Présentation de Sun Cluster pour SE Solaris

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

Référence : 819–0156–10
Septembre 2004, Révision A

Copyright 2004 Sun Microsystems, Inc. 4150 Network Circle, Santa Clara, CA 95054 U.S.A. Tous droits réservés.

041129@10536

Table des matières

Préface

Le document *Présentation de Sun™ Cluster pour SE Solaris* introduit le produit Sun Cluster en expliquant son intérêt et la façon dont il s'utilise. Il explique également les notions fondamentales de Sun Cluster. Les informations fournies permettent de se familiariser avec les fonctions et caractéristiques de Sun Cluster.

Documentation connexe

Vous trouverez dans le tableau suivant les manuels contenant des informations sur des sujets connexes associés à Sun Cluster. Toute la documentation relative à Sun Cluster est disponible à l'adresse http://docs.sun.com.

Sujet	Documentation
Présentation	Présentation de Sun Cluster pour SE Solaris
Concepts	*Guide des notions fondamentales de Sun Cluster pour SE Solaris*
Installation et administration du matériel	*Sun Cluster 3.x Hardware Administration Manual for Solaris OS*
	Guides d'administration matérielle individuelle
Installation du logiciel	*Guide d'installation du logiciel Sun Cluster pour SE Solaris*
Installation et administration des services de données	*Sun Cluster Data Services Planning and Administration Guide for Solaris OS*
	Guides des services de données individuels
Développement de services de données	*Guide des développeurs pour les services de données Sun Cluster pour SE Solaris*

Sujet	Documentation
Administration du système	*Guide d'administration système de Sun Cluster pour SE Solaris*
Messages d'erreur	*Sun Cluster Error Messages Guide for Solaris OS*
Références sur les commandes et les fonctions	*Sun Cluster Reference Manual for Solaris OS*

Pour obtenir la liste complète de la documentation Sun Cluster, reportez-vous aux notes de version de votre logiciel Sun Cluster, disponibles à l'adresse `http://docs.sun.com`.

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 est `http://docs.sun.com`.

Commande de documents Sun

Sun Microsystems offre une sélection de documentation produit imprimée. Pour obtenir la liste de ces documents et les commander, reportez-vous à la rubrique "Acheter la documentation imprimée" à l'adresse `http://docs.sun.com`.

Accès à l'aide

Si vous rencontrez des difficultés lors de l'installation ou de l'utilisation du système Sun Cluster, contactez votre fournisseur de services et donnez-lui les informations suivantes :

- votre nom et votre adresse de courrier électronique (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, 3.1 09/04).

Pour obtenir des informations sur chaque nœud 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`	Indique les patchs installés.
`prtdiag -v`	Affiche des informations diagnostiques sur le système.
`scinstall -pv`	Affiche des informations sur la version du package et 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 des informations sur les ressources installées, les groupes de ressources et les types de ressources.

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

Conventions typographiques

Vous trouverez ci-dessous les styles typographiques de cette documentation.

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.`
`AaBbCc123`	Ce que vous entrez, par opposition à ce qui s'affiche à l'écran.	`nom_machine%` **`su`** `Mot de passe :`

TABLEAU P–1 Conventions typographiques *(Suite)*

Type de caractère ou symbole	Signification	Exemple
AaBbCc123	Paramètre substituable de ligne de commande à remplacer par un nom ou une valeur	La commande permettant de supprimer un fichier est `rm>` *nom_fichier*.
AaBbCc123	Titres de manuels, termes nouveaux et mis en évidence.	Reportez-vous au chapitre 6 du *Guide de l'utilisateur*. Effectuez une *analyse de patchs*. *N'*enregistrez *pas* le fichier. [Notez que certains éléments mis en évidence s'affichent en gras sur le site.]

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 Shell

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

CHAPITRE 1

Introduction à Sun Cluster

Le système SunPlex est une solution matérielle et logicielle Sun Cluster intégrée destinée à la création de services à haute disponibilité évolutifs. Ce chapitre donne une vue d'ensemble de haut niveau des fonctionnalités de Sun Cluster.

Ce chapitre comprend les rubriques suivantes :

Haute disponibilité des applications avec Sun Cluster

Un cluster est composé de deux ou plusieurs systèmes (ou nœuds) fonctionnant ensemble pour former un système unique continuellement disponible permettant de fournir des applications, des ressources système et des données aux utilisateurs. Chaque nœud d'un cluster est un système fonctionnel autonome. Cependant, dans un environnement clustérisé, les nœuds sont reliés par une interconnexion et fonctionnent ensemble comme une entité unique pour fournir une disponibilité et des performances accrues.

Les clusters hautement disponibles fournissent un accès quasi-continu aux données et aux applications en préservant l'exécution du cluster lors d'échecs qui mettraient hors fonction un système à serveur unique. Aucune panne isolée (au niveau du matériel, du logiciel ou du réseau) ne peut causer la défaillance d'un cluster. Les systèmes à tolérance de pannes offrent quant à eux un accès permanent aux données et applications, mais leur prix est plus élevé du fait de l'utilisation de matériel spécialisé. Ils sont généralement protégés contre les pannes logicielles.

Chaque système Sun Cluster est un ensemble de nœuds étroitement liés fournissant une vue d'administration unique des services réseau et applications. Le système Sun Cluster permet une haute disponibilité grâce à une combinaison des matériels et logiciels suivants :

- Les systèmes de disques redondants permettent le stockage. Ils sont généralement mis en miroir pour permettre un fonctionnement ininterrompu en cas de défaillance d'un disque ou d'un sous-système. Les connexions redondantes aux systèmes de disques garantissent que les données ne sont pas isolées en cas de panne d'un serveur, d'un contrôleur ou d'un câble. Une interconnexion haute vitesse parmi les nœuds fournit l'accès aux ressources. Tous les nœuds du cluster sont également connectés à un réseau public, permettant aux clients de plusieurs réseaux d'accéder au cluster.
- Les composants redondants enfichables à chaud, comme les prises d'alimentation et les systèmes de refroidissement, optimisent la disponibilité en permettant aux systèmes de continuer à fonctionner après une panne matérielle. Les composants enfichables à chaud offrent la possibilité d'ajouter ou de supprimer les composants matériels d'un système en fonctionnement sans l'arrêter.
- La structure à haute disponibilité du logiciel Sun Cluster détecte rapidement un nœud défaillant et fait migrer l'application ou le service vers un autre nœud tournant dans un environnement identique. À aucun moment les applications ne se retrouvent toutes indisponibles. Les applications non affectées par un nœud défaillant sont totalement disponibles lors de la récupération. En outre, celles du nœud défaillant redeviennent disponibles dès que la récupération est terminée. Après la récupération, l'application n'a pas à attendre que toutes les autres aient terminé la leur.

Gestion de la disponibilité

Une application est dite hautement disponible si elle est capable de survivre à toute panne matérielle ou logicielle unique du système. Les pannes dues aux bogues et à la corruption de données dans l'application elle-même sont exclues. Pour les applications hautement disponibles :

- La récupération est transparente à partir des applications utilisant une ressource.
- L'accès aux ressources est totalement préservé lors d'une panne de nœud.
- Les applications ne peuvent pas détecter le déplacement du nœud hôte vers un autre nœud.
- La panne d'un nœud unique est complètement transparente pour les programmes des nœuds restants utilisant les fichiers, périphériques et volumes de disques attachés à ce nœud.

Services de basculement, services évolutifs et applications parallèles

Les services de basculement, les services évolutifs et les applications parallèles vous permettent de rendre vos applications hautement disponibles et d'améliorer les performances d'une application sur un cluster.

Un service de basculement offre une haute disponibilité à travers la redondance. Lors d'une panne, vous pouvez configurer une application s'exécutant pour qu'elle redémarre sur le même nœud ou qu'elle soit déplacée sur un autre nœud du cluster, sans intervention de l'utilisateur.

Pour améliorer les performances, un service évolutif oriente les différents nœuds d'un cluster pour qu'ils exécutent une application en même temps. Dans une configuration évolutive, chaque nœud du cluster peut fournir des données et traiter les requêtes du client.

Les bases de données parallèles permettent aux différentes instances du serveur de base de données :

- de participer au cluster ;
- de traiter différentes requêtes dans la même base de données en même temps ;
- de rendre possibles les requêtes parallèles lors de requêtes importantes.

Pour plus d'informations sur les services de basculement, les services évolutifs et les applications parallèles, reportez-vous à la section "Types de services de données " à la page 27.

Multiacheminement sur réseau IP

Les clients adressent des requêtes de données au cluster à travers le réseau public. Chaque nœud du cluster est connecté à au moins un réseau public à travers un ou plusieurs adaptateurs de réseau public.

Le Multiacheminement sur réseau IP permet à un serveur d'avoir plusieurs ports réseau connectés au même sous-réseau. Le logiciel Multiacheminement sur réseau IP fournit d'abord une tolérance aux pannes de l'adaptateur réseau en détectant la défaillance ou la réparation d'un adaptateur réseau. Ensuite, il commute simultanément l'adresse réseau vers et depuis l'autre adaptateur. Lorsque plusieurs adaptateurs réseau sont fonctionnels, le Multiacheminement sur réseau IP augmente le débit de données en répartissant des paquets sortants aux adaptateurs.

Gestion du stockage

Le stockage multihôte rend les disques hautement disponibles en les connectant à plusieurs nœuds. Les différents nœuds permettent différents chemins d'accès aux données ; si un chemin est défaillant, un autre est disponible pour prendre sa place.

Les disques multihôtes permettent les processus de cluster suivants :

- Tolérance de pannes pour un nœud unique.
- Centralisation des données d'application, binaires d'applications et fichiers de configuration.
- Protection contre les pannes de nœuds. Si les requêtes client accèdent aux données via un nœud défaillant, elles sont basculées vers un autre nœud ayant une connexion directe aux mêmes disques.
- Accès soit global par un nœud principal« gérant » les disques, soit concurrent direct via les chemins locaux.

Prise en charge de la gestion de volumes

Un gestionnaire de volumes permet de gérer de grands nombres de disques et les données qu'ils contiennent. Les gestionnaires de volumes peuvent augmenter la capacité de stockage et la disponibilité des données en offrant les fonctionnalités suivantes :

- entrelacement et concaténation de l'unité de disque ;
- mise en miroir ;
- disques hot spares ;
- traitement de pannes de disques et remplacements de disques.

Les systèmes Sun Cluster prennent en charge les gestionnaires de volumes suivants :

- Solaris Volume Manager ;
- VERITAS Volume Manager.

Sun StorEdge Traffic Manager

Le logiciel Sun StorEdge Traffic Manager entièrement intégré démarre avec la structure d'E/S centrale du système d'exploitation Solaris 8. Il permet de représenter et de gérer de manière plus efficace des périphériques accessibles à travers plusieurs interfaces de contrôleur d'E/S dans une instance unique de l'environnement d'exploitation Solaris. L'architecture de ce logiciel permet :

- la protection contre les coupures d'E/S dues aux pannes de contrôleurs d'E/S ;
- la commutation automatique vers un autre contrôleur en cas de panne de contrôleur d'E/S ;
- des performances d'E/S accrues grâce à l'équilibrage des charges sur plusieurs canaux d'E/S.

Prise en charge des ensembles redondants de disques indépendants matériels

Les systèmes Sun Cluster prennent en charge l'utilisation de l'ensemble redondant de disques indépendants (RAID) matériel et du RAID logiciel basé sur les hôtes. Le RAID matériel utilise la redondance matérielle du système de stockage ou de la baie de disques de stockage pour garantir que les pannes matérielles indépendantes n'influencent pas la disponibilité des données. Si vous mettez en miroir différents systèmes de stockage, le RAID logiciel basé sur les hôtes garantit que des pannes matérielles indépendantes n'influencent pas la disponibilité des données lorsqu'un système de stockage entier est hors ligne. Même si vous pouvez utiliser en même temps un RAID matériel et un RAID logiciel basé sur les hôtes, vous n'avez besoin que d'une seule solution RAID pour maintenir un niveau élevé de disponibilité des données.

Prise en charge des systèmes de fichiers

Une des propriétés inhérentes aux systèmes clustérisés étant les ressources partagées, un cluster requiert un système de fichiers répondant au besoin de partage cohérent des fichiers. Le système de fichiers Sun Cluster permet aux utilisateurs et aux applications d'accéder à n'importe quel fichier sur n'importe quel nœud du cluster en utilisant les API UNIX standard distantes ou locales. Les systèmes de fichiers Sun Cluster prennent en charge les systèmes de fichiers suivants :

- système de fichiers UNIX (UFS) ;
- système de fichiers Sun StorEdge QFS ;
- système de fichiers VERITAS (VxFS).

Si une application est déplacée d'un nœud à un autre, aucune modification n'est requise pour que l'application accède aux mêmes fichiers. Les applications existantes n'ont pas besoin d'être modifiées pour utiliser pleinement le système de fichiers de cluster.

Clusters de campus

Les systèmes Sun Cluster standard offrent haute disponibilité et fiabilité à partir d'un seul emplacement. Si votre application doit rester disponible après des catastrophes imprévisibles telles qu'un tremblement de terre, des inondations ou une coupure de courant, vous pouvez configurer votre cluster en tant que cluster de campus.

Les clusters de campus vous permettent de placer des composants de cluster, tels que des nœuds et des dispositifs de stockage partagé, dans des locaux différents, distants de plusieurs kilomètres. Vous pouvez séparer les nœuds des dispositifs de stockage partagé en les installant dans différents locaux dispersés au sein du site de votre entreprise ou à l'extérieur, dans un rayon de plusieurs kilomètres. Si l'un des locaux est touché par une catastrophe, les nœuds épargnés peuvent prendre la relève du nœud défaillant. Ainsi, les applications et les données restent disponibles pour vos utilisateurs.

Contrôle de pannes

Le système Sun Cluster rend hautement disponible le chemin entre les utilisateurs et les données en utilisant des disques multihôtes, le multiacheminement et un système de fichiers globaux. Le système Sun Cluster contrôle les pannes des éléments suivants :

- Les applications : la plupart des services de données Sun Cluster comportent un détecteur de pannes sondant périodiquement le service de données pour déterminer son état. Un détecteur de pannes vérifie que le ou les démons d'application s'exécutent et que les clients sont servis. En fonction des informations retournées par les sondes, une action prédéfinie comme le redémarrage des démons ou le déclenchement d'un basculement peut être initié.
- Chemins de disques : le logiciel Sun Cluster prend en charge le contrôle de chemin de disque (CCD). Le CCD améliore la fiabilité générale du basculement et de la commutation en rapportant l'échec du chemin de disque secondaire.
- Multiacheminement sur protocole Internet (IP) : le logiciel Solaris de multiacheminement sur réseau IP des systèmes Sun Cluster fournit le mécanisme de base du contrôle des adaptateurs de réseau public. Le multiacheminement IP permet également le basculement d'adresses IP d'un adaptateur vers un autre lors de la détection d'une panne.

Outils d'administration et de configuration

Vous pouvez installer, configurer et administrer le système Sun Cluster soit via l'interface graphique Gestionnaire SunPlex, soit via l'interface de ligne de commande (CLI).

Le système Sun Cluster a également un module s'exécutant comme élément du logiciel Sun Management Center pour offrir une interface graphique pour certaines tâches de cluster.

Gestionnaire SunPlex

Gestionnaire SunPlex est un outil basé sur le navigateur pour l'administration des systèmes Sun Cluster. Ce logiciel permet aux administrateurs de gérer et de contrôler des systèmes, d'installer des logiciels et de configurer des systèmes.

SunPlex Manager comprend les fonctionnalités suivantes :

- mécanismes de sécurité et d'autorisations intégrés ;
- prise en charge SSL (Secure Sockets Layer) ;
- contrôle d'accès basé sur des rôles (RBAC) ;
- PAM (module d'authentification enfichable) ;
- fonctions d'administration de groupes de multiacheminement sur réseau IP et NAFO ;
- périphériques de quorum, de transport, de stockage partagé et d'administration de groupes de ressources ;
- vérification d'erreur avancée et détection automatique d'interconnexions privées.

interface de ligne de commande

L'interface de ligne de commande Sun Cluster est un ensemble d'utilitaires permettant d'installer et d'administrer les systèmes Sun Cluster, et d'administrer la portion du gestionnaire de volumes du logiciel Sun Cluster.

Vous pouvez effectuer les tâches d'administration SunPlex suivantes via l'interface de ligne de commande de Sun Cluster :

- validation d'une configuration de Sun Cluster ;
- installation et configuration du logiciel Sun Cluster ;
- mise à jour d'une configuration de Sun Cluster ;
- gestion de l'enregistrement des types de ressources, de la création de groupes de ressources et de l'activation des ressources dans un groupe ;
- modification de la maîtrise des nœuds et de l'état des groupes de ressources et des groupes de périphériques de disque ;
- contrôle de l'accès avec des contrôles basés sur des rôles (RBAC) ;
- arrêt de l'ensemble du cluster.

Sun Management Center

Le système Sun Cluster comprend également un module s'exécutant comme une partie du logiciel Sun Management Center. Le logiciel Sun Management Center sert de base au cluster pour les opérations d'administration et de contrôle et permet aux administrateurs système d'effectuer les tâches suivantes au moyen d'une interface graphique ou d'une interface de ligne de commande :

- configuration d'un système distant ;
- contrôle des performances ;
- détection et isolement des pannes matérielles et logicielles.

Le logiciel Sun Management Center peut également être utilisé comme interface de gestion de reconfiguration dynamique au sein des serveurs Sun Cluster. La reconfiguration dynamique comprend la création de domaines, la liaison de cartes dynamiques et la séparation dynamique.

RBAC

Dans les systèmes UNIX classiques, l'utilisateur racine, également appelé superutilisateur, est omnipotent, ayant le droit de lecture et d'écriture sur tous les fichiers, d'exécution de tous les programmes et d'envoi de signaux kill à n'importe quel processus. Le contrôle d'accès basé sur des rôles (RBAC) est une alternative au modèle superutilisateur "tout ou rien". Le RBAC utilise le principe de sécurité du privilège minimum : aucun utilisateur ne doit se voir accorder plus de privilèges que nécessaire pour effectuer son travail.

Le RBAC permet une organisation permettant de séparer les droits du superutilisateur et de les regrouper dans des comptes utilisateurs ou rôles spéciaux pour les affecter à des individus particuliers. Cette séparation et ces groupements permettent d'appliquer une variété de règles de sécurité. Les comptes peuvent être créés pour des administrateurs aux fonctions spéciales dans des domaines comme la sécurité, la mise en réseau, les pare-feux, les sauvegardes et le fonctionnement du système.

CHAPITRE 2

Notions fondamentales de Sun Cluster

Ce chapitre décrit les notions fondamentales liées aux composants matériels et logiciels du système Sun Cluster que vous devez comprendre avant de travailler avec les systèmes Sun Cluster.

Ce chapitre comprend les rubriques suivantes :

Nœuds de cluster

Un nœud de cluster est une machine exécutant à la fois le logiciel Solaris et le logiciel Sun Cluster. Le logiciel Sun Cluster permet d'avoir de deux à huit nœuds dans un cluster.

Les nœuds de cluster sont généralement liés à un ou plusieurs disques. Ceux qui ne le sont pas utilisent le système de fichiers de cluster pour accéder aux disques multihôtes. Les nœuds de configurations de bases de données parallèles partagent l'accès simultané à certains ou à tous les disques.

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

Les nœuds 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 la possibilité de basculement, chaque nœud doit avoir des capacités suffisantes pour satisfaire les critères de niveau de service en cas d'échec d'un nœud.

Interconnexion de cluster

L'interconnexion de cluster est la configuration physique de périphériques utilisés pour transférer des communications de clusters privés et des communications de services de données entre les nœuds du cluster.

Les interconnexions redondantes permettent de poursuivre une opération sur les interconnexions subsistantes alors que les administrateurs système isolent les pannes et réparent une communication. Le logiciel Sun Cluster détecte, répare et réinitie automatiquement la communication sur l'interconnexion réparée.

Pour plus d'informations, reportez-vous à la section "Interconnexion de cluster " à la page 35.

Appartenance au cluster

Le moniteur d'appartenance au cluster (CMM, Cluster Membership Monitor) est un ensemble distribué d'agents échangeant des messages via l'interconnexion de cluster pour effectuer les tâches suivantes :

- appliquer une vue cohérente de l'appartenance sur tous les nœuds (quorum) ;
- assurer la reconfiguration synchronisée suite aux modifications d'appartenance ;
- gérer le partitionnement de cluster ;
- assurer la connectivité complète de tous les membres du cluster en laissant les nœuds défaillants hors du cluster jusqu'à leur réparation.

La fonction principale du CMM est d'établir l'appartenance au cluster, requérant un accord pour tout le cluster de l'ensemble de nœuds participant au cluster à tout moment. Il détecte les principales modifications d'état du cluster sur chaque nœud, comme une perte de communication entre un ou plusieurs nœuds. Il se base sur le module du noyau de transport pour générer les pulsations sur le média de transport vers d'autres nœuds du cluster. Lorsqu'il ne détecte pas de pulsation de la part d'un nœud dans la période de temporisation définie, le CMM considère que le nœud est défaillant et il initie une reconfiguration du cluster pour renégocier l'appartenance au cluster.

Pour déterminer l'appartenance au cluster et pour assurer l'intégrité des données, le CMM effectue les tâches suivantes :

- comptabilisation des modifications d'appartenance au cluster, telles qu'un nœud rejoignant ou quittant le cluster ;
- garantie de la sortie de tout nœud défaillant du cluster ;
- garantie de l'inactivité d'un nœud défaillant jusqu'à sa réparation ;
- protection du cluster contre un partitionnement en sous-ensembles de nœuds.

Pour plus d'informations sur la manière dont le cluster se protège des partitionnements en plusieurs clusters, reportez-vous à la rubrique "Intégrité des données " à la page 21.

Référentiel de configuration de la grappe

Le référentiel de configuration du cluster (CCR, Cluster Configuration Repository) est une base de données privée, distribuée, à l'échelle du cluster pour le stockage d'informations relatives à la configuration et à l'état du cluster. Pour éviter toute corruption des données de configuration, chaque nœud doit connaître l'état actuel des ressources du cluster. Le CCR veille à ce que tous les nœuds aient une vue cohérente du cluster. Il est mis à jour en cas d'erreur ou de récupération, ou lorsque l'état général du cluster est modifié.

Les structures du CCR contiennent les types d'information suivants :

- noms de cluster et de nœud ;
- configuration de transport du cluster ;
- noms des jeux de disques Solaris Volume Manager ou des groupes de disques VERITAS ;
- liste des nœuds pouvant gérer chaque groupe de disques ;
- valeurs de paramètres opérationnels pour les services de données ;
- chemins vers les méthodes de rappel de services de données ;
- configuration des périphériques IDP ;
- état actuel du cluster.

Détecteurs de pannes

Le système Sun Cluster rend tous les composants du « chemin » entre les utilisateurs et les données hautement disponibles en surveillant les applications elles-mêmes, le système de fichiers et les interfaces réseau.

Le logiciel Sun Cluster détecte rapidement une panne de nœud et crée un serveur équivalent pour les ressources sur le nœud défaillant. Grâce au logiciel Sun Cluster, les ressources non affectées par le nœud défaillant restent disponibles lors de la récupération ; les ressources du nœud défaillant sont, une fois réparées, à nouveau disponibles.

Contrôle des services de données

Chaque service de données Sun Cluster intègre un détecteur de pannes le sondant périodiquement pour vérifier son état. Un détecteur de pannes vérifie que le ou les démons d'application s'exécutent 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.

Contrôle de chemin de disque

Le logiciel Sun Cluster prend en charge le contrôle de chemin de disque (CCD). Le CCD améliore la fiabilité générale du basculement et de la commutation en rapportant la panne d'un chemin d'accès aux disques secondaire. Vous pouvez utiliser une des deux méthodes de contrôle de chemin de disque. La première méthode 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. Pour plus d'informations sur les options de ligne de commande, reportez-vous à la page man `scdpm`(1M).

La seconde méthode consiste à utiliser l'interface utilisateur graphique (IUG) de Gestionnaire SunPlex. Gestionnaire SunPlex offre une vue topologique des chemins contrôlés. Cette vue est mise à jour toutes les 10 minutes afin de donner des informations sur le nombre de pings ayant échoué.

Contrôle du multiacheminement sur réseau IP

Chaque nœud du cluster a sa propre configuration de Multiacheminement sur réseau IP, pouvant être différente de celle des autres nœuds. Multiacheminement sur réseau IP contrôle les pannes de communication réseau suivantes :

- Le chemin de transmission et de réception de l'adaptateur réseau a arrêté de transmettre des paquets.
- La liaison de l'adaptateur réseau au lien est désactivée.
- Le port sur le commutateur ne reçoit/transmet pas de paquet.
- L'interface physique d'un groupe n'est pas présente à l'initialisation du système.

Périphériques de quorum

Un périphérique de quorum est un disque, partagé par deux nœuds ou plus, participant aux votes utilisés pour calculer le quorum du cluster utilisé. 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 nœuds pour déterminer quel ensemble de nœuds constitue le nouveau cluster.

Les nœuds du cluster et les périphériques de quorum votent tous pour former un quorum. Par défaut les nœuds du cluster acquièrent un nombre de vote de quorum égal à un lorsqu'ils sont initialisés et deviennent membres du cluster. Les nœuds peuvent aussi avoir un nombre de vote égal à zéro, lorsqu'ils sont en cours d'installation ou ont été placés à l'état de maintenance par un administrateur.

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

Intégrité des données

Le système Sun Cluster tente d'empêcher toute corruption des données et veille à leur intégrité. Comme les nœuds du cluster partagent des données et des ressources, un cluster ne doit jamais se diviser en partitions distinctes actives en même temps. Le CMM veille à ce qu'il n'y ait toujours qu'un seul cluster opérationnel à tout moment.

Des partitions de cluster peuvent provoquer deux types de problèmes : le split brain et l'amnésie. Le split brain a lieu lorsque l'interconnexion entre les nœuds du cluster est perdue et que le cluster est partitionné en sous-clusters, chaque sous-cluster croyant

être la seule partition. Un sous-cluster ignorant l'existence d'autres sous-clusters peut entraîner un conflit au niveau des ressources partagées tel que la duplication des adresses réseau et la corruption de données.

L'amnésie apparaît si tous les nœuds quittent le cluster en groupes successifs. Prenons l'exemple d'un cluster à deux nœuds avec les nœuds A et B. Si le nœud A tombe en panne, les données de configuration du CCR ne sont mises à jour que sur le nœud B, et pas sur le nœud A. Si le nœud B tombe en panne par la suite, et que le nœud A est réinitialisé, ce dernier s'exécutera avec l'ancien contenu du CCR. Cet état est appelé amnésie et peut conduire à exécuter un cluster avec des informations de configuration obsolètes.

Le split brain et l'amnésie peuvent être évités en donnant un vote à chaque nœud 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 majorité de votes fonctionne bien si le cluster compte plus de deux nœuds. Dans un cluster à deux nœuds, la majorité est deux. Si ce cluster est partitionné, un vote externe permet à une partition d'obtenir le 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 nœuds.

Le Tableau 2–1 décrit la manière dont le logiciel Sun Cluster utilise le quorum pour éviter les problèmes de split brain et d'amnésie.

TABLEAU 2–1 Quorum du cluster et problèmes de split brain et d'amnésie

Type de partition	Solution du quorum
Split brain	N'autorise que les partitions (sous-cluster) ayant une majorité de votes à s'exécuter comme le cluster (avec une telle majorité, il ne peut exister qu'une seule partition). Lorsqu'un nœud a perdu la course au quorum, il panique.
Amnésie	Garantit, lors de l'initialisation, que le cluster possède au moins un nœud faisant partie des derniers membres du cluster (possédant donc les données de configuration les plus à jour).

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, les nœuds ne peuvent pas tous communiquer, ainsi, des nœuds individuels ou des sous-ensembles de nœuds risquent de tenter de former des clusters individuels ou des sous-ensembles. Chaque partition ou sous-ensemble peut « croire » qu'il est le seul à être propriétaire des disques multihôtes et à en posséder l'accès. Les tentatives d'écriture des différents nœuds sur les disques peuvent entraîner une corruption des données.

La séparation en cas d'échec limite l'accès des nœuds aux disques multihôtes en les empêchant d'accéder aux disques. Lorsqu'un nœud 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 nœuds ont accès aux disques, garantissant ainsi l'intégrité des données.

Le système Sun Cluster utilise le mode de réservation des disques SCSI pour implémenter la séparation en cas d'échec. Grâce au système de réservation SCSI, les nœuds défectueux sont « isolés » à l'extérieur des disques multihôtes, pour les empêcher d'accéder à ces disques.

Lorsqu'un membre détecte qu'un autre nœud ne communique plus sur l'interconnexion du cluster, il lance une procédure de séparation en cas d'échec pour empêcher le nœud défaillant d'accéder aux disques partagés. Lors de la séparation en cas d'échec, le nœud séparé panique et un message de « conflit de réservation » s'affiche sur la console.

Mécanisme failfast pour la séparation en cas d'échec

Le mécanisme failfast fait paniquer un nœud défaillant, mais ne l'empêche pas de redémarrer. Après la panique, le nœud peut redémarrer et tenter de rejoindre le cluster.

Si un nœud perd la connectivité aux autres nœuds 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 nœud. Un autre nœud faisant partie de la partition pouvant atteindre le quorum effectue des réservations sur les disques partagés. Le nœud n'ayant pas de quorum panique alors en conséquence du mécanisme failfast.

Périphériques

Le système de fichiers global donne à tous les fichiers d'un cluster un même niveau d'accessibilité et de visibilité pour tous les nœuds. De même, le logiciel Sun Cluster rend tous les périphériques d'un cluster accessibles et visibles dans tout le cluster. C'est-à-dire que le sous-système d'E/S rend possible l'accès à tout périphérique du cluster, à partir de n'importe quel nœud, quel que soit le lieu physique de connexion du périphérique. Cet accès est appelé accès périphérique global.

Périphériques globaux

Les systèmes Sun Cluster utilisent des périphériques globaux pour fournir un accès hautement disponible dans tout le cluster à tout périphérique d'un cluster, à partir de n'importe quel nœud. En général, si un nœud est défaillant lorsqu'il permet l'accès à un périphérique global, le logiciel Sun Cluster change le chemin vers ce périphérique et redirige l'accès en utilisant ce nouveau chemin. Cette redirection est facile avec les périphériques globaux car le même nom de périphérique est utilisé, quel que soit le chemin. L'accès aux périphériques distants s'effectue de la même manière que pour des périphériques locaux utilisant le même nom. Ainsi, l'API d'accès à un périphérique global d'un cluster est le même que celui utilisé pour accéder à un périphérique localement.

Les périphériques globaux de Sun Cluster comprennent les disques, les CD et les bandes. Cependant, les disques sont les seuls périphériques globaux à ports multiples pris en charge. Cette prise en charge limitée 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 nœud du cluster.

ID de périphérique

Le logiciel Sun Cluster gère les périphériques globaux à travers une structure appelée 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 pilote d'ID de périphériques (IDP) fait partie intégrante de la fonction d'accès aux périphériques globaux du cluster. Il sonde tous les nœuds du cluster et dresse la liste des périphériques de disques uniques. Le pilote d'IDP affecte également un numéro majeur et mineur unique à chaque périphérique cohérent sur tous les nœuds du cluster. L'accès aux périphériques globaux se fait à travers l'IDP unique affecté par le pilote d'IDP au lieu des IDP Solaris traditionnels.

Cela garantit que toute application accédant à des disques, comme Solaris Volume Manager ou Sun Java System Directory Server, utilisera un chemin cohérent dans le cluster. Cette cohérence est particulièrement importante pour les disques multihôtes, car les numéros majeurs et mineurs de chaque périphérique peuvent varier d'un nœud à l'autre. Ces numéros peuvent également modifier les conventions d'attribution de noms de périphériques Solaris.

Périphériques locaux

Le logiciel Sun Cluster gère également les périphériques locaux. Ces périphériques ne sont accessibles que sur un nœud exécutant un service et ayant une connexion physique vers ce cluster. N'ayant pas à répliquer les informations d'état sur plusieurs nœuds en même temps, les périphériques locaux peuvent être plus avantageux en termes de performances par rapport aux périphériques globaux. La défaillance du domaine d'un périphérique supprime l'accès à ce périphérique, à moins qu'il ne soit partagé par plusieurs nœuds.

Groupes de périphériques de disques

Les groupes de périphériques de disques permettent aux groupes de disques gestionnaires de volumes de devenir « globaux » car ils fournissent des prises en charge multihôtes et de multiacheminement aux disques sous-jacents. Chaque nœud du cluster physiquement relié aux disques multihôtes fournit un chemin d'accès au groupe de périphériques de disques.

Dans le système Sun Cluster, les disques multihôtes peuvent être contrôlés par le logiciel Sun Cluster en étant enregistrés comme groupes de périphériques de disques. Cet enregistrement fournit au système Sun Cluster des informations permettant d'identifier quels nœuds possèdent un chemin d'accès à quels groupes de disques gestionnaires de volumes. Le logiciel Sun Cluster crée dans le cluster un groupe de périphériques de disques bruts pour chaque disque et chaque lecteur de bande. Ces groupes de périphériques du cluster restent à l'état hors ligne jusqu'à ce que vous y accédiez en tant que périphériques globaux, soit en montant un système de fichiers global, soit en accédant à un fichier de base de données brut.

Services de données

Un service de données est la combinaison de fichiers logiciels et de configuration permettant à une application de s'exécuter sans modification dans une configuration de Sun Cluster. Lors de l'exécution d'une configuration de Sun Cluster, une application s'exécute comme une ressource contrôlée par le gestionnaire de groupes de ressources (RGM, Resource Group Manager). Un service de données permet de configurer une application telle que Sun Java System Web Server ou la base de données Oracle pour qu'elle s'exécute sur le cluster et non sur un serveur unique.

Le logiciel d'un service de données fournit des implémentations de méthodes de gestion de Sun Cluster effectuant les opérations suivantes sur l'application :

- démarrage de l'application ;
- arrêt de l'application ;
- contrôle des pannes dans l'application et récupération de ces pannes.

Les fichiers de configuration d'un service de données définissent les propriétés de la ressource représentant l'application au RGM.

Le RGM contrôle la disposition des services de données évolutifs et de basculement du cluster. Il démarre et arrête les services de données sur les nœuds sélectionnés du cluster en réponse aux modifications d'appartenance au cluster. Il permet aux applications de services de données d'utiliser la structure de cluster.

Le RGM contrôle les services de données en tant que ressources. Ces implémentations sont soit fournies par Sun, soit créées par un développeur utilisant un modèle de service de données générique, l'API de bibliothèque de développement de services de données (API DSDL, Data Service Development Library API) ou l'API de gestion de ressources (RMAPI, Resource Management API). L'administrateur du cluster crée et gère les ressources dans des conteneurs appelés groupes de ressources. Les actions du RGM et de l'administrateur peuvent faire passer les ressources et les groupes de ressources de l'état en ligne à l'état hors ligne et inversement.

Types de ressources

Un type de ressource est un ensemble de propriétés décrivant une application à un cluster. Cet ensemble comprend des informations sur la manière dont l'application est démarrée, arrêtée et contrôlée sur les nœuds du cluster. Un type de ressource comprend également des propriétés spécifiques à l'application devant être définies pour pouvoir utiliser l'application dans le cluster. Les services de données de Sun Cluster ont plusieurs types de ressources prédéfinis. Par exemple, Sun Cluster HA pour Oracle est le type de ressource `SUNW.oracle-server` et Sun Cluster HA pour Apache le type de ressource `SUNW.apache`.

Ressources

Une ressource est une instance d'un type de ressource défini au niveau du cluster. Le type de ressource permet d'installer plusieurs instances d'une application sur le cluster. Lorsque vous initialisez une ressource, le RGM affecte des valeurs à des propriétés spécifiques à l'application, et la ressource hérite de ces propriétés au niveau du type de ressources.

Les services de données utilisent plusieurs types de ressources. Les applications, telles qu'Apache Web Server ou Sun Java System Web Server utilisent des adresses 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.

Groupes de ressources

Les ressources gérées par le RGM sont placées dans des groupes de ressources de manière à être gérées comme une unité. Un groupe de ressources est un ensemble de ressources connexes ou interdépendantes, Par exemple, une ressource dérivée d'un type de ressource `SUNW.LogicalHostname` peut être placée dans le même groupe de ressources qu'une ressource dérivée d'un type de ressource de base de données Oracle. Si une commutation ou un basculement est initié sur le groupe de ressources, ce dernier se transforme en unité.

Types de services de données

Les services de données permettent aux applications de devenir hautement disponibles et les services évolutifs aident à éviter une interruption majeure de l'application après une défaillance unique au sein du cluster.

Lorsque vous configurez un service de données, vous devez le configurer comme un des types de services de données suivants :

- service de données de basculement ;
- service de données évolutif ;
- service de données parallèle.

Services de données de basculement

Le basculement est le processus par lequel le cluster déplace automatiquement une application d'un nœud principal défaillant vers un nœud secondaire redondant. Les applications de basculement ont les caractéristiques suivantes :

- capables de s'exécuter sur un seul nœud du cluster ;
- non compatibles avec les clusters ;
- dépendantes de la structure du cluster pour la haute disponibilité.

Si le détecteur de pannes rencontre une erreur, il essaie soit de redémarrer l'instance sur le même nœud, soit de la démarrer sur un autre nœud (basculement), selon la configuration du service de données. Les services de basculement utilisent un groupe de ressources de basculement contenant des ressources d'instances d'application et des ressources réseau (noms d'hôtes logiques). Les noms d'hôtes logiques sont des adresses IP pouvant être configurées sur un nœud puis automatiquement retirées pour être configurées sur un autre nœud.

Les clients risquent de subir une brève interruption de service et de devoir se reconnecter après un basculement. Toutefois, ils ignorent la modification du serveur physique fournissant le service.

Services de données évolutifs

Les services de données évolutifs permettent aux instances d'application de s'exécuter sur plusieurs nœuds en même temps. Ils utilisent deux groupes de ressources. Le groupe de ressources évolutif contient les ressources d'application et le groupe de ressources de basculement contient les ressources réseau (adresses partagées) dont dépend le service évolutif. Le groupe de ressources évolutif peut être connecté à plusieurs nœuds, 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 nœud à la fois. Tous les nœuds hébergeant un service évolutif utilisent la même adresse partagée pour héberger le service.

Le cluster reçoit des requêtes de service à travers une interface réseau unique (interface globale). Ces requêtes sont distribuées aux nœuds, en fonction d'un des algorithmes prédéfinis définis par la règle d'équilibrage de la charge. Le cluster peut utiliser cette règle pour équilibrer la charge de service entre plusieurs nœuds.

Applications parallèles

Les systèmes Sun Cluster fournissent un environnement partageant l'exécution en parallèle d'applications sur tous les nœuds du cluster à l'aide de bases de données parallèles. Prise en charge Sun Cluster pour Oracle Parallel Server/Real Application Clusters est un ensemble de packages qui, une fois installé, permet aux Oracle Parallel Server/Real Application Clusters de s'exécuter sur les nœuds de Sun Cluster. Ce service de données permet également à Prise en charge Sun Cluster pour Oracle Parallel Server/Real Application Clusters d'être géré à l'aide des commandes de Sun Cluster.

Une application parallèle a été instrumentée pour s'exécuter dans un environnement de cluster de sorte que l'application puisse être gérée par deux nœuds ou plus en même temps. Dans un environnement de Oracle Parallel Server/Real Application Clusters, plusieurs instances d'Oracle coopèrent pour fournir l'accès à la même base de données partagée. Les clients d'Oracle peuvent utiliser n'importe quelle instance pour accéder à la base de données. Ainsi, si une instance ou plus sont défaillantes, les clients peuvent se connecter à une instance survivante et continuer à accéder à la base de données.

CHAPITRE 3

Architecture Sun Cluster

L'architecture Sun Cluster permet de déployer, de gérer et d'afficher un groupe de systèmes comme un grand système unique.

Ce chapitre comprend les rubriques suivantes :

Environnement matériel de Sun Cluster

Un cluster est formé des composants matériels suivants :

- Les nœuds de cluster avec disques locaux (non partagés) fournissent la plate-forme principale du cluster.
- Le stockage multihôte fournit des disques partagés entre des nœuds.
- Les médias amovibles sont configurés comme des périphériques globaux, comme des bandes ou des CD-ROM.
- L'interconnexion de cluster offre un canal de communication entre les nœuds.
- Les interfaces de réseau public activent les interfaces réseau utilisées par les systèmes client pour accéder aux services de données sur le cluster.

La Figure 3–1 représente les relations entre les composants matériels.

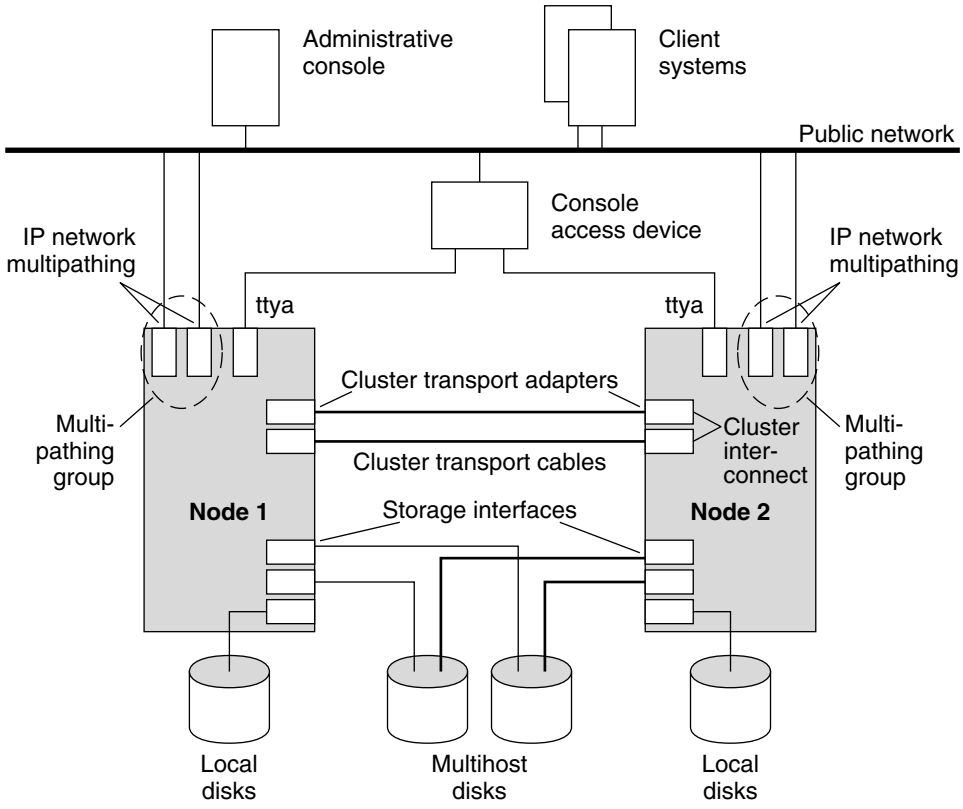

FIGURE 3–1 Composants matériels de Sun Cluster

Environnement logiciel de Sun Cluster

Pour fonctionner comme un membre du cluster, un nœud doit être équipé des logiciels suivants :

- logiciel Solaris ;
- logiciel Sun Cluster ;
- application de service de données ;
- gestion du volume (Solaris™ Volume Manager ou VERITAS Volume Manager).

  La configuration utilisant la gestion du volume constitue une exception. Cette configuration peut se passer d'un logiciel de gestion du volume.

La Figure 3–2 fournit une représentation fonctionnelle des composants logiciels constituant l'environnement logiciel de Sun Cluster.

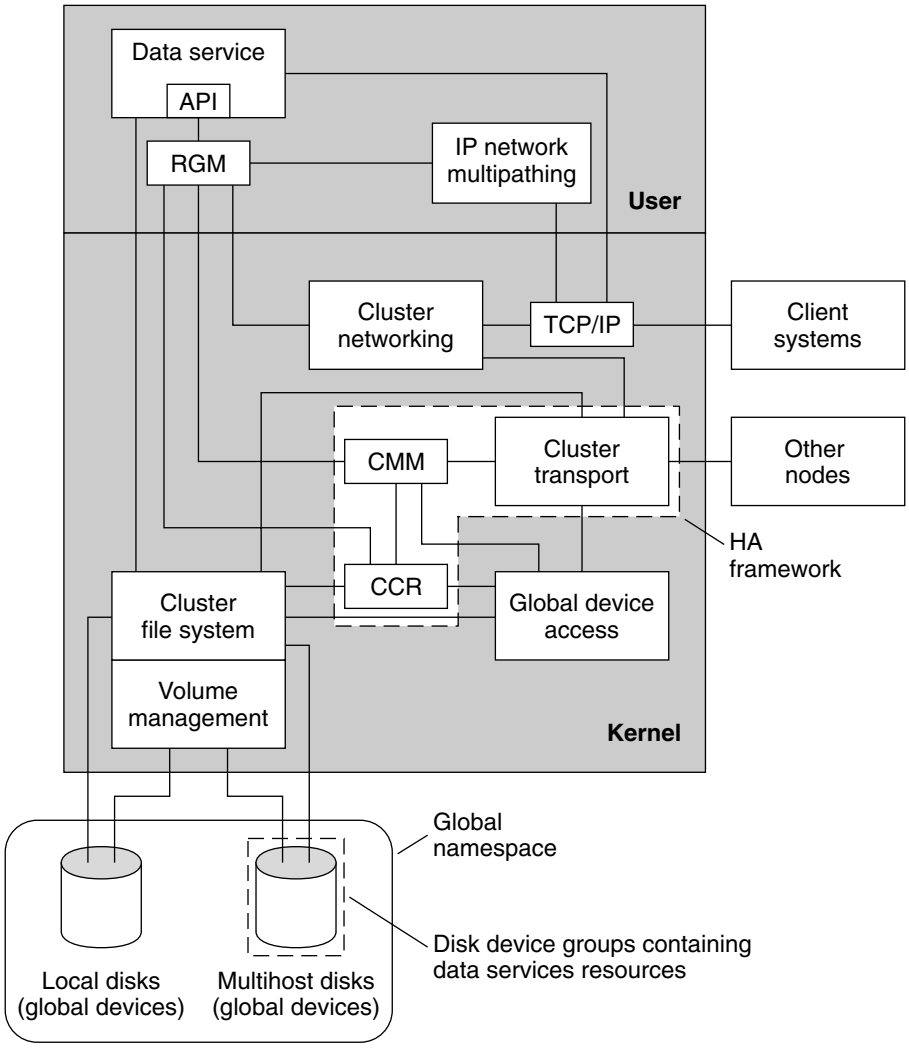

FIGURE 3–2 Architecture logicielle de Sun Cluster

Moniteur d'appartenance à la grappe

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

Le MAC reçoit des informations sur la connectivité aux autres nœuds 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 au cluster, le CMM effectue une configuration synchronisée du cluster. Dans cette configuration, les ressources du cluster peuvent être redistribuées, en fonction de la nouvelle appartenance au cluster.

Le CMM s'exécute entièrement dans le noyau.

Référentiel de configuration de la grappe (CCR)

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.

Systèmes de fichiers de grappe

Un système de fichiers de cluster est un proxy entre les éléments suivants :

- le noyau sur un nœud et le système de fichiers sous-jacent ;
- le gestionnaire de volumes s'exécutant sur un nœud ayant une connexion physique vers le ou les disques.

Les systèmes de fichiers de cluster dépendent des périphériques globaux (disques, bandes, CD-ROM). Les périphériques globaux sont accessibles à partir de n'importe quel nœud du cluster à travers le même nom de fichier (par exemple, `/dev/global/`). Ce nœud n'a pas besoin de connexion physique au périphérique de stockage. Vous pouvez utiliser un périphérique global comme un périphérique normal , c'est-à-dire que vous pouvez créer un système de fichiers sur un périphérique global à l'aide de la commande `newfs` ou `mkfs`.

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 sur le système. De même, les processus sur tous les nœuds peuvent utiliser le même nom de chemin 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 nœuds.

- 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 informatif hautement disponible par le biais des interfaces `fcntl(2)`. Les applications exécutées sur plusieurs nœuds peuvent synchroniser l'accès aux données en utilisant le verrouillage de fichiers informatif sur le fichier d'un système de fichiers du cluster. Les verrous de fichiers sont immédiatement récupérés à partir de nœuds 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 de 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.

Services de données évolutifs

L'objectif principal d'un réseau en cluster est de fournir des services de données évolutifs. 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 nœuds au cluster et à l'exécution de nouvelles instances de serveur. 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 nœuds différents du cluster. Ensemble, ces instances se comportent comme un service unique face à un client distant de ce service et implémentent la fonctionnalité de ce service. Dans un service Web évolutif avec plusieurs démons `httpd` s'exécutant sur différents nœuds, n'importe quel démon peut servir la requête d'un 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, ce qui préserve l'apparence d'un service unique.

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

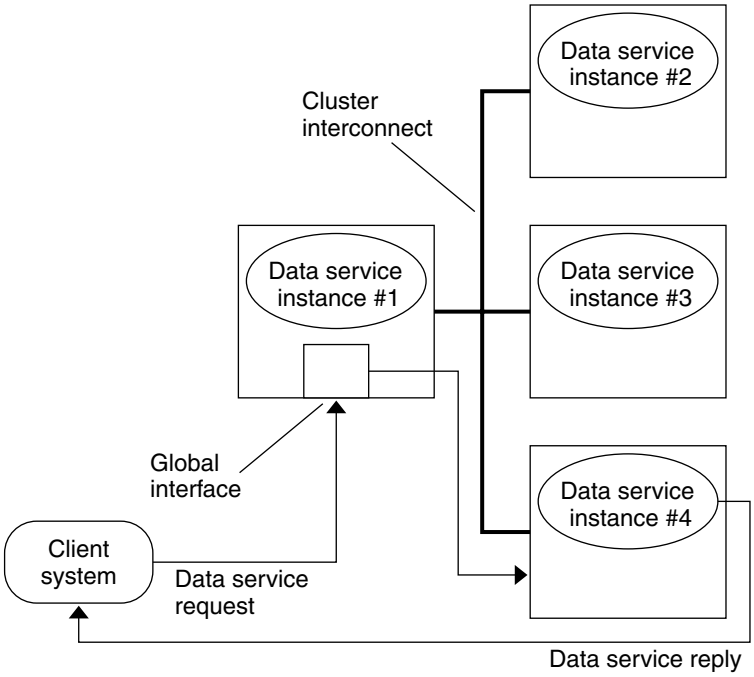

FIGURE 3–3 Architecture d'un service de données évolutif

Les nœuds n'hébergeant pas l'interface globale (nœuds proxy) hébergent l'adresse partagée sur leur interface de bouclage. Les paquets arrivant dans l'interface globale sont distribués à d'autres nœuds de cluster, en fonction de 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 sortes 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 cluster équilibre la charge des requêtes au nœud. 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. Par exemple, dans un cluster à trois nœuds où chaque nœud a le poids 1, chaque nœud traite un tiers des requêtes de n'importe quel client pour le service. Les poids peuvent être modifiés à tout moment grâce à l'interface de commande `scrgadm(1M)` ou à l'interface graphique Gestionnaire SunPlex.

Il y a deux types de services sticky : *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" envers l'instance du serveur écoutant sur un port unique. 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.

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 nœud. Le client est dit "sticky joker" sur les ports et associé à la même adresse IP.

Stockage sur disques multihôtes

Le logiciel Sun Cluster rend les disques hautement disponibles en utilisant un stockage sur disques multihôtes, pouvant être connecté à plus d'un nœud à la fois. Le logiciel de gestion du volume peut être utilisé pour organiser ces disques dans un emplacement de stockage partagé géré par un nœud du cluster. Les disques sont ensuite configurés pour passer à un autre nœud en cas de panne. L'utilisation de disques multihôtes dans les systèmes Sun Cluster offre de nombreux avantages, notamment :

- l'accès global aux systèmes de fichiers ;
- les chemins d'accès multiples aux systèmes de fichiers et aux données ;
- la tolérance aux pannes de nœuds uniques.

Interconnexion de cluster

Tous les nœuds 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. Alors que deux interconnexions sont requises pour la redondance, l'expansion du traffic peut en utiliser jusqu'à six afin d'éviter les goulots d'étranglement et améliorer la redondance et l'évolutivité. L'interconnexion de Sun Cluster utilise Fast Ethernet, Gigabit-Ethernet, Sun Fire Link ou l'interface SCI (SCI, IEEE 1596-1992), pour permettre des communications privées de cluster de haute performance.

Dans les environnements clustérisés, les interconnexions et protocoles à grande vitesse et faible latence sont essentiels pour les communications entre les nœuds. L'interconnexion SCI dans les systèmes Sun Cluster améliore les performances sur les cartes d'interface réseau typiques (NIC). Sun Cluster utilise l'interface Remote Shared Memory (RSM™) pour la communication internodale sur un réseau Sun Fire Link. La RSM est une interface de messagerie Sun, extrêmement efficace pour les opérations de mémoire à distance.

Le pilote RSM Reliable Datagram Transport (RSMRDT) se compose d'un pilote créé sur l'API RSM et d'une bibliothèque qui exporte l'interface API RSMRDT. Ce pilote fournit des performances améliorées Oracle Parallel Server/Real Application Clusters. Les fonctions d'équilibrage de charge et de haute disponibilité sont également améliorées car étant fournies directement avec le pilote, elles sont disponibles de manière transparente pour les clients.

L'interconnexion de cluster se compose des éléments matériels suivants :

- *Adaptateurs* : cartes d'interface réseau résidant sur chaque nœud de cluster. Un adaptateur réseau à plusieurs interfaces peut entraîner un point de panne unique s'il tombe en panne.
- *Jonctions* : commutateurs résidant à l'extérieur des nœuds de cluster. Ils remplissent des fonctions d'intercommunication et de commutation vous permettant de connecter plus de deux nœuds entre eux. Dans un cluster à deux nœuds, vous n'avez pas besoin de jonction, car les nœuds peuvent être directement connectés entre eux à l'aide de câbles physiques redondants. Ces câbles redondants sont connectés à des adaptateurs redondants sur chaque nœud. Les configurations à plus de deux nœuds requièrent des jonctions.
- *Câbles* : connexions placées entre deux adaptateurs réseau ou entre un adaptateur et une jonction.

La Figure 3–4 représente l'interconnexion des trois composants.

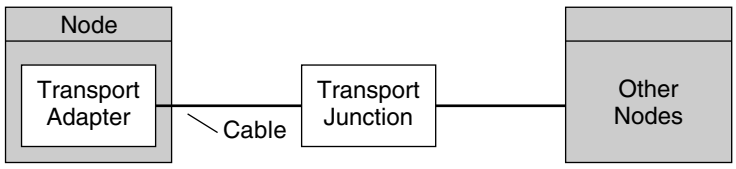

FIGURE 3–4 Interconnexion de cluster

Groupes de multiacheminement sur réseau IP

Les adaptateurs de réseau public sont organisés en groupes IPMP (groupes de multiacheminement). Chaque groupe de multiacheminement possède un ou plusieurs adaptateurs de réseau public. Chaque adaptateur d'un groupe de multiacheminement peut être actif. Vous pouvez aussi configurer des interfaces de réserve demeurant inactives jusqu'au moment d'un basculement.

Les groupes de multiacheminement fournissent à la base des ressources de nom d'hôte logique et d'adresse partagée. Sur un nœud, le même groupe de multiacheminement peut héberger un nombre indéfini de ressources de nom d'hôte logique ou d'adresse partagée. Pour contrôler la connectivité de réseau public des nœuds du cluster, vous pouvez créer un multiacheminement.

Pour plus d'informations sur les ressources de nom d'hôte logique et d'adresse partagée, reportez-vous au manuel *Sun Cluster Data Services Planning and Administration Guide for Solaris OS*.

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 nœuds 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 de défaillance d'un des adaptateurs, le logiciel Solaris de multiacheminement sur réseau IP sur Sun Cluster est invoqué pour effectuer le basculement de l'interface défaillante vers un autre adaptateur dans le groupe.

Index

T

V