application clusterisée peut être plus rapide et que son niveau de disponibilité est plus élevé.

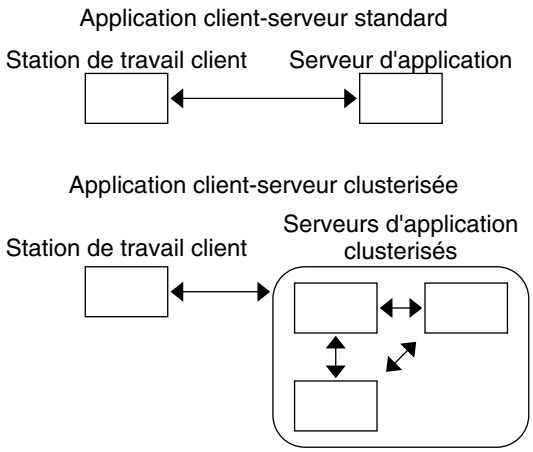


FIGURE 3–4 Configuration standard et configuration client/serveur clusterisée

Dans le modèle de serveur unique, l'application est configurée pour accéder au serveur à travers une interface de réseau public particulière (nom d'hôte). Le nom d'hôte est associé à ce serveur physique.

Dans le modèle de serveur clusterisé, l'interface de réseau public est un *nom d'hôte logique* ou une *adresse partagée*. Le terme *ressources réseau* se rapporte à la fois aux noms d'hôtes logiques et aux adresses partagées.

Certains services de données vous demandent de spécifier soit les noms d'hôtes logiques, soit les adresses partagées comme interfaces réseau : ils ne sont pas interchangeables. D'autres services de données vous permettent de spécifier les noms d'hôtes logiques comme les adresses partagées. Reportez-vous à l'installation et à la configuration de chaque service de données pour déterminer plus précisément le type d'interface que vous devez spécifier.

Une ressource réseau n'est pas associée à un serveur physique spécifique : elle peut se déplacer entre les serveurs physiques.

Une ressource réseau est initialement associée à un noeud, le *noeud principal*. Si le noeud principal échoue, la ressource réseau et la ressource d'application basculent sur un autre noeud du cluster (noeud secondaire). Lorsque la ressource réseau bascule, la ressource d'application continue de fonctionner sur le noeud secondaire après un bref laps de temps.

La Figure 3–5 compare le modèle de serveur unique au modèle de serveur clusterisé. Notez que dans le modèle de serveur clusterisé, une ressource réseau (nom d'hôte logique dans cet exemple) peut se déplacer entre plusieurs noeuds du cluster. L'application est configurée pour utiliser ce nom d'hôte logique à la place d'un nom d'hôte associé à un serveur particulier.

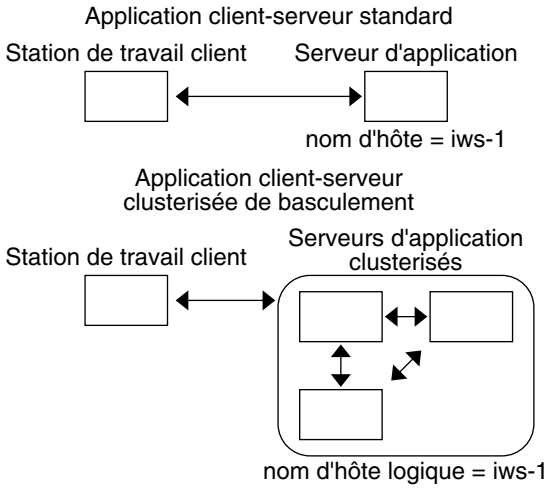


FIGURE 3–5 Nom d'hôte fixe et nom d'hôte logique

Une adresse partagée est aussi initialement associée à un noeud. Ce noeud s'appelle le noeud d'interface globale (ou noeud GIF) . Une adresse partagée est utilisée comme interface de réseau simple vers le cluster. Elle est connue sous le nom d'*interface globale*.

La différence entre le modèle de nom d'hôte logique et le modèle de service évolutif est que dans ce dernier, chaque noeud possède aussi une adresse partagée activement configurée sur son interface de bouclage. Cette configuration permet d'avoir plusieurs instances d'un service de données actives sur plusieurs noeuds simultanément. Le terme « service évolutif » signifie que vous pouvez augmenter la puissance de calcul de l'application en ajoutant des noeuds de cluster supplémentaires afin d'améliorer les performances.

En cas d'échec du noeud GIF, l'adresse partagée peut être mise sur un autre noeud où tourne aussi une instance de l'application (faisant ainsi de cet autre noeud le nouveau noeud GIF). L'adresse partagée peut également basculer sur un autre noeud du cluster n'ayant pas exécuté l'application auparavant.

La Figure 3–6 compare la configuration de serveur unique à la configuration de service évolutif clusterisé. Notez que dans la configuration service évolutif, l'adresse partagée est présente sur tous les noeuds. De la même façon qu'un nom d'hôte logique est utilisé dans un service de données à basculement, l'application est configurée pour utiliser cette adresse partagée à la place d'un nom d'hôte associé à un serveur particulier.

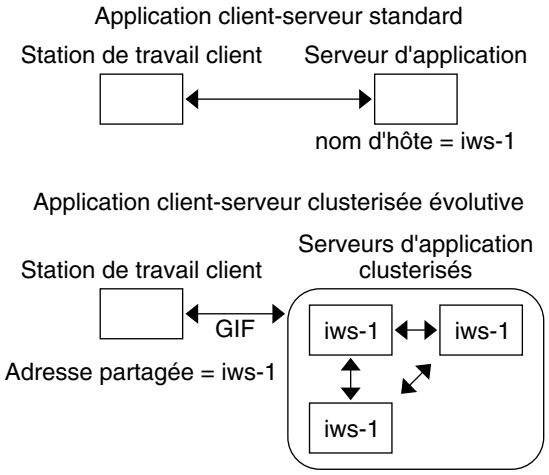


FIGURE 3–6 Nom d'hôte fixe et adresse partagée

Méthodes de services de données

Le logiciel Sun Cluster intègre un ensemble de méthodes de gestion de services. Ces méthodes fonctionnent sous le contrôle du Gestionnaire du groupe de ressources (RGM), qui les utilise pour démarrer, arrêter et contrôler l'application sur les noeuds du cluster. Ces méthodes, avec les logiciels de cluster et les disques multihôtes, permettent aux applications de devenir des services de données évolutifs ou de basculement.

Le gestionnaire du groupe de ressources (RGM) gère aussi les ressources du cluster, notamment les instances d'une application et les ressources réseau (noms d'hôtes logiques et adresses partagées).

En plus des méthodes fournies par le logiciel Sun Cluster, le système SunPlex intègre également une API (interface de programme d'application) et plusieurs outils de développement de services de données. Ces outils permettent aux programmeurs de développer les méthodes de services de données nécessaires pour que les autres applications fonctionnent comme des services de données à haute disponibilité avec le logiciel Sun Cluster.

Services de données de basculement

Si le noeud sur lequel le service de données fonctionne (noeud principal) échoue, le service est déplacé vers un autre noeud opérationnel sans intervention de l'utilisateur. Les services de basculement utilisent un *groupe de ressources de basculement* contenant les ressources des instances d'application et les ressources réseau (*noms d'hôtes logiques*). Les noms d'hôtes logiques sont des adresses IP pouvant être configurées sur un noeud puis automatiquement retirées pour être configurées sur un autre noeud.

Avec les services de données, les instances d'application ne tournent que sur un seul noeud. Si le détecteur de pannes rencontre une erreur, il essaie soit de redémarrer l'instance sur le même noeud, soit de la démarrer sur un autre noeud (basculement), selon la configuration du service de données.

Services de données évolutifs

Le service de données évolutif permet d'avoir des instances fonctionnant sur plusieurs noeuds. Les services évolutifs utilisent deux groupes de ressources : un *groupe de ressources évolutif* contenant les ressources d'application, et un groupe de ressources de basculement contenant les ressources réseau (*adresses partagées*) dont dépend le service évolutif. Le groupe de ressources évolutif peut être connecté à plusieurs noeuds, permettant ainsi à plusieurs instances du service de fonctionner en même temps. Le groupe de ressources de basculement hébergeant les adresses partagées ne peut être connecté qu'à un seul noeud à la fois. Tous les noeuds hébergeant un service évolutif utilisent la même adresse partagée pour héberger le service.

Les demandes de services arrivent au cluster à travers une interface réseau unique (interface globale) et sont réparties entre les noeuds en fonction de l'un des algorithmes prédéfinis par la *règle d'équilibrage de la charge*. Le cluster peut utiliser cette règle pour équilibrer la charge de service entre plusieurs noeuds. Notez qu'il peut y avoir plusieurs interfaces globales sur des noeuds différents hébergeant d'autres adresses partagées.

Avec les services évolutifs, les instances d'application tournent sur plusieurs noeuds en même temps. Si le noeud hébergeant l'interface globale échoue, celle-ci bascule sur un autre noeud. Si une instance d'application en cours de fonctionnement échoue, elle essaie de redémarrer sur le même noeud.

S'il lui est impossible de redémarrer sur le même noeud et qu'un autre noeud non utilisé est configuré pour exécuter le service, celui-ci bascule sur le noeud non utilisé. Si aucun noeud n'est disponible, l'instance d'application continue de fonctionner sur les noeuds restants et peut ainsi entraîner une dégradation des performances du service.

Remarque – l'état TCP de chaque instance d'application est conservé sur le noeud contenant l'instance et non sur le noeud d'interface globale. Ainsi, l'échec du noeud d'interface globale n'a pas d'incidence sur la connexion.

La Figure 3–7 montre un exemple de groupe de ressources évolutif et de groupe de ressources de basculement avec leurs dépendances dans le cadre de services évolutifs. Cet exemple présente trois groupes de ressources. Le groupe de ressources de basculement contient les ressources d'application des serveurs DNS à haute disponibilité et les ressources réseau utilisées à la fois par les serveurs DNS et les serveurs Web Apache à haute disponibilité. Les groupes de ressources évolutifs ne contiennent que les instances d'application du serveur Web Apache. Notez qu'il existe des dépendances entre les groupes de ressources évolutifs et de basculement (lignes continues) et que toutes les ressources d'application Apache dépendent de la ressource réseau `schost-2`, qui est une adresse partagée (ligne en trait tireté).

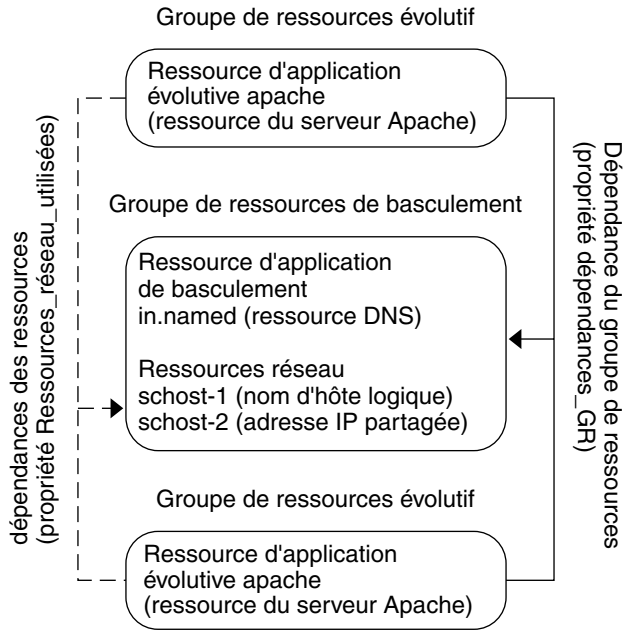


FIGURE 3–7 Exemple de groupe de ressources évolutif et de basculement

Architecture de service évolutif

L'objectif principal d'un réseau en cluster est de fournir de l'évolutivité aux services de données. L'évolutivité signifie qu'en dépit de l'accroissement de la charge offerte à un service, celui-ci peut maintenir un temps de réponse constant grâce à l'ajout de

nouveaux noeuds au cluster et à l'exécution de nouvelles instances de serveur. Ce service est appelé service de données évolutif. Un service Web est un bon exemple de ce type de service. Un service de données est généralement composé de plusieurs instances tournant chacune sur des noeuds différents du cluster. Ensemble, ces instances se comportent comme un service unique du point de vue d'un client distant et implémentent la fonctionnalité du service. On peut par exemple avoir un service Web évolutif composé de plusieurs démons `httpd` fonctionnant sur des noeuds différents. Chaque démon `httpd` peut servir une requête client. Le démon servant la requête dépend d'une *règle d'équilibrage de la charge*. La réponse au client semble venir du service, et non du démon servant la requête, et l'apparence d'un service unique est ainsi préservée.

Un service évolutif comporte :

- la prise en charge d'infrastructures de réseau pour les services évolutifs ;
- l'équilibrage de la charge ;
- la prise en charge de services de données et de gestion de réseaux (à l'aide du Gestionnaire du groupe de ressources).

La figure suivante représente l'architecture d'un service de données évolutif :

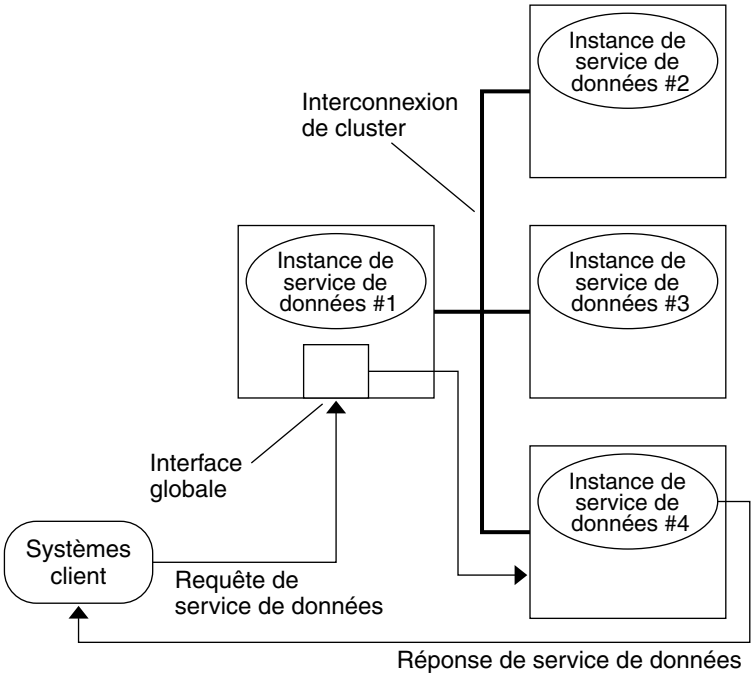


FIGURE 3–8 Architecture d'un service de données évolutif

Les noeuds n'étant pas hébergés par l'interface globale (noeuds proxy) hébergent l'adresse partagée sur leur interface de bouclage. Les paquets arrivant dans l'interface globale sont distribués à d'autres noeuds du cluster en fonction des règles d'équilibrage de la charge configurables. Les différentes règles d'équilibrage de la charge sont décrites ci-dessous.

Règles d'équilibrage de la charge

L'équilibrage de la charge permet d'améliorer les performances du service évolutif, tant en matière de temps de réponse que de rendement.

Il existe deux classes de services de données évolutifs : *pur* et *sticky*. Sur un service pur, n'importe quelle instance peut répondre aux requêtes du client. Sur un service sticky, le client envoie les requêtes à la même instance. Ces dernières ne sont pas redirigées vers d'autres instances.

Un service pur utilise une règle d'équilibrage de la charge pondérée. Sous cette règle, les requêtes du client sont réparties de manière uniforme entre les instances de serveur du cluster. Supposons par exemple que dans un cluster à trois noeuds, chaque noeud supporte une charge de 1. Chaque noeud servira 1/3 des requêtes d'un client pour le compte de ce service. Les charges peuvent être modifiées à tout moment par l'administrateur à l'aide de l'interface de la commande `scrgadm(1M)` ou de l'interface utilisateur graphique de gestion de SunPlex.

Un service sticky peut être de deux types : *sticky ordinaire* et *sticky joker*. Les services sticky permettent à plusieurs sessions d'application simultanées utilisant plusieurs connexions TCP de partager la mémoire d'état (état de la session d'application).

Les services sticky ordinaires permettent à un client de partager l'état entre plusieurs connections TCP simultanées. Le client est dit « sticky » par rapport à cette instance de serveur écoutant sur un seul port. Le client a la garantie que toutes ses requêtes vont vers la même instance de serveur, sous réserve que cette instance soit active et accessible et que la règle d'équilibrage de la charge ne soit pas modifiée alors que le service est en ligne.

Par exemple, un navigateur Web du client se connecte à une adresse IP partagée sur le port 80 en utilisant trois connexions TCP différentes, mais les connexions échangent entre elles des informations de session en mémoire cache lors de l'exécution du service.

La généralisation d'une règle sticky s'étend à plusieurs services évolutifs échangeant des informations de session en arrière-plan au même moment. Lorsque ces services échangent des informations de session en arrière plan au même moment, le client est dit « sticky » par rapport à ces multiples instances de serveur sur le même noeud écoutant sur des ports différents.

Par exemple, le client d'un site de commerce électronique remplit sa carte d'achats en utilisant le protocole HTTP ordinaire sur le port 80, mais il bascule vers le protocole SSL sur le port 443 pour envoyer les données sécurisées relatives au paiement de ces achats par carte de crédit.

Les services sticky joker utilisent des numéros de ports assignés de manière dynamique, mais attendent toujours des requêtes du client qu'elles aillent vers le même noeud. Le client est dit « sticky joker » sur les ports par rapport à la même adresse IP.

Un bon exemple de cette règle est le protocole FTP en mode passif. Un client se connecte à un serveur FTP sur le port 21 et est ensuite informé par le serveur de se connecter de nouveau à un serveur de ports d'écoute sur une plage de ports dynamiques. Toutes les requêtes pour cette adresse IP sont envoyées au même noeud ; celui que le serveur a désigné au client à travers les informations de contrôle.

Notez que pour chacune de ces règles sticky, la règle de la charge pondérée est activée par défaut, ainsi, la requête initiale d'un client est dirigée vers l'instance désignée par l'équilibreur de la charge. Après que le client a défini une affinité pour le noeud où fonctionne l'instance, les requêtes futures sont dirigées vers cette instance sous réserve que le noeud soit accessible et que la règle d'équilibrage de la charge ne soit pas modifiée.

Vous trouverez ci-après des détails complémentaires concernant des règles d'équilibrage de la charge spécifiques.

- Pondérée. La charge est répartie entre les différents noeuds en fonction de valeurs spécifiées. Cette règle est définie par la valeur `équilibrage_lignes_pondéré` de la propriété `poids_équilibrage_lignes`. Si la charge d'un noeud n'est pas explicitement définie, elle sera par défaut égale à 1.

  La règle pondérée redirige un certain pourcentage du trafic provenant des clients vers un noeud particulier. Si on a X = charge et A = la charge totale de tous les noeuds en activité, un noeud actif supportera approximativement X/A du total des nouvelles connexions, lorsque le nombre total de connexions est suffisamment élevé. Cette règle ne s'applique pas à des requêtes individuelles.

  Notez qu'elle ne fonctionne pas à tour de rôle. Si c'était le cas, chaque requête d'un client irait vers un noeud différent : la première requête vers le noeud 1, la seconde vers le noeud 2, etc.
- Sticky. Avec cette règle, l'ensemble de ports est connu au moment de la configuration des ressources d'application. Cette règle est définie par la valeur `équilibrage_lignes_sticky` de la propriété de la ressource `règle_équilibrage_charge`.
- Sticky joker. Cette règle est un sur-ensemble de la règle « sticky » ordinaire. Dans un service évolutif identifié par l'adresse IP, les ports sont assignés par le serveur (et ne sont pas connus à l'avance). Les ports peuvent varier. Cette règle est définie par la valeur `équilibrage_lignes_sticky_joker` de la propriété de la ressource `règle_équilibrage_charge`.

Paramètres de rétablissement

Les groupes de ressources basculent d'un noeud vers un autre noeud. Lorsque ce basculement se produit, le noeud secondaire d'origine devient principal. Les paramètres de rétablissement spécifient les actions qui auront lieu lorsque le noeud principal d'origine sera de nouveau en ligne. Vous pouvez soit autoriser le noeud principal d'origine à reprendre son rôle (rétablissement), soit permettre au noeud principal actuel de rester. L'option choisie peut être spécifiée à l'aide du paramètre `rétablissement` de la propriété du groupe de ressources.

Dans certains cas, si le noeud d'origine hébergeant le groupe de ressources échoue et se réinitialise souvent, la définition du paramètre de rétablissement pourrait réduire la disponibilité du groupe de ressources.

Détecteurs de pannes des services de données

Chaque service de données SunPlex intègre un détecteur de pannes sondant périodiquement le service de données pour vérifier son état. Un détecteur de pannes vérifie que le(s) démon(s) de l'application fonctionnent et que les clients sont servis. Sur la base des informations retournées par les sondes, des actions prédéfinies telles que le redémarrage des démons ou le déclenchement d'un basculement peuvent être initiées.

Développement de nouveaux services de données

Sun propose des fichiers de configuration et des modèles de méthodes de gestion vous permettant de faire fonctionner différentes applications comme service évolutif ou de basculement au sein d'un cluster. Si l'application que vous voulez faire fonctionner comme service évolutif ou de basculement ne fait pas partie de celles proposées par Sun, vous pouvez utiliser une API ou l'API DSET pour la configurer comme service évolutif ou de basculement.

Un ensemble de critères permet de déterminer si une application peut ou non devenir un service de basculement. Les critères spécifiques sont présentés dans la documentation SunPlex décrivant les API susceptibles d'être utilisées pour votre application.

Nous présentons ici quelques directives pour vous aider à déterminer si votre service peut ou non tirer profit d'une architecture de services de données évolutifs. Reportez-vous à la rubrique « Services de données évolutifs » à la page 61 pour de plus amples informations sur les services évolutifs.

Les nouveaux services remplissant les critères présentés ci-après peuvent utiliser les services évolutifs. Si un service existant ne correspond pas exactement à ces critères, certaines parties risquent de devoir être réécrites.

Un service de données évolutifs a les caractéristiques suivantes : premièrement, il est composé d'une ou plusieurs *instances* de serveur. Chaque instance tourne sur un noeud différent du cluster et plusieurs instances du même service ne peuvent pas fonctionner sur le même noeud.

Deuxièmement, si le service procure un stockage de données logique externe, l'accès simultané de plusieurs instances de serveur à ce stockage doit être synchronisé pour éviter de perdre des mises à jour ou de lire des données tandis qu'il est modifié. Notez que nous utilisons le terme « externe » pour distinguer le dispositif de stockage de la mémoire interne et le terme « logique » parce que le dispositif de stockage apparaît comme une entité unique, bien qu'il puisse lui-même être répliqué. En outre, ce stockage de données logique possède une propriété grâce à laquelle chaque fois qu'une instance de serveur met à jour les données stockées, la mise à jour est immédiatement vue par les autres instances.

Le système SunPlex fournit ce stockage externe à travers son système de fichiers de cluster et ses partitions de données brutes globales. Supposons par exemple qu'un service écrive des données sur un fichier journal externe ou qu'il modifie les données en place. Lorsque plusieurs instances de ce service fonctionnent, chacune d'elles a accès à ce journal externe et elles peuvent y accéder simultanément. Les instances doivent alors synchroniser leur accès pour éviter toute interférence entre elles. Le service peut utiliser le système de verrouillage de fichiers ordinaire de Solaris via `fcntl(2)` et `lockf(3C)` pour effectuer cette synchronisation.

Une base de données back-end, telle les bases de données à haute disponibilité Oracle ou Oracle Parallel Server/Real Application Clusters est un autre exemple de ce type de stockage. Notez que ce type de serveur de base de données back-end possède un dispositif de synchronisation muni d'un système de requêtes à la base de données ou de transactions de mises à jour ; ainsi les instances de serveur n'ont pas besoin d'implémenter leur propre synchronisation.

Le serveur Sun IMAP tel qu'il se présente à l'heure actuelle est un exemple de service non évolutif. Le service met à jour une mémoire, mais cette mémoire est privée et lorsque les instances IMAP écrivent dans cette mémoire, elles s'écrasent mutuellement parce que les mises à jour ne sont pas synchronisées. Le serveur IMAP doit être réécrit pour synchroniser les accès simultanés.

Enfin, notez que les instances peuvent avoir des données privées se détachant des données d'autres instances. Dans ce cas, le service n'a pas à se préoccuper de la synchronisation des accès, car les données sont privées et seule cette instance peut les manipuler. Vous devez alors prendre garde à ne pas stocker ces données privées sous le système de fichiers du cluster car elles ont la capacité de devenir globalement accessibles.

API de services de données et de bibliothèque de développement de services de données

Le système SunPlex intègre les éléments suivants permettant de rendre les applications hautement disponibles :

- des services de données intégrés au système SunPlex ;
- une API (interface de programme d'application) de services de données ;
- une API de bibliothèque de développement de services de données ;
- un service de données « générique ».

Le document *Sun Cluster 3.1 Data Services Installation and Configuration Guide* décrit la procédure d'installation et de configuration des services de données intégrés au système SunPlex. Le document *Sun Cluster 3.1 Data Services Developer's Guide* explique la façon de rendre d'autres applications hautement disponibles dans le cadre de Sun Cluster.

Les API Sun Cluster permettent aux programmeurs de développer des détecteurs de pannes et des scripts démarrant et arrêtant les instances des services de données. Grâce à ces outils, une application peut devenir un service de données évolutif ou de basculement. En outre, le système SunPlex intègre un service de données « générique » pouvant être utilisé pour générer rapidement les méthodes de démarrage et d'arrêt d'une application afin de la faire fonctionner comme service évolutif ou de basculement.

Utilisation de l'interconnexion de cluster pour le trafic des services de données

Un cluster doit avoir de multiples connexions réseau entre les noeuds pour former une interconnexion de cluster. Le logiciel de clustering fait appel à de nombreuses interconnexions pour optimiser la disponibilité et les performances. Pour le trafic interne (par exemple, les données des systèmes de fichiers ou des services évolutifs), les messages sont répartis à tour de rôle sur toutes les interconnexions disponibles.

L'interconnexion de cluster est également mise à la disposition des applications pour garantir une communication hautement disponible entre les noeuds. Par exemple, une application répartie peut avoir des composants exécutés sur différents noeuds et ayant besoin de communiquer entre eux. En utilisant l'interconnexion de cluster plutôt que le transport public, ces connexions peuvent résister à l'échec d'un lien individuel.

Pour utiliser l'interconnexion de cluster dans le cadre des communications, l'application doit adopter les noms d'hôtes privés configurés à l'installation du cluster. Par exemple, si le nom d'hôte privé du noeud 1 est `clusternode1-priv` , il faut utiliser ce nom pour communiquer sur l'interconnexion de cluster vers le noeud 1. Les sockets TCP ouverts à l'aide de ce nom sont dirigés vers l'interconnexion de cluster et peuvent être redirigés de manière transparente en cas de panne du réseau.

Veuillez noter que comme les noms d'hôte privés peuvent être configurés durant l'installation, l'interconnexion de cluster peut utiliser n'importe quel nom choisi à ce moment-là. Le nom réel peut être obtenu à l'aide de la commande `scha_cluster_get`(3HA) suivie de l'argument `scha_privatelink_hostname_node`.

Lors de l'utilisation de l'interconnexion du cluster pour les applications, une seule interconnexion est établie entre deux noeuds d'une paire, mais des interconnexions distinctes sont, si possible, utilisées entre les différentes paires de noeuds. Supposons par exemple qu'une application tournant sur trois noeuds communique sur l'interconnexion de cluster. La communication entre les noeuds 1 et 2 peut avoir lieu sur l'interface `hme0` et la communication entre les noeuds 1 et 3 sur l'interface `qfe1`. Autrement dit, les communications d'une application entre deux noeuds sont limitées à une simple interconnexion, tandis que les communications sur cluster interne sont réparties sur toutes les interconnexions.

Veuillez noter que l'application partage l'interconnexion avec le trafic de cluster interne et, de ce fait, que la bande passante à la disposition de l'application dépend de la bande passante utilisée par le reste du trafic. En présence d'une défaillance, le trafic interne est acheminé tour à tour vers les autres interconnexions et les connexions des applications sont basculées vers une interconnexion en état de marche.

Deux types d'adresse prennent en charge l'interconnexion du cluster, et la commande `gethostbyname`(3N) sur un nom d'hôte privé renvoie normalement deux adresses IP. La première adresse est appelée *adresse de paire logique* et la seconde *adresse logique par noeud*.

Une adresse de paire logique distincte est attribuée à chaque paire de noeuds. Ce petit réseau logique prend en charge la reprise des connexions. Une adresse par noeud fixe est également attribuée à chaque noeud. Autrement dit, les adresses de paires logiques de `clusternode1-priv` sont différentes sur chaque noeud, tandis que l'adresse logique par noeud de `clusternode1-priv` est la même sur chaque noeud. Un noeud ne comporte toutefois pas d'adresse de paire logique le désignant, de ce fait, la commande `gethostbyname`(`clusternode1-priv`) sur le noeud 1 ne renvoie que l'adresse par noeud logique.

Veuillez noter que les applications acceptant des connexions sur l'interconnexion de cluster et vérifiant ensuite l'adresse IP pour des raisons de sécurité doivent procéder à une nouvelle vérification de toutes les adresses IP renvoyées par la commande `gethostbyname`, et non pas simplement la première.

Si vous avez besoin d'adresses IP cohérentes à tout moment dans votre application, configurez-la pour être liée à l'adresse par noeud du côté client et du côté serveur de sorte que toutes les connexions semblent provenir et se diriger vers cette adresse par noeud.

Ressources, groupes de ressources et types de ressources

Les services de données utilisent plusieurs types de *ressources* : les applications telles que le serveur Web Apache ou le serveur Web Sun ONE utilisent des adresses de réseau (noms d'hôtes logiques et adresses partagées) dont dépendent les applications. Les ressources de l'application et du réseau constituent une unité de base que gère le RGM.

Les services de données sont des types de ressource. Par exemple, Sun Cluster HA pour Oracle est le type de ressource `SUNW.oracle-server` et Sun Cluster HA pour Apache est le type de ressource `SUNW.apache`.

Une ressource est une instanciation d'un *type de ressource* défini au niveau du cluster. Plusieurs types de ressource sont définis.

Les ressources réseau sont de type `SUNW.LogicalHostname` ou `SUNW.SharedAddress`. Ces deux types de ressource sont préenregistrés dans le logiciel Sun Cluster.

Les types de ressources `SUNW.HAStorage` et `HAStoragePlus` servent à synchroniser le démarrage des ressources et les groupes de périphériques de disques dont dépendent les ressources. Vous avez ainsi l'assurance qu'avant qu'un service de données ne démarre, les chemins vers les points de montage du système de fichiers du cluster, les périphériques globaux et les noms des groupes de périphériques sont disponibles. Pour de plus amples informations, reportez-vous à la rubrique "Synchronizing the Startups Between Resource Groups and Disk Device Groups" du document *Data Services Installation and Configuration Guide* (le type de ressource `HAStoragePlus` est désormais disponible dans Sun Cluster 3.0 5/02 et intègre une autre fonction permettant aux systèmes de fichiers locaux d'être hautement disponibles ; pour de plus amples informations sur cette fonction, reportez-vous à la rubrique « Type de ressource `HAStoragePlus` » à la page 47).

Les ressources administrées par le RGM sont réparties dans des groupes, appelés *groupes de ressources*, afin de pouvoir être gérées comme des unités. Si une commutation ou un basculement est initié sur un groupe de ressources, ce dernier se transforme en unité.

Remarque – lorsque vous introduisez un groupe de ressources contenant des ressources d'application en ligne, l'application s'exécute. La méthode de démarrage du service de données attend que l'application soit lancée et qu'elle fonctionne avant de se fermer. La méthode permettant de définir à quel moment l'application est opérationnelle est identique à la méthode utilisée par le contrôleur de panne pour déterminer si un service de données est en train de servir des clients. Reportez-vous au document *Sun Cluster 3.1 Data Services Installation and Configuration Guide* pour de plus amples informations sur cette méthode.

Gestionnaire du groupe de ressources (RGM)

Le gestionnaire du groupe de ressources (RGM) contrôle les services de données (applications) comme des ressources gérées par des implémentations de *type de ressource*. Ces implémentations peuvent être fournies par Sun ou créées par un développeur à l'aide d'un modèle de service de données générique, l'API de la bibliothèque de développement d'un service de données (API BDSD) ou l'API de la gestion de ressources (API GR). L'administrateur du cluster crée et gère des ressources dans des conteneurs appelés *groupes de ressources*. Le RGM arrête et démarre les groupes de ressources des noeuds sélectionnés en réponse aux modifications des membres du cluster.

Le RGM agit sur les *ressources* et les *groupes de ressources.* Sous l'action du RGM les ressources et les groupes de ressources passent de l'état en ligne à l'état hors ligne. Vous trouverez une description complète des états et paramètres applicables aux ressources et groupes de ressources à la rubrique « Paramètres et états des ressources et groupes de ressources » à la page 71. Reportez-vous au document « Ressources, groupes de ressources et types de ressources » à la page 70 pour de plus amples informations sur la procédure de lancement d'un projet de gestion de ressources sous le contrôle du gestionnaire du groupe de ressources.

Paramètres et états des ressources et groupes de ressources

Un administrateur applique des paramètres statiques aux ressources et groupes de ressources. Ces paramètres ne peuvent être modifiés qu'à travers des actions d'administration. Le RGM met les groupes de ressources à différents « états » dynamiques. Ces paramètres et états sont décrits ci-après.

- Gérés ou non gérés : ces paramètres s'appliquent uniquement aux groupes de ressources d'un cluster. Les groupes de ressources sont gérés par le RGM. La commande `scrgadm(1M)` permet de demander qu'un groupe de ressources soit géré ou non géré par le RGM. Ces paramètres ne changent pas au cours d'une reconfiguration du cluster.

  Au moment de sa création, un groupe de ressources n'est pas géré. Il doit être géré avant que toute ressource placée dans le groupe ne devienne active.

  Dans certains serveurs de données comme les serveurs Web évolutifs, des tâches doivent être effectuées avant le démarrage et après l'arrêt des ressources réseau. Ces tâches s'accomplissent par le biais des méthodes initialisation (INIT) et fin (FINI) des services de données. Les méthodes INIT ne fonctionnent que si le groupe de ressources dans lequel réside les ressources est à l'état « géré ».

  Lorsqu'un groupe de ressources passe de l'état « non géré » à « géré », toutes les méthodes INIT enregistrées pour le groupe sont exécutées sur les ressources du groupe.

  Lorsqu'un groupe de ressources passe de l'état « géré » à « non géré », toutes les méthodes FINI enregistrées sont appelées pour procéder à un nettoyage.

Les méthodes INIT et FINI sont le plus souvent utilisées pour les ressources réseau de services évolutifs, mais elles peuvent aussi être utilisées pour toute opération d'initialisation ou de nettoyage non faite par l'application.

- Activé ou désactivé : ces paramètres s'appliquent aux ressources d'un cluster. La commande `scrgadm(1M)` permet d'activer ou de désactiver une ressource. Ces paramètres ne changent pas après une reconfiguration du cluster.

  Une ressource est normalement activée et fonctionne dans le système.

  Si pour une raison ou une autre, vous souhaitez rendre cette ressource indisponible sur tous les noeuds du cluster, vous devez la désactiver. Un ressource désactivée devient inutilisable.

- En ligne ou hors ligne : états dynamiques s'appliquant tant aux ressources qu'aux groupes de ressources.

  Ces états changent au cours des reconfigurations du cluster au moment des commutations et des basculements. Ils peuvent aussi être modifiés à travers des actions d'administration. La commande `scswitch(1M)` permet de mettre une ressource ou un groupe de ressources en ligne ou hors ligne.

  Une ressource de basculement ou un groupe de ressources ne peuvent être connectés qu'à un seul noeud à la fois. Une ressource évolutive ou un groupe de ressources peuvent être en ligne sur certains noeuds et hors ligne sur d'autres. Durant une commutation ou un basculement, les groupes de ressources et les ressources qu'ils contiennent sont déconnectés d'un noeud puis reconnectés à un autre noeud.

  Si un groupe de ressources est déconnecté, toutes ses ressources le sont aussi. Si un groupe de ressources est connecté, toutes ses ressources activées le sont aussi.

  Les groupes de ressources peuvent contenir plusieurs ressources ayant des dépendances entre elles. Ces dépendances nécessitent que les ressources soient mises en ligne et hors ligne dans un ordre particulier. Le temps nécessaire aux méthodes utilisées pour connecter et déconnecter les ressources peut varier pour chaque ressource. Du fait de l'existence de dépendances entre les ressources et du caractère variable des temps de démarrage et d'arrêt, les ressources d'un même groupe peuvent avoir différents états « en ligne » et « hors ligne » durant une reconfiguration du cluster.

Propriétés des ressources et des groupes de ressources

Vous pouvez attribuer des propriétés aux ressources et groupes de ressources de vos services de données SunPlex. Il existe des propriétés standard communes à tous les services de données et des propriétés d'extension propres à chaque service de données. Certaines propriétés standard et propriétés d'extension sont définies par des paramètres par défaut et vous n'avez pas à les modifier. D'autres propriétés doivent être définies au moment de la création et de la configuration des ressources. La documentation de chaque service de données indique quelles propriétés de ressources peuvent être définies et comment les définir.

Les propriétés standard permettent de configurer les propriétés des ressources et groupes de ressources habituellement indépendantes de tout service de données. Les propriétés standard sont décrites en annexe du document *Sun Cluster 3.1 Data Services Installation and Configuration Guide*.

Les propriétés d'extension du RGM (gestionnaire du groupe de ressources) donne des informations sur l'emplacement des binaires d'application et des fichiers de configuration. Les propriétés d'extension peuvent être modifiées lorsque vous configurez vos services de données. Elles sont décrites au chapitre individuel consacré au service de données du document *Sun Cluster 3.1 Data Services Installation and Configuration Guide*.

Configuration de projets de services de données

Les services de données peuvent être configurés pour démarrer sous un nom de projet Solaris lorsqu'ils sont connectés à l'aide du RGM. La configuration associe une ressource ou un groupe de ressources géré par le RGM à une ID de projet Solaris. La correspondance entre votre ressource ou groupe de ressources et une ID de projet vous permet d'utiliser les outils de contrôle sophistiqués de l'environnement Solaris pour gérer les charges de travail et la consommation au sein du cluster.

Remarque – cette configuration n'est possible que si la version actuelle du logiciel Sun Cluster fonctionne sous Solaris 9.

L'utilisation des fonctions de gestion de Solaris dans un environnement de cluster vous permet d'assurer que les applications les plus importantes gardent la priorité lorsqu'un noeud est partagé avec d'autres applications. Il arrive que les applications partagent un noeud si vous avez des services consolidés ou si les applications ont basculé. L'utilisation des fonctions de gestion décrites ci-dessus permettent d'améliorer la disponibilité d'une application critique en empêchant d'autres applications non prioritaires de consommer trop de ressources du système tel que le temps UC.

Remarque – la documentation Solaris se rapportant à cette fonction désigne le temps UC, les processus, les tâches et les composants similaires sous le terme de « ressources ». La documentation Sun Cluster utilise quant à elle le terme de « ressources » pour désigner les entités sous le contrôle du RGM. Dans la rubrique suivante, on utilisera le terme de« ressource » en référence aux entités de Sun Cluster qui sont sous le contrôle du RGM et le terme de « ressources système » en référence au temps UC, aux processus et aux tâches.

Cette rubrique décrit la procédure de configuration des services de données pour le lancement de processus dans un `project`(4) Solaris 9 spécifié. Cette rubrique présente également plusieurs scénarios de basculement et des suggestions pour l'utilisation des fonctions de gestion de l'environnement Solaris. Pour consulter la documentation complète de la fonction de gestion, reportez-vous à la rubrique *System Administration Guide: Resource Management and Network Services* de la *Solaris 9 System Administrator Collection.*

Pour configurer des ressources ou groupes de ressources en vue de l'utilisation des fonctions de gestion Solaris dans un cluster, procédez de la manière suivante :

1. Configurez les applications dans la ressource.
2. Configurez les ressources dans le groupe de ressources.
3. Activez les ressources du groupe de ressources.
4. Mettez le groupe de ressource à l'état géré.
5. Créez un projet Solaris pour votre groupe de ressources.
6. Configurez les propriétés standard pour associer le nom du groupe de ressources au projet que vous avez créé à l'étape 5.
7. Mettez le groupe de ressources en ligne.

Pour configurer les propriétés standard `Nom_projet_ressource` ou `Nom_projet_GR` afin d'associer l'ID de projet Solaris à la ressource ou au groupe de ressources, utilisez l'option `-y` avec la commande `scrgadm`(1M). Définissez les valeurs des propriétés de la ressource ou du groupe de ressources. Pour consulter les définitions des propriétés, reportez-vous à la rubrique "Standard Properties" dans le document *Sun Cluster 3.1 Data Services Installation and Configuration Guide*. Reportez-vous aux documents `r_properties`(5) et `rg_properties`(5) pour de plus amples informations sur les propriétés.

Le nom de projet spécifié doit figuré dans la base de données des projets (`/etc/project`) et le superutilisateur doit être configuré comme membre du projet nommé. Reportez-vous à la rubrique "Projects and Tasks" du document *System Administration Guide: Resource Management and Network Services* de la collection *Solaris 9 System Administrator*pour de plus amples informations concernant la conception de la base données de noms de projets. Reportez-vous à la rubrique `project`(4) pour consulter la description de la syntaxe des fichiers de projets.

Lorsque le RGM met les ressources ou les groupes de ressources en ligne, il lance les processus connexes sous le nom du projet.

Remarque – les utilisateurs peuvent à tout moment associer la ressource ou le groupe de ressources à un projet. Toutefois, le nom de projet ne sera effectif qu'au moment où la ressource ou le groupe de ressources aura été mis hors ligne puis de nouveau en ligne à l'aide du RGM.

Le lancement des ressources ou des groupes de ressources sous le nom de projet vous permet de configurer les fonctions indiquées ci-après pour gérer les ressources système au sein du cluster.

- Comptabilité étendue : elle constitue un moyen souple d'enregistrer la consommation d'une tâche ou d'un procédé. La comptabilité étendue vous permet d'examiner l'historique de l'utilisation et d'évaluer les besoins en capacité des charges de travail futures.
- Contrôles : ce mécanisme permet d'appliquer des contraintes aux ressources système. On peut empêcher les processus, tâches et projets de surconsommer certaines ressources système.
- Fair Share Scheduling (FSS) : cette fonction offre la possibilité de contrôler l'allocation de temps UC aux charges de travail, en fonction de leur importance. L'importance des charges de travail est définie par le nombre de parts de temps UC attribué à chaque charge. Reportez-vous à la rubrique `dispadmin` (1M) pour consulter la procédure de définition de FSS comme programmateur par défaut. Consultez également les documents `priocntl`(1), `ps`(1) et `FSS`(7) pour de plus amples informations.
- Pools : cette fonction offre la possibilité d'utiliser des partitions pour les applications interactives en fonction des besoins de l'application. Les pools permettent de segmenter un serveur qui prend en charge un certain nombre d'applications différentes. Grâce à l'utilisation de pools les réponses des applications deviennent plus prévisibles.

Détermination des besoins pour la configuration de projets

Avant de configurer les services de données pour l'utilisation des fonctions de contrôle de Solaris dans un environnement Sun Cluster, vous devez décider de la manière dont vous souhaitez contrôler et suivre les ressources au cours des commutations et des basculements. Veillez à identifier les dépendances au sein du cluster avant de configurer un nouveau projet. Par exemple, les ressources et groupes de ressources dépendent des groupes de périphériques de disques. Pour identifier les priorités de la liste de noeuds de votre groupe de ressources, utilisez les propriétés de groupes de ressources `liste_noeuds`, `rétablissement`, `éléments_principaux_max` et `éléments_principaux_souhaités` configurées avec `scrgadm` (1M). Reportez-vous à la rubrique "Relationship Between Resource Groups and Disk Device Groups" dans le document *Sun Cluster 3.1 Data Services Installation and Configuration Guide* pour obtenir une brève description des dépendances entre les groupes de ressources et les groupes de périphériques de disques dans la liste de noeuds. Pour obtenir une description détaillée des propriétés, reportez-vous à la rubrique `rg_properties` (5).

Pour déterminer les priorités des groupes de périphériques de disques dans la liste de noeuds, utilisez les propriétés `prédilection` et `rétablissement` configurées avec `scrgadm`(1M) et `scsetup`(1M). Pour des informations relatives à la procédure, reportez-vous à la rubrique "How To Change Disk Device Properties" de "

Administering Disk Device Groups" dans le document *Sun Cluster 3.1 System Administration Guide*. Reportez-vous à la rubrique « Les composants matériels du système SunPlex » à la page 19 pour obtenir des informations théoriques concernant la configuration des noeuds et le comportement des services de données évolutifs ou de basculement.

Si vous configurez tous les noeuds du cluster de la même façon, les limites d'utilisation sont appliquées de manière identique aux noeuds principaux et aux noeuds secondaires. Les paramètres de configuration des projets doivent être identiques pour toutes les applications dans les fichiers de configuration de tous les noeuds. Tous les projets associés à l'application doivent au moins être accessibles par la base de données du projet sur tous les maîtres potentiels de cette application. Supposons que l'application 1 est contrôlée par *phys-schost-1* mais qu'elle peut être potentiellement commutée ou basculée sur *phys-schost-2* ou *phys-schost-3*. Le projet associé à l'application 1 doit être accessible sur les trois noeuds (*phys-schost-1* , *phys-schost-2* et *phys-schost-3*).

Remarque – les informations de la base de données de projet peuvent être placées dans un fichier de base de données `/etc/project` ou stockées dans la mappe NIS ou le service de répertoire LDAP.

L'environnement Solaris permet une configuration souple des paramètres d'utilisation et peu de restrictions sont imposées par Sun Cluster. Les choix de configuration sont tributaires des besoins du site. Tenez compte des indications précisées aux rubriques suivantes pour la configuration de vos systèmes.

Définition des limites de mémoire virtuelle par processus

Pour limiter la mémoire virtuelle par processus, définissez le paramètre `process.max-address-space`. Reportez-vous à la rubrique `rctladm`(1M) pour de plus amples informations sur la définition de la valeur de `process.max-address-space`.

Lorsque vous utilisez les fonctions de contrôle de gestion avec Sun Cluster, veillez à configurer les limites de la mémoire de manière à éviter les basculements intempestifs des applications et un effet « ping-pong » sur celles-ci. En général :

- Ne définissez pas de limites de mémoire trop basses.

  Lorsqu'une application atteint sa limite de mémoire, elle peut basculer. Cet aspect est particulièrement important pour les applications de base de données, où les conséquences liées à l'atteinte de la limite de mémoire virtuelle sont imprévisibles.
- Les limites de mémoire des noeuds principaux et des noeuds secondaires ne doivent pas être identiques.

Si elles l'étaient, un effet ping-pong pourrait se produire au moment où l'application atteint sa limite de mémoire et bascule sur un noeud secondaire ayant une limite de mémoire identique. Définissez une limite de mémoire légèrement supérieure sur le noeud secondaire. Vous préviendrez ainsi des effets ping-pong et l'administrateur système aura un laps de temps lui permettant de régler les paramètres si nécessaire.

- Utilisez les limites de la mémoire de gestion de ressources pour l'équilibrage de la charge.

  Vous pouvez par exemple utiliser les limites de la mémoire pour empêcher une application errante de consommer trop d'espace de swap.

Scénarios de basculement

Vous pouvez configurer les paramètres de gestion de sorte que l'allocation de la configuration de projet (`/etc/project`) soit opérationnelle durant le fonctionnement normal du cluster et dans les situations de commutation et de basculement.

Les rubriques suivantes présentent des exemples de scénarios.

- Les deux premières rubriques, "Cluster à deux noeuds avec deux applications" et "Cluster à deux noeuds avec trois applications", présentent des scénarios de basculement de noeuds complets.
- La rubrique "Basculement de groupes de ressources uniquement" illustre les opérations de basculement d'une application seulement.

Dans un environnement de cluster, une application est configurée dans la ressource et une ressource est configurée dans un groupe de ressources (GR). Lorsqu'un échec se produit, le groupe de ressources et ses applications associées basculent sur un autre noeud. Dans les exemples suivants les ressources n'apparaissent pas de manière explicite. Supposons que chaque ressource n'a qu'une seule application.

Remarque – un basculement intervient en fonction de l'ordre de préférence de la liste de noeuds définie par le RGM.

Les exemples présentés ci-après comportent les contraintes suivantes :

- Application 1 (App-1) est configurée dans le groupe de ressources GR-1.
- Application 2 (App-2) est configurée dans le groupe de ressources GR-2.
- Application 3 (App-3) est configurée dans le groupe de ressources GR-3.

Bien que le nombre de parts attribuées reste le même, le pourcentage de temps UC alloué à chaque application change après un basculement. Ce pourcentage dépend du nombre d'applications tournant sur le noeud et du nombre de parts attribuées à chaque application active.

Pour ces scénarios, supposons que nous avons les configurations suivantes :

- Toutes les applications sont configurées sous un même projet.
- Chaque ressource n'a qu'une seule application.
- Les applications ne sont pas les seuls processus actifs sur les noeuds.
- Les bases de données de projets sont configurées de la même manière sur tous les noeuds du cluster.

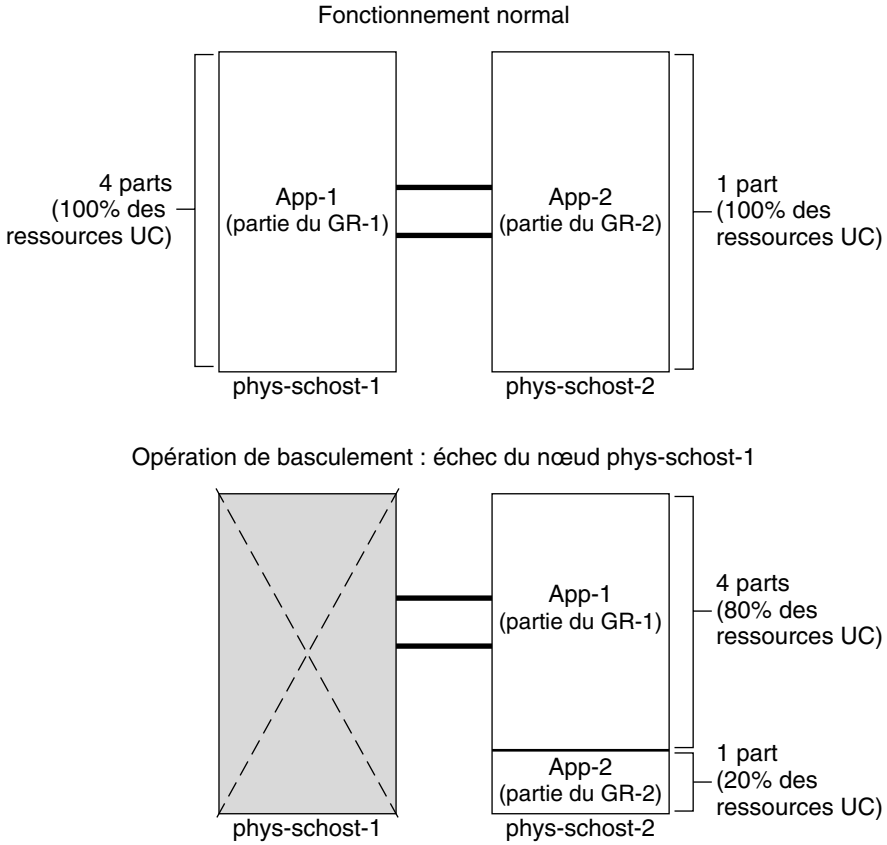
Cluster à deux noeuds avec deux applications

Vous pouvez configurer deux applications sur un cluster à deux noeuds pour assurer que chaque hôte physique (*phys-schost-1*, *phys-schost-2*) fonctionne comme maître par défaut d'une application. Chaque hôte physique sert de noeud secondaire à l'autre hôte physique. Tous les projets associés à l'application 1 et l'application 2 doivent être représentés dans les fichiers de la base de données de projets des deux noeuds. Lorsque le cluster fonctionne normalement, chaque application tourne sur son maître par défaut, où le temps UC lui est attribué par la fonction de gestion.

Lorsqu'un basculement ou une commutation a lieu, les deux applications tournent sur un seul noeud, où les parts précisées dans le fichier de configuration leur sont attribuées. Par exemple, cette entrée dans le fichier `/etc/project` spécifie que 4 parts sont attribuées à l'application 1 et 1 part à l'application 2.

```
Prj_1:100:project for App-1:root::project.cpu-shares=(privileged,4,none)
Prj_2:101:project for App-2:root::project.cpu-shares=(privileged,1,none)
```

Le schéma indiqué ci-après présente le fonctionnement de cette configuration en situation normale et en cas de basculement. Le nombre de parts assignées ne change pas. Toutefois, le pourcentage de temps UC disponible pour chaque application peut changer en fonction du nombre de parts attribuées à chaque processus demandant du temps UC.



Cluster à deux noeuds avec trois applications

Sur un cluster à deux noeuds avec trois applications, vous pouvez configurer un hôte physique (*phys-schost-1*) comme maître par défaut d'une application et le deuxième hôte (*phys-schost-2*) comme maître par défaut des deux autres applications. Dans l'exemple suivant, supposons que le fichier de base de données des projets se trouve sur chaque noeud. Le fichier de base de données des projets ne change pas lorsqu'un basculement ou une commutation a lieu.

```
Prj_1:103:project for App-1:root::project.cpu-shares=(privileged,5,none)
Prj_2:104:project for App_2:root::project.cpu-shares=(privileged,3,none)
Prj_3:105:project for App_3:root::project.cpu-shares=(privileged,2,none)
```

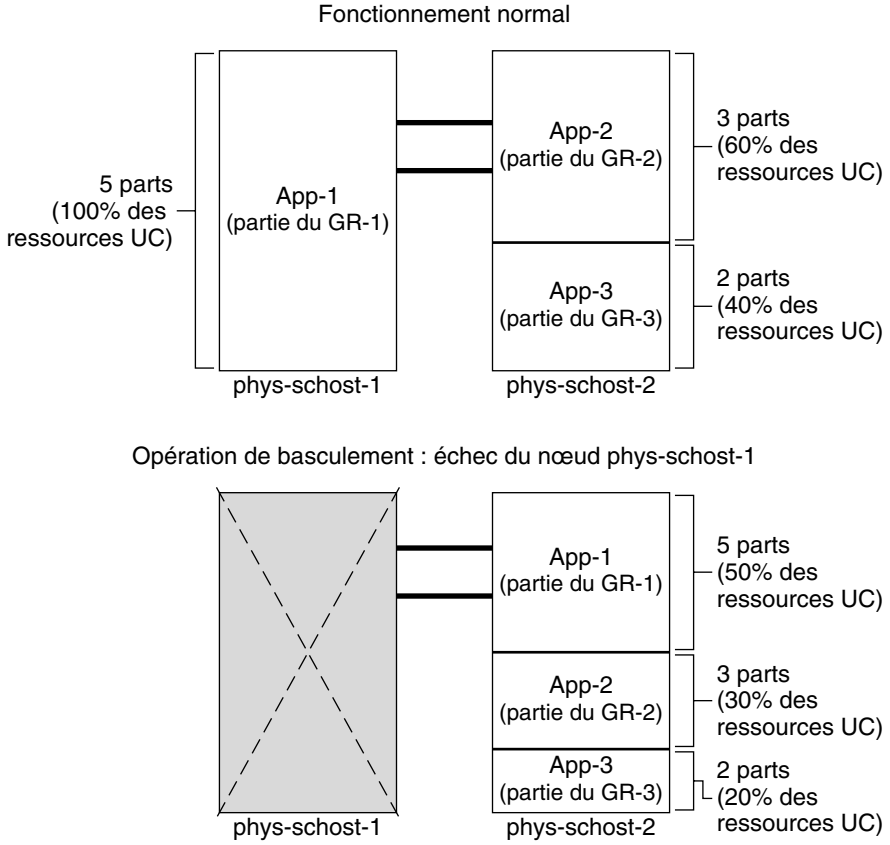
Lorsque le cluster fonctionne normalement, 5 parts sont allouées à l'application 1 sur son maître par défaut, *phys-schost-1*. Ce nombre équivaut à 100 pour cent du temps UC parce que c'est la seule application qui demande du temps UC sur ce noeud. Trois et

deux parts sont respectivement allouées aux applications 2 et 3 sur leur maître par défaut, *phys-schost-2* . L'application 2 reçoit 60 pour cent du temps UC et l'application 3 reçoit 40 pour cent du temps UC au cours du fonctionnement normal.

Si un basculement ou une commutation se produit et que l'application 1 bascule sur *phys-schost-2*, les parts des trois applications restent les mêmes. Toutefois, les pourcentages de ressources UC sont allouées en fonction du fichier de la base de données des projets.

- L'application 1, avec 5 parts, reçoit 50 pour cent de l'UC.
- L'application 2, avec 3 parts, reçoit 30 pour cent de l'UC.
- L'application 3, avec 2 parts, reçoit 20 pour cent de l'UC.

Le schéma indiqué ci-après présente le fonctionnement de cette configuration en situation normale et en cas de basculement.



Basculement du groupe de ressources uniquement

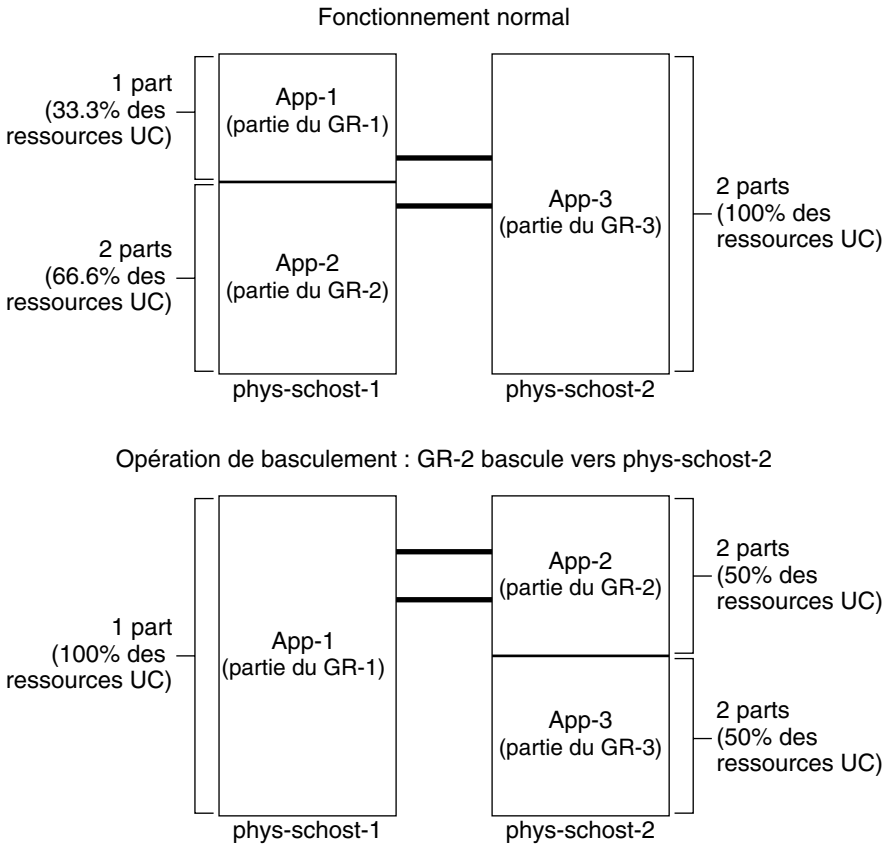
Dans une configuration où plusieurs groupes de ressources ont le même maître par défaut, un groupe de ressources (et ses applications associées) peuvent basculer ou commuter sur un noeud secondaire. Pendant ce temps, le maître par défaut continue de fonctionner dans le cluster.

Remarque – durant un basculement, l'application basculant se voit attribuer des ressources tel que spécifié dans le fichier de configuration du noeud secondaire. Dans cet exemple, les fichiers de la base de données de projets des noeuds principaux et secondaires ont la même configuration.

Par exemple, cet échantillon de fichier de configuration spécifie que 1 part est allouée à l'application, 2 parts sont allouées à l'application 2 et 2 parts à l'application 3.

```
Prj_1:106:project for App_1:root::project.cpu-shares=(privileged,1,none)
Prj_2:107:project for App_2:root::project.cpu-shares=(privileged,2,none)
Prj_3:108:project for App_3:root::project.cpu-shares=(privileged,2,none)
```

Le schéma suivant illustre le fonctionnement de cette configuration en situation normale et en cas de basculement, où GR-2 contenant l'application 2 bascule sur *phys-schost-2*. Notez que le nombre de parts assignées ne change pas. Toutefois, le pourcentage de temps UC disponible pour chaque application peut varier en fonction du nombre de parts assignées à chaque application nécessitant du temps UC.



Adaptateurs de réseau public et multi-acheminement sur réseau IP

Les clients adressent des requêtes de données au cluster à travers le réseau public. Chaque noeud du cluster est connecté à au moins un réseau public à travers une paire d'adaptateurs de réseau public.

Le logiciel IPMP (Internet Protocol Network Multipathing) Solaris de Sun Cluster intègre les mécanismes de base permettant le contrôle des adaptateurs de réseaux publics et le basculement des adresses IP d'un adaptateur à l'autre, lorsqu'une panne est détectée. Chaque noeud du cluster a sa propre configuration de multi-acheminement sur réseau IP , pouvant être différente de celle des autres noeuds.

Les adaptateurs de réseau public sont organisés en *groupes IPMP* (groupes de multi-acheminement). Chaque groupe de multi-acheminement possède un ou plusieurs adaptateurs de réseau public. Chaque adaptateur d'un groupe de multi-acheminement peut être actif ; vous pouvez aussi configurer des interfaces de réserve qui restent inactives jusqu'au moment d'un basculement. Le démon de multi-acheminement `in.mpathd` utilise une adresse IP de test pour détecter les pannes et réparations. S'il détecte une panne sur l'un des adaptateurs, un basculement a lieu. Tous les accès réseau basculent de l'adaptateur défaillant sur un adaptateur en état de marche du groupe de multi-acheminement et la connexion du noeud au réseau public est ainsi maintenue. Si une interface de réserve a été configurée, le démon choisit l'interface de réserve. Sinon, `in.mpathd` choisit l'interface ayant le moins d'adresses IP. Le basculement se produisant au niveau de l'interface de l'adaptateur, les connexions de niveaux supérieurs telles que TCP ne sont pas affectées, sauf pendant un laps de temps très bref, au moment du basculement. Lorsque le basculement des adresses IP se fait dans de bonnes conditions, des diffusions ARP sont envoyées en supplément. La connexion aux clients distants est donc maintenue.

Remarque – le protocole TCP possédant des fonctions de récupération des encombrements, ses extrémités peuvent souffrir de retards supplémentaires après un basculement réussi, car certains segments peuvent se perdre au cours du basculement et provoquer l'activation du mécanisme de contrôle des encombrements.

Les groupes de multi-acheminement procurent aux blocs assemblés des ressources de nom d'hôte logique et d'adresse partagée. Vous pouvez aussi créer des groupes de multi-acheminement indépendamment des ressources de nom d'hôte logique et d'adresse partagée pour contrôler la connexion des noeuds du cluster au réseau public. Sur un noeud, le même groupe de multi-acheminement peut héberger un nombre indéfini de ressources de nom d'hôte logique ou d'adresse partagée. Pour de plus amples informations sur les ressources de nom d'hôtes logiques et d'adresses partagées, reportez-vous au document *Sun Cluster 3.1 Data Services Installation and Configuration Guide*.

Remarque – le mécanisme de multi-acheminement sur réseau IP a été conçu pour détecter et masquer les pannes des adaptateurs. Il n'est pas destiné à récupérer d'une erreur d'un administrateur utilisant la commande `ifconfig(1M)` pour supprimer une des adresses IP logiques (ou partagées). Le logiciel Sun Cluster affiche les adresses IP logiques et partagées comme des ressources gérées par le RGM. L'administrateur doit ajouter ou supprimer une adresse IP à l'aide de la commande `scrgadm`(1M) pour modifier le groupe de ressources contenant la ressource.

Pour de plus amples informations sur la fonction de multi-acheminement sur réseau IP de Solaris, reportez-vous à la documentation adéquate de l'environnement d'exploitation Solaris installé sur votre cluster.

Version de l'environnement d'exploitation	Pour les instructions, voir...
Environnement d'exploitation Solaris 8	*IP Network Multipathing Administration Guide*
Environnement d'exploitation Solaris 9	"IP Network Multipathing Topics" dans le document *System Administration Guide: IP Services*

Prise en charge de la reconfiguration dynamique

La prise en charge de la fonction de reconfiguration dynamique (DR) par Sun Cluster 3.1 est développée par phases successives. Cette rubrique propose des explications et des observations concernant la prise en charge de la fonction DR par Sun Cluster 3.1.

Notez que toutes les exigences de configuration, procédures et restrictions applicables à la reconfiguration dynamique (DR) de Solaris s'appliquent également à la DR de Sun Cluster (à l'exception de l'opération de quiescence de l'environnement d'exploitation). Reportez-vous donc à la documentation relative à la DR de Solaris *avant* d'utiliser la fonction DR du logiciel Sun Cluster. Relisez surtout les conditions applicables aux périphériques ES hors réseau dans le cadre d'une opération DR de détachement. Les documents *Sun Enterprise 10000 Dynamic Reconfiguration User Guide* et *Sun Enterprise 10000 Dynamic Reconfiguration Reference Manual* (des collections *Solaris 8 on Sun Hardware* ou *Solaris 9 on Sun Hardware*) peuvent tous deux être téléchargés à l'adresse http://docs.sun.com.

Description générale de la reconfiguration dynamique

La fonction de reconfiguration dynamique permet de supprimer des éléments matériels sur des systèmes en cours de fonctionnement. Les processus de la reconfiguration dynamique ont pour but d'assurer la continuité du fonctionnement du système. Il n'est pas nécessaire d'arrêter le système et d'interrompre la disponibilité du cluster.

La reconfiguration dynamique (DR) opère au niveau de la carte. Elle affecte ainsi tous les composants de la carte. Chaque carte peut contenir plusieurs composants, notamment l'unité centrale, la mémoire et les jonctions de périphériques des lecteurs de disques, lecteurs de bandes et connexions réseau.

La suppression d'une carte contenant des composants actifs entraînerait des erreurs système. Avant la suppression d'une carte, le sous-système de la DR interroge d'autres systèmes, tels que Sun Cluster, pour déterminer si les composants de la carte sont en cours d'utilisation. Si le sous-système de la DR découvre qu'une carte est en cours d'utilisation, l'opération de suppression de la carte n'est pas effectuée. Ainsi, la suppression d'une carte à l'aide de la DR n'est jamais risquée puisque le sous-système de la DR rejette l'opération si la carte contient des composants actifs.

Une opération d'ajout de carte avec la DR est également sans danger. Les unités centrales et la mémoire d'une nouvelle carte sont automatiquement mises en service par le système. Toutefois, l'administrateur système doit configurer manuellement le cluster pour pouvoir utiliser les composants de la nouvelle carte de manière active.

Remarque – le sous-système de la DR comporte plusieurs niveaux. Si un niveau inférieur rapporte une erreur, le niveau supérieur fait de même. Toutefois, alors que le niveau inférieur décrit l'erreur, le niveau supérieur se contente d'indiquer "Unknown error." Cette dernière doit être ignorée des administrateurs système.

Les rubriques suivantes présentent quelques observations sur les conséquences de l'utilisation de la reconfiguration dynamique avec différents types de périphériques.

Reconfiguration dynamique et unités centrales

Le logiciel Sun Cluster ne rejette pas une opération de suppression de carte par reconfiguration automatique en raison de la présence d'unités centrales.

Lorsqu'une opération d'ajout de carte par reconfiguration dynamique a réussi, les unités centrales de la carte ajoutée sont automatiquement incorporées au système.

Reconfiguration dynamique et mémoire

Dans le cadre de la reconfiguration dynamique, il faut distinguer deux types de mémoire. Ces deux types ne diffèrent qu'au niveau de l'utilisation. Le matériel utilisé est le même.

La mémoire utilisée par le système d'exploitation est appelée panier de la mémoire du noyau. Le logiciel Sun Cluster ne prend pas en charge la suppression d'une carte contenant le panier de la mémoire du noyau, il rejette ce type d'opération. Lorsque l'opération de suppression concerne une autre mémoire, Sun Cluster ne rejette pas l'opération.

Lorsqu'une carte qui contient de la mémoire est ajoutée par reconfiguration dynamique, la mémoire de cette carte est automatiquement incorporée au système.

Reconfiguration dynamique et lecteurs de disques et de bandes

Sun Cluster rejette les opérations de suppression de cartes par reconfiguration dynamique sur les lecteurs actifs dans le noeud principal. Ces opérations peuvent être exécutées sur des lecteurs non actifs dans le noeud principal et sur n'importe quel lecteur du noeud secondaire. Après l'opération DR, l'accès aux données du cluster se poursuit comme auparavant.

Remarque – Sun Cluster rejette les opérations DR ayant un impact sur la disponibilité des périphériques de quorum. Pour obtenir des informations sur les périphériques de quorum et sur la procédure de reconfiguration dynamique les concernant, reportez-vous à la rubrique « Reconfiguration dynamique et périphériques de quorum » à la page 86.

Reportez-vous au *Guide d'administration système du logiciel Sun Cluster 3.1* pour des directives détaillées concernant la réalisation de ces actions.

Reconfiguration dynamique et périphériques de quorum

Si l'opération de suppression de carte par reconfiguration dynamique concerne une carte contenant une interface vers un périphérique de quorum, Sun Cluster rejette l'opération et identifie le périphérique de quorum qui serait affecté par l'opération. Vous devez désactiver la fonction de quorum du périphérique avant d'effectuer une opération de suppression de carte par reconfiguration dynamique.

Reportez-vous au *Guide d'administration système du logiciel Sun Cluster 3.1* pour des directives détaillées concernant la réalisation de ces actions.

Reconfiguration dynamique et interfaces d'interconnexion du Cluster

Si l'opération de suppression de carte par reconfiguration dynamique concerne une carte contenant une interface d'interconnexion du cluster active, Sun Cluster rejette l'opération et identifie l'interface qui serait affectée par l'opération. Vous devez utiliser un outil d'administration de Sun Cluster pour désactiver l'interface active avant l'opération de reconfiguration dynamique (voir également les précautions à prendre répertoriées ci-dessous).

Reportez-vous au *Guide d'administration système du logiciel Sun Cluster 3.1* pour des directives détaillées concernant la réalisation de ces actions.

Attention – Sun Cluster exige que chaque noeud du cluster possède au moins un chemin fonctionnel vers les autres noeuds du cluster. Ne désactivez pas une interface d'interconnexion privée prenant en charge le dernier chemin d'accès à un noeud du cluster.

Reconfiguration dynamique et interfaces de réseau public

Si l'opération de suppression de carte par reconfiguration dynamique concerne une carte contenant une interface de réseau public active, Sun Cluster rejette l'opération et identifie l'interface qui serait affectée par l'opération. Avant de supprimer une carte où se trouve une interface de réseau active, tout le trafic sur cette interface doit être basculé vers une autre interface active du groupe de multi-acheminement à l'aide la commande `if_mpadm`(1M).

Attention – si l'adaptateur de réseau restant échoue au moment où vous supprimez l'adaptateur de réseau désactivé par reconfiguration dynamique, la disponibilité du cluster peut être menacée. L'adaptateur restant ne peut pas effectuer de basculement pendant toute la durée de l'opération DR.

Reportez-vous au *Guide d'administration système du logiciel Sun Cluster 3.1* pour obtenir des directives détaillées sur la procédure de suppression par reconfiguration dynamique d'une interface de réseau public.

CHAPITRE 4

Questions récurrentes

Ce chapitre contient les réponses aux questions les plus fréquemment posées sur le système SunPlex. Ces questions sont classées par thème.

Questions récurrentes concernant la haute disponibilité

- **Qu'est-ce qu'un système hautement disponible ?**

  Sur le système SunPlex, la haute disponibilité (HD) est la capacité d'un cluster à maintenir une application active, même lors d'une panne entraînant normalement l'indisponibilité du système serveur.

- **Quel est le processus grâce auquel le cluster peut fournir la haute disponibilité ?**

  Le cluster fournit un environnement hautement disponible à travers un processus appelé basculement. Le basculement est une série d'actions exécutées par le cluster pour déplacer les ressources d'un service de données d'un noeud défectueux vers un autre noeud actif du cluster.

- **Quelle est la différence entre un service de données évolutif et un service de données de basculement ?**

  Il existe deux types de services de données à haute disponibilité : le service de basculement et le service évolutif.

  Un service de données de basculement exécute une application sur un seul noeud principal du cluster à la fois. Les autres noeuds peuvent exécuter d'autres applications, mais chaque application ne tourne que sur un seul noeud. Si un noeud principal tombe en panne, les applications fonctionnant sur ce noeud basculent vers un autre noeud et continuent de fonctionner.

Un service évolutif exécute une application sur plusieurs noeuds pour créer un service logique unique. Les services évolutifs utilisent tous les noeuds et processeurs du cluster sur lequel ils tournent.

Pour chaque application, un noeud héberge l'interface physique pour le cluster. Ce noeud s'appelle noeud d'interface globale (ou noeud GIF). Il peut y avoir plusieurs noeuds GIF dans le cluster. Chacun héberge une ou plusieurs interfaces logiques pouvant être utilisées par les services évolutifs. Ces interfaces logiques sont appelées *interfaces globales*. Un noeud GIF héberge une interface globale pour toutes les requêtes d'une application particulière et les distribue aux noeuds sur lesquels tourne le serveur d'application. Si le noeud GIF tombe en panne, l'interface globale bascule sur un noeud encore actif.

Si l'un des noeuds sur lequel tourne l'application tombe en panne, celle-ci continue de fonctionner sur les autres noeuds avec une certaine dégradation des performances jusqu'à ce que le noeud défectueux retourne dans le cluster.

Questions récurrentes concernant les systèmes de fichiers

- **Puis-je faire fonctionner un ou plusieurs noeuds de cluster comme serveur(s) NFS hautement disponible(s) avec d'autres noeuds comme clients ?**

  Non, il ne faut pas faire de montage en boucle.

- **Puis-je utiliser un système de fichiers de cluster pour des applications n'étant pas sous le contrôle du Gestionnaire du groupe de ressources ?**

  Oui. Toutefois, sans le contrôle du RGM, les applications doivent être relancées manuellement après l'échec du noeud sur lequel elles fonctionnent.

- **Tous les systèmes de fichiers de cluster doivent-ils avoir un point de montage sous le répertoire /global ?**

  Non. Toutefois, en plaçant tous les systèmes de fichiers de cluster sous le même point de montage, par exemple `/global`, vous en faciliterez l'organisation et la gestion.

- **Quelle différence y-a-t-il entre l'utilisation des systèmes de fichiers de cluster et l'exportation de systèmes de fichiers NFS ?**

  Il y en a plusieurs :

  1. Le système de fichiers de cluster prend en charge les périphériques globaux. NFS ne prend pas en charge l'accès distant aux périphériques.
  2. Le système de fichiers de cluster a un espace de noms global. Seulement une commande de montage est requise. Avec NFS, vous devez monter le système de fichiers sur chaque noeud.

3. Le système de fichiers de cluster met en cache des fichiers plus souvent que ne le fait NFS. Par exemple, chaque fois que plusieurs noeuds accèdent à un fichier pour des opérations de lecture, écriture, verrouillage, E/S asynchrones.
4. Le système de fichiers de cluster prend en charge les basculements sans interruption du service si un serveur échoue. NFS prend en charge plusieurs serveurs, mais le basculement n'est possible que pour les systèmes de fichiers à lecture seule.
5. Le système de fichiers de cluster a été conçu pour exploiter les futures capacités d'interconnexion rapide du cluster apportant des fonctions de DMA distant et de zéro copie.
6. Si vous modifiez les attributs d'un fichier (à l'aide de `chmod(1M)` , par exemple) d'un système de fichiers de cluster, la modification se reflète immédiatement sur tous les noeuds. Avec un système de fichiers NFS exporté, cette opération peut prendre beaucoup plus de temps.

- **Le système de fichiers /global/.devices/node@<nodeID> apparaît sur les noeuds de mon cluster. Puis-je l'utiliser pour stocker des données destinées à être hautement disponibles et globales ?**

  Ces systèmes de fichiers stockent l'espace de noms des périphériques globaux. Ils ne sont pas destinés à une utilisation générale. Bien qu'ils soient globaux, on n'y accède jamais de manière globale -- chaque noeud accède uniquement à son propre espace de noms de périphérique global. Si un noeud a échoué, les autres ne peuvent pas accéder à son espace de noms. Ces systèmes de fichiers ne sont pas hautement disponibles. Ils ne doivent pas être utilisés pour stocker des données accessibles globalement ou hautement disponibles.

Questions récurrentes concernant la gestion de volumes

- **Dois-je mettre en miroir tous les périphériques de disque ?**

  Pour qu'un périphérique de disque soit considéré comme hautement disponible, il doit être mis en miroir ou utiliser le matériel RAID 5. Tous les services de données doivent utiliser des périphériques de disque hautement disponibles ou des systèmes de fichiers de cluster montés sur ce type de périphériques. Ces configurations peuvent supporter des pannes de disque uniques.

- **Puis-je utiliser deux gestionnaires de volumes différents pour les disques locaux (disque d'initialisation) et pour les disques multihôtes ?**

  Cette configuration est prise en charge par le logiciel Solaris Volume Manager, gérant les disques locaux et VERITAS Volume Manager, gérant les disques multihôtes. Aucune autre combinaison n'est prise en charge.

Questions récurrentes concernant les services de données

- **Quels services de données SunPlex sont disponibles ?**

  Vous trouverez une liste des services de données pris en charge dans les *Notes de version de Sun Cluster 3.1*.

- **Quelles versions d'application sont prises en charge par les services de données SunPlex ?**

  Vous trouverez une liste des versions d'application prises en charge dans les *Notes de version de Sun Cluster 3.1*.

- **Puis-je écrire sur mon propre service de données ?**

  Oui. Pour de plus amples informations, reportez-vous au document *Sun Cluster 3.1 Data Services Developer's Guide* et à la documentation concernant les technologies d'activation des services de données fournie avec l'API de la bibliothèque de développement des services de données (API DSDL).

- **Lors de la création des ressources réseau, dois-je spécifier les adresses IP numériques ou les noms d'hôtes ?**

  Le meilleur moyen de spécifier les ressources réseau est d'utiliser le nom d'hôte UNIX plutôt que l'adresse IP numérique.

- **Lors de la création de ressources réseau, quelle est la différence entre l'utilisation d'un nom d'hôte logique (ressource LogicalHostname) et l'utilisation d'une adresse partagée (ressource SharedAddress) ?**

  Excepté pour le cas de Sun Cluster HA pour NFS, lorsque la documentation requiert l'utilisation d'une ressource `LogicalHostname` dans un groupe de ressources en mode `basculement`, les ressources `SharedAddress` et `LogicalHostname` peuvent être utilisées indifféremment. L'utilisation d'une ressource `SharedAddress` implique un temps système accru car le cluster est configuré pour une ressource `SharedAddress` et non pour `LogicalHostname`.

  Vous avez intérêt à utiliser une ressource `SharedAddress` lorsque vous configurez à la fois des services de données évolutifs et de basculement, et souhaitez que les clients puissent y accéder à l'aide du même nom d'hôte. Dans ce cas, la/les ressource(s) `SharedAddress` et les ressources d'application de basculement sont contenues dans un groupe de ressources, tandis que la ressource du service de données évolutif est contenue dans un groupe de ressources à part et configuré pour utiliser `SharedAddress`. Les services évolutifs et de basculement peuvent ensuite utiliser le même ensemble de noms d'hôtes/adresses configurés dans la ressource `SharedAddress`.

Questions récurrentes concernant le réseau public

- **Quels adaptateurs de réseau public le système SunPlex prend-il en charge ?**

  À l'heure actuelle, le système SunPlex prend en charge les adaptateurs de réseau public Ethernet (10/100BASE-T et 1000BASE-SX Gb). De nouvelles interfaces pouvant être prises en charge dans le futur, informez-vous auprès de votre représentant Sun.

- **Quel est le rôle de l'adresse MAC au cours d'un basculement ?**

  Lorsqu'un basculement se produit, de nouveaux paquets du protocole ARP (Address Resolution Protocol) sont générés et diffusés. Ils contiennent la nouvelle adresse MAC (du nouvel adaptateur physique vers lequel le noeud a basculé) et l'ancienne adresse IP. Lorsqu'une autre machine sur le réseau reçoit un de ces paquets, elle débarrasse sa mémoire cache ARP de l'ancienne correspondance MAC-IP et utilise la nouvelle.

- **Le système SunPlex prend-il en charge le paramètre local-mac-address?=true dans la PROM OpenBoot™ (OBP) d'un adaptateur d'hôte ?**

  Oui. En fait, le multi-acheminement sur réseau IP requiert la définition de `local-mac-address?` sur `true`.

- **Quel délai d'attente faut-il compter lorsque le multi-acheminement sur réseau IP commute des adaptateurs ?**

  Ce délai d'attente peut être de quelques minutes. Il est dû au fait qu'en cas de commutation du multi-acheminement sur réseau IP , un ARP est envoyé en supplément. Il n'y a cependant aucune garantie que le routeur entre le client et le cluster utilise cet ARP. En effet, tant que l'entrée du cache ARP pour cette adresse IP n'est pas arrivée au bout de son délai d'exécution sur le routeur, il est possible que celui-ci utilise l'ancienne adresse MAC.

- **Combien de temps faut-il pour détecter la panne d'un adaptateur réseau ?**

  Le temps de détection de la panne par défaut est de 10 secondes. L'algorithme essaie de se conformer à ce temps de détection, mais le temps réel dépend de la charge du réseau.

Questions récurrentes concernant les membres du cluster

- **Tous les membres du cluster doivent-ils avoir le même mot de passe racine ?**

Non, ce n'est pas obligatoire. Toutefois, en adoptant le même mot de passe, vous simplifierez l'administration du cluster.

- **L'ordre dans lequel les noeuds sont initialisés est-il important ?**

  Dans la plupart des cas, non. L'ordre d'initialisation des noeuds est cependant important pour prévenir des phénomènes d'amnésie (reportez-vous à la rubrique « Quorum et périphériques de quorum » à la page 52 pour de plus amples détails sur l'amnésie). Par exemple, si le noeud 2 est propriétaire du périphérique de quorum et que le noeud 1 est arrêté, si vous arrêtez ensuite le noeud 2, vous devrez l'activer de nouveau avant d'activer le noeud 1. Ceci vous empêche d'activer accidentellement un noeud où les informations de configuration ne sont pas à jour.

- **Dois-je mettre en miroir les disques locaux d'un noeud du cluster ?**

  Oui. Bien que cette mise en miroir ne soit pas obligatoire, elle empêche que l'échec d'un disque non mis en miroir n'arrête le noeud. Cette mise en miroir a cependant pour conséquence d'augmenter la charge d'administration du système.

- **Quels sont les problèmes liés à la sauvegarde des membres du cluster ?**

  Il existe plusieurs méthodes de sauvegarde d'un cluster. Une des méthodes consiste à avoir comme noeud de sauvegarde un noeud auquel est rattaché un lecteur/bibliothèque de bandes. Le système de fichiers de cluster est ensuite utilisé pour sauvegarder les données. Ne connectez pas ce noeud aux disques partagés.

  Reportez-vous au *Guide d'administration système de Sun Cluster 3.1* pour obtenir des informations complémentaires sur les procédures de sauvegarde et de restauration.

- **Quand un noeud est-il suffisamment sain pour être utilisé comme noeud secondaire ?**

  Après une réinitialisation, un noeud est suffisamment sain pour devenir noeud secondaire lorsqu'il affiche l'invite de connexion.

Questions récurrentes concernant le stockage du cluster

- **Qu'est-ce qui rend le stockage multihôte hautement disponible ?**

  Le stockage multihôte est hautement disponible parce qu'il peut survivre à la perte d'un disque, grâce à la mise en miroir (ou grâce aux contrôleurs matériels RAID-5). Un périphérique de stockage multihôte ayant plus d'une connexion hôte, il peut aussi résister à la perte d'un noeud auquel il est connecté. En outre, chaque noeud possède des chemins d'accès redondants vers le périphérique de stockage auquel il est relié afin de supporter les pannes des adaptateurs de bus hôtes, des câbles ou des contrôleurs de disques.

Questions récurrentes concernant l'interconnexion de cluster

- **Quelles interconnexions de cluster le système SunPlex prend-il en charge ?**

  À l'heure actuelle, le système SunPlex prend en charge Ethernet (100BASE-T Fast Ethernet et 1000BASE-SX Gb) et les interconnexions de cluster de l'interface de réseau SCSI.

- **Quelle est la différence entre un « câble » et un « chemin » de transport ?**

  Les câbles de transport de cluster sont configurés pour utiliser des adaptateurs et commutateurs de transport. Les câbles relient les adaptateurs et les commutateurs d'un composant à l'autre. Le gestionnaire de la topologie du cluster utilise les câbles disponibles pour établir des chemins d'accès de bout en bout entre les noeuds. Un câble ne correspond pas directement à un chemin d'accès.

  Les câbles sont « activés » et « désactivés » de manière statique par un administrateur. Les câbles ont un « état », (activé ou désactivé) mais pas de « statut ». Si un câble est désactivé, c'est comme s'il n'était pas configuré. Les câbles désactivés ne peuvent pas être utilisés comme chemins d'accès. Ils ne sont pas sondés et il est ainsi impossible de connaître leur statut. L'état d'un câble peut être affiché à l'aide de la commande `scconf -p`.

  Les chemins d'accès sont définis de manière dynamique par le gestionnaire de la topologie du cluster, déterminant également leur « statut ». Un chemin peut avoir un statut « en ligne » ou « hors ligne », consultable à l'aide de la commande `scstat(1M)`.

  Prenons l'exemple suivant d'un cluster à deux noeuds avec quatre câbles.

```
noeud 1 : adaptateur 0     vers commutateur 1, port 0
noeud 1 : adaptateur 1     vers commutateur 2, port 0
noeud 2 : adaptateur 0     vers commutateur 1, port 1
noeud 2 : adaptateur 1     vers commutateur 2, port 1
```

  Deux chemins d'accès possibles peuvent être formés à partir de ces quatre câbles.

```
noeud 1 : adaptateur 0    vers noeud 2 : adaptateur 0
noeud 2 : adaptateur 1    vers noeud 2 : adaptateur 1
```

Questions récurrentes concernant les systèmes client

- **Dois-je prendre en compte certains besoins ou restrictions spécifiques au client dans le cadre de l'utilisation d'un cluster ?**

Les systèmes client se connectent au cluster comme à n'importe quel autre serveur. En fonction de l'application du service de données, vous pouvez dans certains cas avoir à installer un logiciel client ou procéder à d'autres modifications de configuration pour que le client puisse se connecter à l'application du service de données. Reportez-vous aux chapitres individuels du document *Sun Cluster 3.1 Data Services Installation and Configuration Guide* pour de plus amples informations sur les exigences de configuration client.

Questions récurrentes concernant la console administrative

- **Le système SunPlex requiert-il une console administrative ?**

  Oui.

- **La console administrative doit-elle être dédiée au cluster ou peut-elle être utilisée pour d'autres tâches ?**

  Le système SunPlex n'exige pas l'utilisation d'une console administrative dédiée ; cette dernière présente cependant les avantages suivants :

  - Elle permet une gestion centralisée des clusters en regroupant les outils de gestion et de console sur la même machine.
  - Elle peut permettre une résolution plus rapide des problèmes par votre fournisseur de services matériels.

- **La console administrative a-t-elle besoin d'être située « près » du cluster, dans la même pièce par exemple ?**

  Consultez votre fournisseur de services. Ce dernier peut demander à ce que la console soit placée près du cluster. Il n'existe cependant aucune raison technique pour que la console soit placée dans la même pièce.

- **Une console administrative peut-elle servir plus d'un cluster, si les contraintes relatives à la distance sont respectées ?**

  Oui. Il est possible de contrôler plusieurs clusters à partir d'une seule console administrative. Vous pouvez aussi partager un seul concentrateur de terminaux entre les clusters.

Questions récurrentes concernant les concentrateurs de terminaux et les SSP

- **Le système SunPlex requiert-il un concentrateur de terminaux ?**

  Les versions Sun Cluster 3.0 et ultérieures peuvent fonctionner sans concentrateur de terminaux. Ce n'était pas le cas de la version Sun Cluster 2.2, nécessitant un concentrateur de terminaux pour la fonction de séparation en cas d'échec.

- **Je constate que la plupart des serveurs SunPlex utilisent un concentrateur de terminaux, mais le Serveur Sun Enterprise E10000 n'en utilise pas. Pour quelle raison ?**

  Le concentrateur de terminaux constitue en effet un convertisseur série/Ethernet pour la plupart des serveurs. Son port de console est un port série. Le Serveur Sun Enterprise E10000 n'a pas de console série. C'est le SSP (System Service Processor) qui sert de console, à travers un port Ethernet ou un port `jtag`. Dans le Serveur Sun Enterprise E10000, le SSP est toujours utilisé comme console.

- **Quels sont les avantages de l'utilisation d'un concentrateur de terminaux ?**

  L'utilisation d'un concentrateur de terminaux fournit un accès au niveau de la console à chaque noeud à partir d'une station de travail distante située n'importe où sur le réseau, même lorsque le noeud est sur la PROM OpenBoot (OBP).

- **Que dois-je savoir si je veux utiliser un concentrateur de terminaux non pris en charge par Sun ?**

  La principale différence entre le concentrateur de terminaux pris en charge par Sun et les autres consoles est que le premier contient des microprogrammes spéciaux empêchant le concentrateur de terminaux d'envoyer une coupure (ou break) à la console au moment où il s'initialise. Notez que si vous possédez une console susceptible d'envoyer une coupure, ou un signal pouvant être interprété comme tel à la console, le noeud est arrêté.

- **Puis-je libérer le port verrouillé d'un concentrateur de terminaux pris en charge par Sun sans pour autant le réinitialiser ?**

  Oui. Notez que le numéro de port qui a besoin d'être réinitialisé et procédez de la manière suivante :

  ```
  telnet tc
  Entrez le nom ou le numéro de port annexe : cli
  annexe : su -
  annex# admin
  admin : reset numéro_port
  admin : quit
  annex# hangup
  #
  ```

  Reportez-vous au *Guide d'administration système de Sun Cluster 3.1* pour obtenir de plus amples informations sur la configuration et l'administration du concentrateur de terminaux pris en charge par Sun.

- **Que se passe-t-il si le concentrateur de terminaux tombe en panne ? Dois-je en avoir un autre à disposition ?**

  Non. Le cluster garde toute sa disponibilité en cas de panne du concentrateur de terminaux. Vous perdez la capacité de vous connecter aux consoles du noeud jusqu'à ce que le concentrateur soit de nouveau en service.

- **Qu'en est-il de la sécurité lorsque j'utilise un concentrateur de terminaux ?**

  Généralement, le concentrateur de terminaux est relié à un petit réseau utilisé par les administrateurs système et non accessible aux autres clients. Vous pouvez contrôler la sécurité en limitant l'accès à ce réseau particulier.

- **Comment utiliser la reconfiguration dynamique avec un lecteur de disques ou de bandes ?**

  - Déterminez si le lecteur de disques ou de bandes fait partie d'un groupe de périphériques actif. Si ce n'est pas le cas, des opérations de suppression par reconfiguration dynamique peuvent être effectuées sur ce lecteur.
  - Si l'opération de suppression par reconfiguration dynamique touche un lecteur de disques ou de bandes actif, le système rejette l'opération et identifie les lecteurs susceptibles d'être affectés par cette opération. Si le lecteur fait partie d'un groupe de périphériques actif, allez à la rubrique « Reconfiguration dynamique et lecteurs de disques et de bandes » à la page 85.
  - Déterminez si le lecteur est un composant du noeud principal ou du noeud secondaire. Si le lecteur est un composant du noeud secondaire, vous pouvez y effectuer l'opération de suppression par reconfiguration dynamique.
  - Si le lecteur est un composant du noeud principal, vous devez commuter les noeuds principal et secondaire avant d'effectuer l'opération de suppression par reconfiguration dynamique sur ce périphérique.

Attention – tout échec sur le noeud principal, alors que vous effectuez une opération DR sur un noeud secondaire, a une incidence sur la disponibilité du cluster. Le noeud principal n'a pas d'endroit pour basculer, jusqu'à ce qu'un nouveau noeud secondaire soit défini.

Glossaire

Ce glossaire s'applique à la documentation de SunPlex 3.1.

A

adaptateur de transport de cluster — Adaptateur réseau résidant sur un noeud et connectant ce noeud à l'interconnexion de cluster. Voir aussi « interconnexion de cluster ».

amnésie — État dans lequel redémarre un cluster après un arrêt, avec des données de configuration (CCR) obsolètes. Par exemple, sur un cluster à deux noeuds où seul le noeud 1 est opérationnel, si un changement de configuration a lieu sur le noeud 1, le CCR du noeud 2 devient obsolète. Si le cluster est arrêté puis relancé sur le noeud 2, on se trouve dans une situation d'amnésie du fait de l'obsolescence du CCR du noeud 2.

API de gestion de ressources (API GR) — Interface de programme d'application dans un système SunPlex rendant une application hautement disponible dans un environnement de cluster.

appel de ressource — Instance d'un type de ressource fonctionnant sur un noeud. Concept abstrait représentant une ressource ayant été démarrée sur le noeud.

B

basculement Déplacement automatique d'un groupe de ressources ou de périphériques d'un noeud principal défectueux vers un nouveau noeud principal.

C

câbles de transport de cluster Connexion réseau reliant les extrémités. Connexion entre les adaptateurs et les jonctions de transport ou entre deux adaptateurs de transport de cluster. Voir aussi « interconnexion de cluster ».

CCR (Cluster Configuration Repository) Mémoire de données répliquées à haute disponibilité utilisée par le logiciel Sun Cluster pour stocker en continu les informations de configuration du cluster.

cluster Au moins deux noeuds ou domaines interconnectés partageant un système de fichiers de cluster et étant configurés ensemble pour exécuter des ressources parallèles, évolutives ou de basculement.

co-localisation Propriété d'être sur le même noeud. Ce concept est utilisé au moment de la configuration du cluster dans le but d'améliorer les performances.

commutation Transfert ordonné d'un groupe de ressources ou de périphériques d'un maître (noeud) du cluster vers un autre maître (ou plusieurs maîtres, si les groupes de ressources sont configurés pour plusieurs noeuds principaux). Une commutation peut être effectuée par l'administrateur à l'aide de la commande `scswitch(1M)`.

concentrateur de terminaux Périphérique externe au cluster, spécialement utilisé pour communiquer avec les membres du cluster dans une configuration autre que Enterprise 10000.

console administrative Station de travail permettant de faire fonctionner les logiciels d'administration du cluster.

D

détecteur de pannes Démon et programmes utilisés pour analyser diverses parties des services de données et prendre des mesures si nécessaire. Voir aussi « moniteur de ressources ».

disque local Disque physiquement réservé à un noeud de cluster donné.

disque multihôte Disque physiquement connecté à plusieurs noeuds.

E

ensemble de disques Voir « groupe de périphériques ».

équilibrage de la charge S'applique uniquement aux services évolutifs. Processus visant à répartir la charge de l'application entre les noeuds du cluster de façon à ce que les requêtes client soient servies rapidement. Reportez-vous à la rubrique « Services de données évolutifs » à la page 61 pour de plus amples informations.

espace de noms de périphérique global Espace de noms contenant les noms logiques des périphériques globaux dans le cluster. Dans l'environnement Solaris, les périphériques locaux sont définis dans les répertoires `/dev/dsk`, `/dev/rdsk` et `/dev/rmt`. L'espace de noms de périphérique global définit les périphériques globaux dans les répertoires `/dev/global/dsk`, `/dev/global/rdsk` et `/dev/global/rmt`.

état de la ressource État d'une ressource du Gestionnaire du groupe de ressources sur un noeud donné.

état du groupe de ressources État du groupe de ressources sur un noeud donné.

évènement Modification de l'état, du contrôle, de la gravité ou de la description d'un objet géré.

extrémité Port physique situé sur un adaptateur ou une jonction de transport du cluster.

G

gestionnaire de verrouillage réparti (GVR) Le logiciel de verrouillage est utilisé dans un environnement Serveur Oracle Parallel (OPS) de disques partagés. Le GVR permet aux processus Oracle exécutés sur des noeuds différents de synchroniser l'accès à la base de données. Il est conçu pour fonctionner en haute disponibilité. Lorsqu'un processus ou un noeud tombe en panne, il n'est pas nécessaire d'arrêter et de redémarrer les noeuds restants. Une reconfiguration rapide du GVR est alors exécutée pour assurer la récupération.

gestionnaire de volumes Logiciel assurant la fiabilité des données à travers l'entrelacement, la concaténation, la mise en miroir et la croissance dynamique des métapériphériques ou volumes.

gestionnaire du groupe de ressources (RGM) Logiciel permettant de rendre les ressources du cluster hautement disponibles et évolutives en les démarrant et les arrêtant automatiquement sur des noeuds de cluster déterminés. Le RGM agit en fonction de politiques pré-configurées, dans le cas de pannes ou de réinitialisations matérielles ou logicielles.

groupe de disques Voir « groupe de périphériques ».

groupe de multi-acheminement sur réseau IP Ensemble d'un ou plusieurs adaptateurs réseau sur le même noeud et le même sous-réseau, configurés pour se servir mutuellement de secours en cas de panne d'un adaptateur.

groupe de périphériques Groupe de périphériques, tels que les disques, défini par l'utilisateur et pouvant être contrôlé à partir de différents noeuds dans une configuration de cluster à haute disponibilité. Ce groupe peut comporter des disques, des ensembles de disques Solaris Volume Manager et des groupes de disques VERITAS Volume Manager.

groupe de périphériques de disques Voir « groupe de périphériques ».

groupe de ressources Ensemble de ressources gérées par le gestionnaire du groupe de ressources (RGM) comme une unité. Toute ressource devant être gérée par le RGM doit être configurée dans un groupe de ressources. En général, les ressources connexes et interdépendantes sont regroupées.

groupe de sauvegarde Voir « groupe de multi-acheminement sur réseau IP ».

H

haute disponibilité Arrêt ou suppression ordonnée d'un noeud défectueux du cluster avant que son mauvais fonctionnement puisse endommage quoi que ce soit.

hôte logique Concept lié au logiciel Sun Cluster version 2.0 (minimum) incluant une application, les ensembles de disques ou les groupes de disques sur lesquels résident les données d'application et les adresses réseau utilisées pour accéder au cluster. Ce concept n'existe plus dans le système SunPlex. Reportez-vous aux rubriques « Groupes de périphériques de disques » à la page 40 et « Ressources, groupes de ressources et types de ressources » à la page 70 pour voir la nouvelle implémentation de ce concept dans le système SunPlex.

hôte multirésidant Hôte résidant sur plusieurs réseaux publics.

I

ID de périphérique (IDP) Mécanisme permettant d'identifier les périphériques accessibles via Solaris. Les ID de périphériques sont décrits dans la page de manuel `devid_get(3DEVID)`.

Le pilote d'IDP de Sun Cluster utilise les ID de périphériques pour déterminer la corrélation entre les noms d'hôtes logiques Solaris sur différents noeuds du cluster. Le pilote d'IDP sonde chaque périphérique pour trouver son ID. Si cet ID correspond à un autre périphérique situé ailleurs sur le cluster, le même IDP est donné à chacun d'eux. Si l'ID de périphérique n'existe pas dans le cluster, un nouveau nom IDP est attribué. Voir aussi « nom logique Solaris » et « Pilote d'IDP ».

instance Voir « appel de ressource ».

instance de service parallèle Instance d'un type de ressource parallèle fonctionnant sur un noeud individuel.

interconnexion de cluster Infrastructure de gestion de réseaux matérielle comprenant les câbles, les jonctions de transport et les adaptateurs de transport du cluster. Le logiciel Sun Cluster et les services de données utilisent cette infrastructure pour la communication au sein du cluster.

interface globale Interface de réseau globale hébergeant physiquement les adresses partagées. Voir aussi « adresse partagée ».

interface réseau logique Dans l'architecture internet, un hôte peut avoir une ou plusieurs adresses IP. Le logiciel Sun Cluster configure des interfaces réseau logiques supplémentaires pour établir une correspondance entre plusieurs interfaces réseau logiques et une interface réseau physique. Chaque interface réseau logique a une seule adresse IP. Cette correspondance permet à une interface réseau physique de répondre à plusieurs adresses IP. Cette correspondance permet aussi à l'adresse IP de se déplacer d'un membre du cluster à l'autre en cas de reprise ou de commutation sans interface matérielle supplémentaire.

interface SCI Matériel d'interconnexion à haute vitesse utilisé dans l'interconnexion du cluster.

IPv4 Protocole Internet, version 4. Il est parfois simplement appelé IP. Cette version prend en charge un espace d'adresses à 32 bits.

J

jonction de transport de cluster Commutateur matériel utilisé dans l'interconnexion de cluster. Voir aussi « interconnexion de cluster ».

L

(logiciel) Sun Cluster Partie logicielle du système SunPlex. (Voir SunPlex.)

M

maître Voir « noeud principal ».

maître par défaut Membre du cluster par défaut sur lequel un type de ressources de basculement est mis en ligne.

maître potentiel Voir « noeud principal potentiel ».

membre du cluster Membre actif de la représentation courante du cluster. Ce membre est capable de partager des ressources avec d'autres membres et de leur fournir des services ainsi qu'aux clients du cluster. Voir aussi « noeud de cluster ».

mise à niveau progressive Dans une configuration Sun Cluster, mise à niveau effectuée séquentiellement sur un noeud à la fois. Au cours d'une mise à niveau progressive, le cluster reste en production et les services continuent de fonctionner sur les autres noeuds.

mode non cluster État obtenu en initialisant un membre du cluster à l'aide de l'option d'initialisation `-x`. À cet état, le noeud n'est plus membre du cluster mais il demeure un noeud de cluster. Voir aussi « membre du cluster » et « noeud de cluster ».

moniteur d'appartenance au cluster (MAC) Logiciel maintenant une liste consistante des membres du cluster. Ces informations sont utilisées par le reste des logiciels de clustering pour décider où placer les services à haute disponibilité. Le moniteur MAC veille à ce que des membres n'appartenant pas au cluster n'altèrent pas les données et qu'ils ne transmettent pas de données altérées ou inconsistantes aux clients.

moniteur de ressources Partie optionnelle de l'implémentation d'un type de ressource sondant périodiquement les ressources afin de vérifier qu'elles fonctionnent correctement et quelles sont leurs performances.

multi-acheminement sur réseau IP Logiciel utilisant la détection de pannes et le basculement pour empêcher qu'un noeud ne devienne indisponible à cause d'une panne d'adaptateur ou câble réseau. Le système de basculement du multi-acheminement sur réseau IP utilise des ensembles d'adaptateurs réseau appelés groupes de multi-acheminement sur réseau IP (groupes de multi-acheminement) afin d'assurer des connexions redondantes entre un noeud du cluster et le réseau public. Les systèmes de détection de pannes et de basculement travaillent de concert pour garantir la disponibilité des ressources. Voir aussi « groupe de multi-acheminement sur réseau IP ».

N

noeud Machine ou domaine physique (dans le Serveur Sun Enterprise E10000) pouvant faire partie d'un système SunPlex. Il est aussi appelé « hôte ».

noeud de cluster Noeud configuré pour être membre du cluster. Un noeud de cluster peut ou non être un membre courant. Voir aussi « membre du cluster ».

noeud GIF Voir « noeud d'interface globale ».

noeud d'interface globale (GIF) Noeud hébergeant une interface globale.

noeud de réserve Noeud de cluster pouvant être converti en noeud secondaire si un basculement a lieu. Voir aussi « noeud secondaire ».

noeud principal Noeud sur lequel un groupe de ressources ou de périphériques est en ligne. Il s'agit donc du noeud hébergeant ou implémentant le service associé à la ressource. Voir aussi « noeud secondaire ».

noeud principal potentiel Membre du cluster pouvant contrôler un type de ressource de basculement en cas d'échec du noeud principal. Voir aussi « maître par défaut ».

noeud secondaire Membre du cluster pouvant contrôler les groupes de périphériques de disques et de ressources en cas d'échec du noeud principal. Voir aussi « noeud principal ».

nom d'hôte principal Nom d'un noeud du réseau public principal. Il s'agit toujours du nom de noeud spécifié dans `/etc/nodename`. Voir aussi « nom d'hôte secondaire ».

nom d'hôte privé Alias du nom d'hôte utilisé pour communiquer avec le noeud sur l'interconnexion de cluster.

nom d'hôte secondaire Nom utilisé pour accéder au noeud d'un réseau public secondaire. Voir aussi « nom d'hôte principal ».

nom IDP Permet d'identifier des périphériques globaux dans un système SunPlex. Il s'agit d'un identifiant de clustering ayant des relations avec un ou plusieurs noms logiques Solaris. Il prend la forme d*X*s*Y*, où *X* est un nombre entier et *Y* le nom de la tranche. Voir aussi « nom logique Solaris ».

nom logique Solaris Noms généralement utilisés pour gérer les périphériques Solaris. Pour les disques, ces noms ont généralement la forme `/dev/rdsk/c0t2d0s2`. À chacun de ces noms de périphériques logiques Solaris correspond un nom de périphérique physique Solaris sous-jacent. Voir aussi « nom IDP » et « nom physique Solaris ».

nom physique Solaris Nom donné à un périphérique par son pilote dans Solaris. Il apparaît sur une machine Solaris comme un chemin de l'arborescence de `/devices`. Un disque SCSI traditionnel a par exemple un nom physique Solaris similaire à :
`/devices/sbus@1f,0/SUNW,fas@e,8800000/sd@6,0:c,raw`

Voir aussi « nom logique Solaris ».

P

périphérique global Périphérique accessible depuis tous les membres du cluster, tels que les disques, CD et bandes.

périphérique de quorum Disque partagé entre plusieurs noeuds contribuant aux votes utilisés pour établir un quorum pour le fonctionnement du cluster. Le cluster ne peut fonctionner que si un quorum de votes a été établi. Le périphérique de quorum est utilisé lorsqu'un cluster est segmenté en plusieurs ensembles de noeuds pour déterminer quel ensemble de noeuds constitue le nouveau cluster.

pilote d'IDP Pilote intégré au logiciel Sun Cluster permettant d'avoir des espaces de noms de périphériques consistants au sein du cluster. Voir aussi « nom IDP ».

point de contrôle Notification envoyée par un noeud principal à un noeud secondaire pour maintenir le logiciel synchronisé entre eux. Voir aussi « noeud principal » et « noeud secondaire ».

propriété de type de ressource Couple clé-valeur, stocké par le RGM comme partie du type de ressource, utilisée pour décrire et gérer les ressources d'un type donné.

pulsation Message périodique envoyé à travers tous les chemins de transport d'interconnexion de cluster disponibles. Un manque de pulsation après un intervalle de temps et un nombre d'essais spécifiés peut entraîner le basculement interne d'une communication vers un autre chemin. En cas d'échec de tous les chemins d'accès vers un membre du cluster, le moniteur d'appartenance au cluster réévalue le quorum du cluster.

R

règle d'équilibrage de la charge S'applique uniquement aux services évolutifs. Manière retenue pour répartir la charge des requêtes d'applications entre les noeuds. Reportez-vous à la rubrique « Services de données évolutifs » à la page 61 pour de plus amples informations.

réplique de la base de données d'état des métapériphériques (réplique) Base de données stockée sur disque et enregistrant les informations relatives à la configuration et à l'état de tous les métapériphériques, ainsi que les situations d'erreur. Ces informations sont nécessaires au bon fonctionnement de l'ensemble de disques Solaris Volume Manager et sont répliquées.

reprise	Voir « basculement ».
ressource	Instance d'un type de ressource. Il peut y avoir plusieurs ressources du même type, chaque ressource ayant son propre nom et ensemble de valeurs de propriété, de façon à ce que plusieurs instances de l'application sous-jacente puissent tourner sur le cluster.
ressource d'adresse partagée	Adresse réseau pouvant être liée aux services évolutifs fonctionnant sur les noeuds du cluster afin de les mettre en corrélation avec ces noeuds. Un cluster peut avoir plusieurs adresses partagées et un service peut être lié à plusieurs adresses partagées.
ressource d'adresse réseau	Voir « ressource réseau ».
ressource de basculement	Ressource ne pouvant être correctement contrôlée que par un seul noeud à la fois. Voir aussi « ressource d'instance unique » et « ressource évolutive ».
ressource de nom d'hôte logique	Ressource contenant un ensemble de noms d'hôtes logiques représentant des adresses réseau. Elles ne peuvent être contrôlées que par un seul noeud à la fois. Voir aussi « hôte logique ».
ressource d'instance unique	Ressource pour laquelle une ressource au plus peut être active dans le cluster.
ressource évolutive	Ressource fonctionnant sur plusieurs noeuds (une instance sur chaque noeud) et utilisant l'interconnexion du cluster pour donner aux clients distants l'apparence d'un service unique.
ressource générique	Démon d'application et son processus fils mis sous le contrôle du gestionnaire du groupe de ressources comme type de ressource générique.
ressource globale	Ressource à haute disponibilité fournie au niveau du noyau du logiciel Sun Cluster. Les ressources globales peuvent inclure les disques (groupes de périphériques HD), le système de fichiers de cluster et la gestion de réseaux globale.
ressource réseau	Ressource contenant un ou plusieurs noms d'hôtes logiques ou adresses partagées. Vois aussi « ressource de nom d'hôte logique » et « ressource d'adresse partagée ».
rétablissement	Voir « rétablissement automatique ».
rétablissement automatique	Processus par lequel un groupe de ressources ou de périphériques est renvoyé vers son noeud principal après une panne de celui-ci et a été redémarré comme membre du cluster.
retour	Voir « basculement ».

S

service de données Application équipée pour fonctionner comme ressource hautement disponible sous le contrôle du gestionnaire du groupe de ressources (RGM).

service de données HD Voir « service de données ».

service évolutif Service de données implémenté fonctionnant simultanément sur plusieurs noeuds.

Solaris Volume Manager Gestionnaire de volumes utilisé par le système SunPlex. Voir aussi « gestionnaire de volumes ».

split brain Situation dans laquelle un cluster se divise en plusieurs partitions ; chaque partition se formant sans connaître l'existence des autres.

SSP (system service processor) Dans les configurations Enterprise 10000, périphérique externe au cluster, spécialement utilisé pour communiquer avec les membres du cluster.

statut de la ressource Situation des ressources rapportée par le détecteur de pannes.

SunPlex Système matériel et logiciel Sun Cluster intégré permettant de créer des services évolutifs et à haute disponibilité.

système de fichiers de cluster Service du cluster donnant un accès hautement disponible aux systèmes de fichiers locaux existants du cluster.

T

type de ressource Nom unique donné à l'objet d'un service de données, LogicalHostname ou SharedAddress du cluster. Les types de ressources des services de données peuvent être évolutifs ou de basculement. Voir aussi « service de données », « ressource de basculement » et « ressource évolutive ».

type de ressource générique Modèle pour un service de données. Un type de ressource générique peut être utilisé pour faire d'une simple application un service de données de basculement (arrêt sur un noeud, démarrage sur un autre). Ce type ne nécessite pas de programmation par l'API SunPlex.

type de ressource parallèle Type de ressource, tel qu'une base de données parallèle, ayant été équipé pour fonctionner dans un environnement de cluster afin de pouvoir être contrôlé par plusieurs (deux ou davantage) noeuds simultanément.

V

VERITAS Volume Manager Gestionnaire de volumes utilisé par le système SunPlex. Voir aussi « gestionnaire de volumes ».

Index

D

E

F

N

O

P

Q

R

S

T

U

V