MetroCluster

Enterprise applications

NetApp
May 09, 2024

Sommaire

MetroCluster . . . 1
- Architecture physique MetroCluster et bases de données Oracle . . . 1
- Architecture logique MetroCluster et bases de données Oracle . . . 5
- Les bases de données Oracle avec SyncMirror . . . 12
- Basculement de base de données Oracle avec MetroCluster . . . 13
- Bases de données Oracle, MetroCluster et NVFAIL . . . 14
- Instance unique Oracle sur MetroCluster . . . 16
- Oracle RAC étendu sur MetroCluster . . . 17

MetroCluster

Architecture physique MetroCluster et bases de données Oracle

Pour comprendre le fonctionnement des bases de données Oracle dans un environnement MetroCluster, il est nécessaire d'expliquer la conception physique d'un système MetroCluster.

Cette documentation remplace le rapport technique *TR-4592 : Oracle on MetroCluster.*

MetroCluster est disponible dans 3 configurations différentes

- Paires HAUTE DISPONIBILITÉ avec connectivité IP
- Paires HAUTE DISPONIBILITÉ avec connectivité FC
- Contrôleur unique avec connectivité FC

[REMARQUE]le terme « connectivité » fait référence à la connexion en cluster utilisée pour la réplication entre sites. Il ne fait pas référence aux protocoles hôtes. Tous les protocoles côté hôte sont pris en charge comme d'habitude dans une configuration MetroCluster, quel que soit le type de connexion utilisé pour les communications entre clusters.

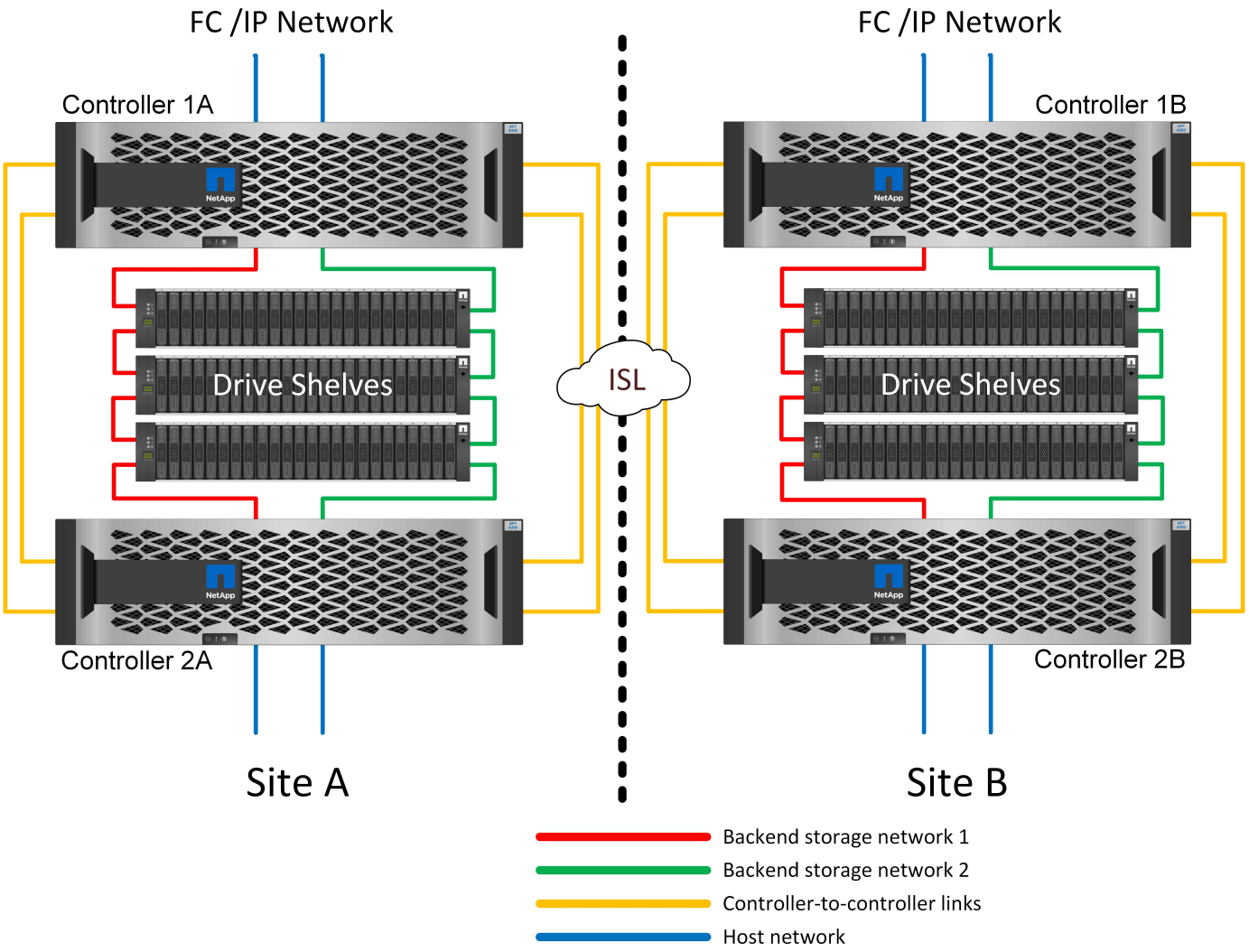
IP MetroCluster

La configuration IP MetroCluster à paire haute disponibilité utilise deux ou quatre nœuds par site. Cette option de configuration augmente la complexité et les coûts liés à l'option à deux nœuds, mais elle offre un avantage important : la redondance intrasite. Une simple panne de contrôleur ne nécessite pas l'accès aux données via le WAN. L'accès aux données reste local via l'autre contrôleur local.

La plupart des clients choisissent la connectivité IP, car les exigences d'infrastructure sont plus simples. Auparavant, la connectivité inter-sites à haut débit était généralement plus facile à provisionner avec des commutateurs FC et fibre noire. Cependant, les circuits IP à haut débit et à faible latence sont aujourd'hui plus facilement disponibles.

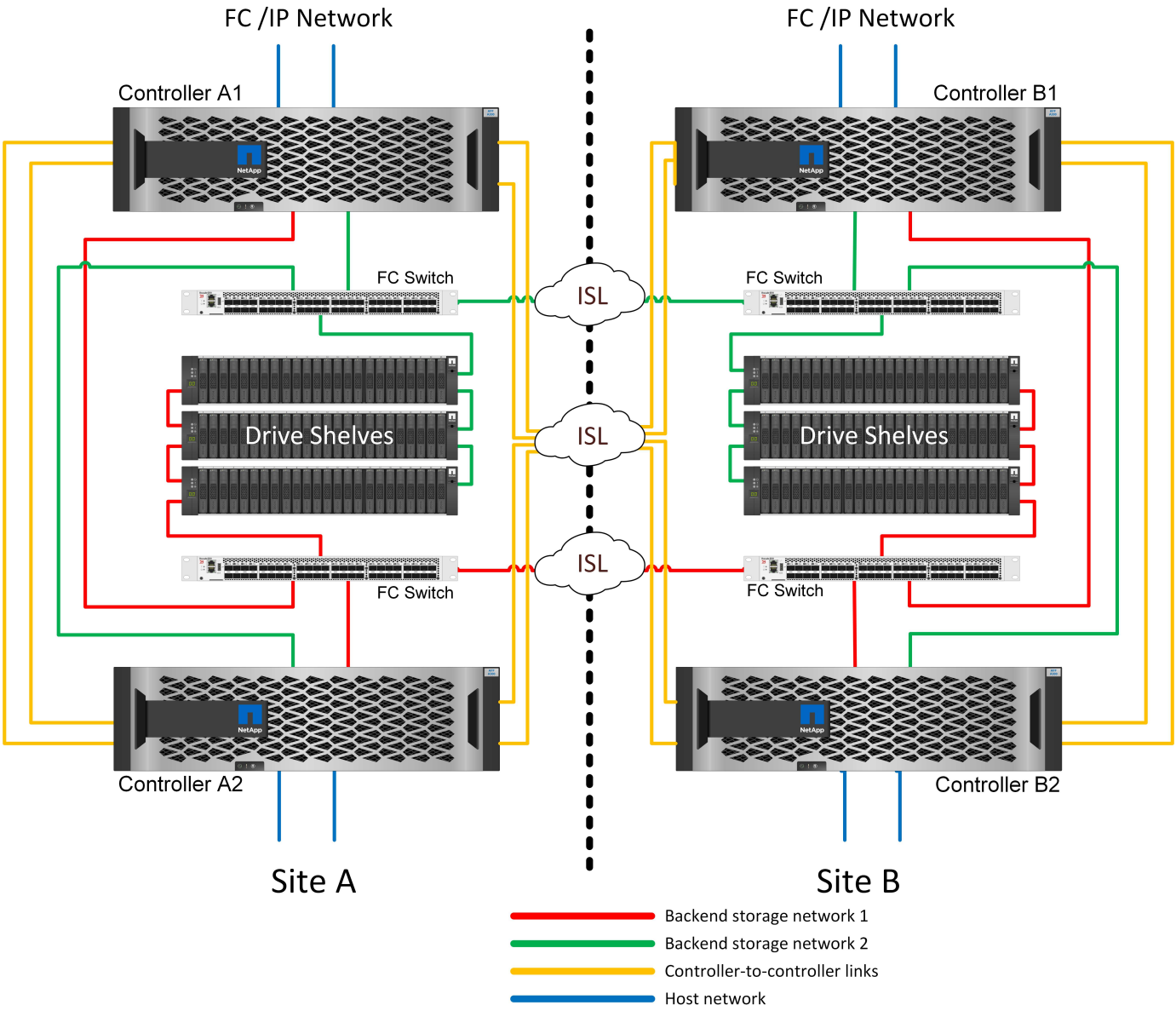
L'architecture est également plus simple, car les contrôleurs disposent des seules connexions entre les sites. Dans les MetroCluster FC, un contrôleur écrit directement sur les disques du site opposé et requiert ainsi des connexions SAN, des commutateurs et des ponts supplémentaires. En revanche, un contrôleur dans une configuration IP écrit sur les lecteurs opposés via le contrôleur.

Pour plus d'informations, consultez la documentation officielle de ONTAP et "Architecture et conception de la solution IP de MetroCluster".



MetroCluster FC à connexion SAN HA-pair

La configuration MetroCluster FC à paire haute disponibilité utilise deux ou quatre nœuds par site. Cette option de configuration augmente la complexité et les coûts liés à l'option à deux nœuds, mais elle offre un avantage important : la redondance intrasite. Une simple panne de contrôleur ne nécessite pas l'accès aux données via le WAN. L'accès aux données reste local via l'autre contrôleur local.



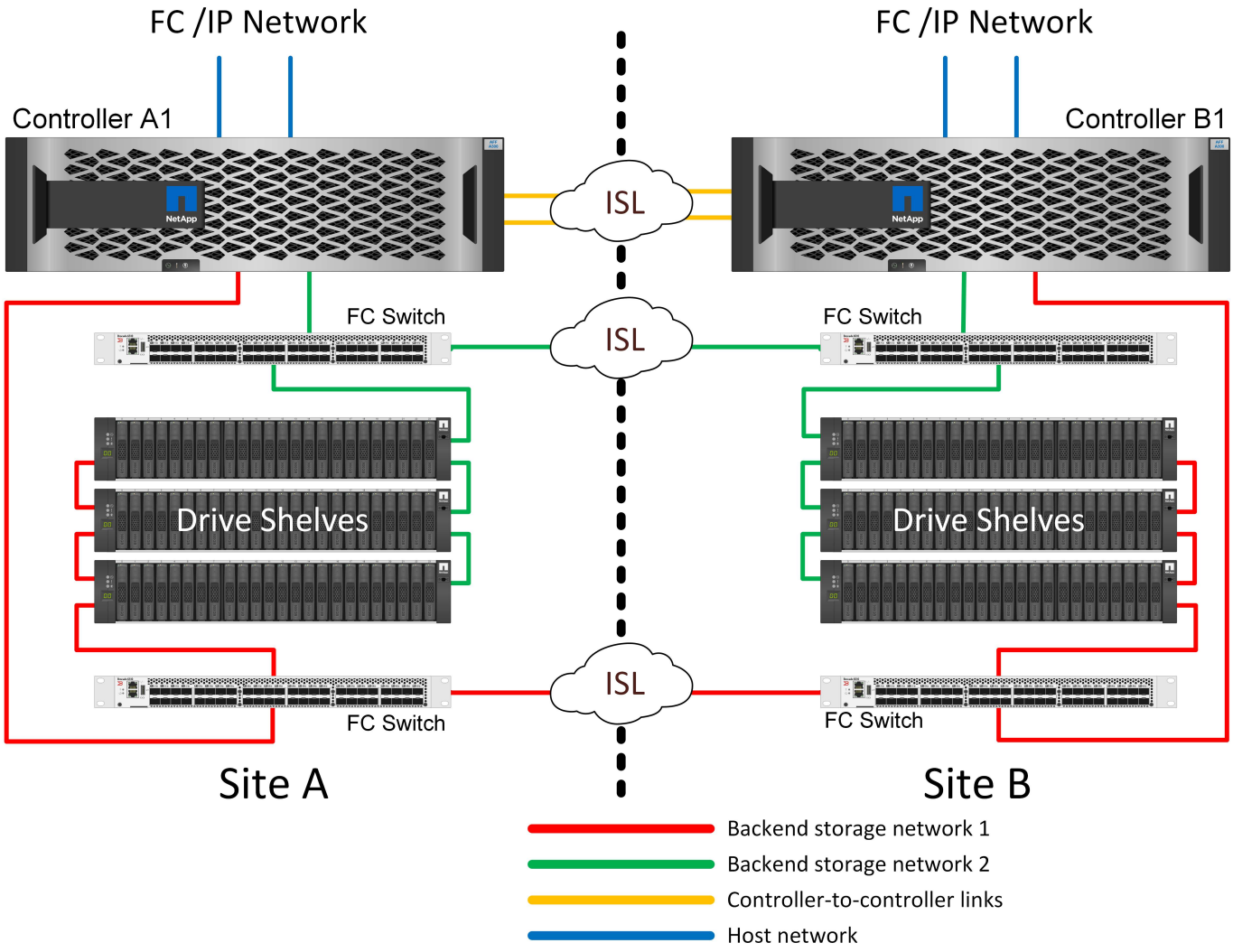
Certaines infrastructures multisites ne sont pas conçues pour les opérations en mode actif-actif. Elles sont plutôt utilisées comme site principal et site de reprise après incident. Dans ce cas, il est généralement préférable d'utiliser une option MetroCluster à paire HA pour les raisons suivantes :

- Bien qu'un cluster MetroCluster à deux nœuds soit un système haute disponibilité, toute panne inattendue d'un contrôleur ou une maintenance planifiée implique que les services de données soient en ligne sur le site opposé. Si la connectivité réseau entre les sites ne prend pas en charge la bande passante requise, les performances sont affectées. La seule option serait également de basculer les différents systèmes d'exploitation hôtes et les services associés vers le site secondaire. Le cluster MetroCluster de paire haute disponibilité élimine ce problème, car la perte d'un contrôleur simplifie le basculement au sein du même site.
- Certaines topologies réseau ne sont pas conçues pour l'accès intersite, mais utilisent des sous-réseaux différents ou des SAN FC isolés. Dans ce cas, le cluster MetroCluster à deux nœuds ne fonctionne plus comme un système haute disponibilité, car le contrôleur secondaire ne peut plus transmettre de données aux serveurs sur le site opposé. L'option MetroCluster de paire haute disponibilité est nécessaire pour assurer une redondance complète.
- Si une infrastructure à deux sites est considérée comme une seule infrastructure extrêmement disponible, la configuration MetroCluster à deux nœuds est adaptée. Toutefois, si le système doit fonctionner pendant

une période prolongée après une panne sur le site, une paire haute disponibilité est recommandée, car la haute disponibilité continue d'être disponible sur un seul site.

MetroCluster FC à deux nœuds avec connexion SAN

La configuration MetroCluster à deux nœuds n'utilise qu'un nœud par site. Cette conception est plus simple que l'option de paire haute disponibilité, car le nombre de composants à configurer et à gérer est inférieur. Elle a également réduit les besoins en infrastructure en termes de câblage et de commutation FC. Enfin, il réduit les coûts.



L'impact évident de cette conception est que la défaillance du contrôleur sur un seul site signifie que les données sont disponibles depuis le site opposé. Cette restriction n'est pas nécessairement un problème. De nombreuses entreprises disposent d'opérations de data Center multisites avec des réseaux étendus, ultra-rapides et à faible latence qui fonctionnent essentiellement comme une infrastructure unique. Dans ce cas, la version à deux nœuds de MetroCluster est la configuration préférée. Plusieurs fournisseurs de services utilisent actuellement des systèmes à deux nœuds de plusieurs pétaoctets.

Fonctions de résilience MetroCluster

Une solution MetroCluster ne présente aucun point de défaillance unique :

- Chaque contrôleur dispose de deux chemins d'accès indépendants aux tiroirs disques sur le site local.

- Chaque contrôleur dispose de deux chemins d'accès indépendants aux tiroirs disques du site distant.
- Chaque contrôleur dispose de deux chemins d'accès indépendants aux contrôleurs sur le site opposé.
- Dans la configuration HA-pair, chaque contrôleur dispose de deux chemins vers son partenaire local.

En résumé, n'importe quel composant de la configuration peut être supprimé sans compromettre la capacité de MetroCluster à transmettre des données. La seule différence en termes de résilience entre les deux options est que la version à paire haute disponibilité reste un système de stockage haute disponibilité global après une panne de site.

Architecture logique MetroCluster et bases de données Oracle

Comprendre le fonctionnement des bases de données Oracle dans un environnement MetroCluster alsop nécessite une explication de la fonctionnalité logique d'un système MetroCluster.

Protection contre les défaillances de site : NVRAM et MetroCluster

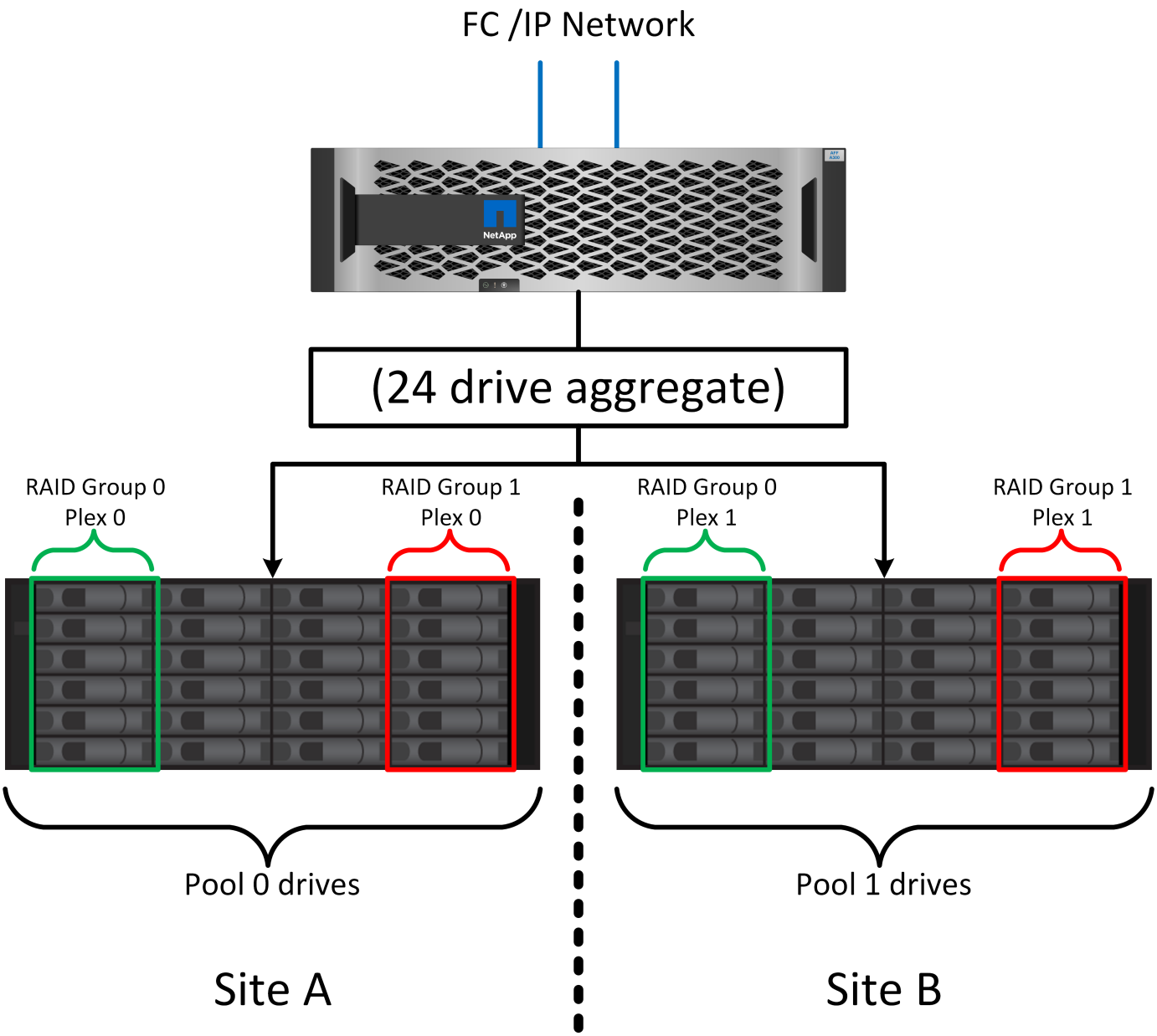
MetroCluster étend la protection des données NVRAM de plusieurs manières :

- Dans une configuration à deux nœuds, les données NVRAM sont répliquées au partenaire distant à l'aide des liens ISL (Inter-Switch Links).
- Dans une configuration de paire haute disponibilité, les données NVRAM sont répliquées à la fois vers le partenaire local et vers un partenaire distant.
- Une écriture n'est pas validée tant qu'elle n'est pas répliquée à tous les partenaires. Cette architecture protège les E/S à la volée contre les défaillances de site en répliquant les données NVRAM sur un partenaire distant. Ce processus n'est pas impliqué dans la réplication des données au niveau des disques. Le contrôleur propriétaire des agrégats est responsable de la réplication des données en écrivant dans les deux plexes de l'agrégat. Cependant, il doit toujours assurer une protection contre les pertes d'E/S à la volée en cas de perte du site. Les données NVRAM répliquées sont uniquement utilisées si un contrôleur partenaire doit prendre le relais en cas de défaillance d'un contrôleur.

Protection contre les pannes de site et de tiroir : SyncMirror et plexes

SyncMirror est une technologie de mise en miroir qui améliore, mais ne remplace pas, RAID DP ou RAID-TEC. Il met en miroir le contenu de deux groupes RAID indépendants. La configuration logique est la suivante :

1. Les disques sont configurés en deux pools en fonction de leur emplacement. Un pool est composé de tous les disques du site A et le second est composé de tous les disques du site B.
2. Un pool de stockage commun, appelé agrégat, est ensuite créé à partir de jeux en miroir de groupes RAID. Un nombre égal de lecteurs est tiré de chaque site. Par exemple, un agrégat SyncMirror de 20 disques se compose de 10 disques du site A et de 10 disques du site B.
3. Chaque jeu de disques d'un site donné est automatiquement configuré comme un ou plusieurs groupes RAID DP ou RAID-TEC entièrement redondants, indépendamment de l'utilisation de la mise en miroir. Cette utilisation de la mise en miroir RAID assure la protection des données même après la perte d'un site.



La figure ci-dessus illustre un exemple de configuration SyncMirror. Un agrégat de 24 disques a été créé sur le contrôleur avec 12 disques à partir d'un tiroir alloué sur le site A et 12 disques à partir d'un tiroir alloué sur le site B. Les disques ont été regroupés en deux groupes RAID en miroir. Le groupe RAID 0 comprend un plex de 6 disques sur le site A mis en miroir sur un plex de 6 disques sur le site B. De même, le groupe RAID 1 comprend un plex de 6 disques sur le site A mis en miroir sur un plex de 6 disques sur le site B.

SyncMirror est généralement utilisé pour assurer la mise en miroir à distance avec les systèmes MetroCluster, avec une copie des données sur chaque site. Il a parfois été utilisé pour fournir un niveau supplémentaire de redondance dans un seul système. Il assure en particulier la redondance au niveau du tiroir. Un tiroir disque contient déjà deux blocs d'alimentation et contrôleurs. Dans l'ensemble, il ne s'agit pas d'une simple tôlerie, mais dans certains cas, une protection supplémentaire peut être garantie. Par exemple, un client NetApp a déployé SyncMirror sur une plateforme mobile d'analytique en temps réel utilisée lors des tests automobiles. Le système a été séparé en deux racks physiques fournis avec des alimentations indépendantes et des systèmes UPS indépendants.

Échec de la redondance : NVFAIL

Comme nous l'avons vu précédemment, une écriture n'est pas validée tant qu'elle n'a pas été connectée à la NVRAM et à la NVRAM locales sur au moins un autre contrôleur. Cette approche évite toute panne matérielle ou de courant qui entraîne une perte des E/S à la volée En cas de panne de la mémoire NVRAM locale ou de la connectivité aux autres nœuds, les données ne seront plus mises en miroir.

Si la mémoire NVRAM locale signale une erreur, le nœud s'arrête. Cet arrêt entraîne le basculement vers un contrôleur partenaire lorsque des paires haute disponibilité sont utilisées. Avec MetroCluster, le comportement dépend de la configuration globale choisie, mais il peut entraîner un basculement automatique vers la note distante. Dans tous les cas, aucune donnée n'est perdue parce que le contrôleur qui connaît la défaillance n'a pas acquitté l'opération d'écriture.

Une défaillance de connectivité site à site qui bloque la réplication NVRAM sur des nœuds distants est une situation plus compliquée. Les écritures ne sont plus répliquées sur les nœuds distants, ce qui crée un risque de perte de données en cas d'erreur catastrophique sur un contrôleur. Plus important encore, une tentative de basculement vers un autre nœud dans ces conditions entraîne une perte de données.

Le facteur de contrôle est de savoir si la NVRAM est synchronisée. Si la mémoire NVRAM est synchronisée, le basculement nœud à nœud peut se poursuivre sans risque de perte de données. Dans une configuration MetroCluster, si la mémoire NVRAM et les plexes d'agrégats sous-jacents sont synchronisés, vous pouvez procéder au basculement sans risque de perte de données.

ONTAP n'autorise pas le basculement ou le basculement lorsque les données ne sont pas synchronisées, sauf si le basculement ou le basculement est forcé. Le fait de forcer une modification des conditions de cette manière reconnaît que les données peuvent être laissées pour compte dans le contrôleur d'origine et que la perte de données est acceptable.

Les bases de données et autres applications sont particulièrement vulnérables à la corruption en cas de basculement ou de basculement forcé, car elles conservent des caches internes de données plus volumineux sur disque. En cas de basculement forcé ou de basculement forcé, les modifications précédemment reconnues sont effectivement supprimées. Le contenu de la baie de stockage recule dans le temps et l'état du cache ne reflète plus l'état des données sur le disque.

Afin d'éviter ce genre de situation, ONTAP permet de configurer les volumes pour une protection spéciale contre les défaillances de mémoire NVRAM. Lorsqu'il est déclenché, ce mécanisme de protection entraîne l'entrée d'un volume dans un état appelé NVFAIL. Cet état entraîne des erreurs d'E/S qui provoquent une panne de l'application. Cette panne provoque l'arrêt des applications, qui n'utilisent donc pas de données obsolètes. Les données ne doivent pas être perdues car des données de transaction validées doivent être présentes dans les journaux. Les étapes suivantes habituelles sont qu'un administrateur arrête complètement les hôtes avant de remettre manuellement en ligne les LUN et les volumes. Bien que ces étapes puissent impliquer un certain travail, cette approche est le moyen le plus sûr d'assurer l'intégrité des données. Toutes les données n'ont pas besoin de cette protection. C'est pourquoi NVFAIL peut être configuré volume par volume.

Paires HAUTE DISPONIBILITÉ et MetroCluster

MetroCluster est disponible dans deux configurations : deux nœuds et paire haute disponibilité. La configuration à deux nœuds se comporte de la même manière qu'une paire haute disponibilité par rapport à la mémoire NVRAM. En cas de défaillance soudaine, le nœud partenaire peut relire les données NVRAM pour assurer la cohérence des disques et garantir la perte d'aucune écriture reconnue.

La configuration HA-pair réplique également la mémoire NVRAM sur le nœud partenaire local. Une simple défaillance de contrôleur entraîne une relecture NVRAM sur le nœud partenaire, comme c'est le cas avec une paire haute disponibilité autonome sans MetroCluster. En cas de perte complète soudaine d'un site, le site

distant dispose également de la mémoire NVRAM requise pour assurer la cohérence des disques et commencer à transmettre les données.

Un aspect important de MetroCluster est que les nœuds distants ne peuvent pas accéder aux données des partenaires dans des conditions de fonctionnement normales. Chaque site fonctionne essentiellement comme un système indépendant qui peut assumer la personnalité du site opposé. Ce processus est connu sous le nom de basculement et inclut un basculement planifié dans lequel les opérations sur site sont migrées sans interruption vers le site opposé. Il comprend également les situations non planifiées où un site est perdu et un basculement manuel ou automatique est nécessaire dans le cadre de la reprise d'activité.

Basculement et rétablissement

Les termes « switchover and switchoback » font référence au processus de transition des volumes entre des contrôleurs distants dans une configuration MetroCluster. Ce processus s'applique uniquement aux nœuds distants. Lorsque MetroCluster est utilisé dans une configuration à quatre volumes, le basculement de nœud local est le même processus de basculement et de rétablissement que celui décrit précédemment.

Basculement et rétablissement planifiés

Un basculement ou rétablissement planifié est similaire à un basculement ou un rétablissement entre les nœuds. Ce processus comporte plusieurs étapes et peut sembler prendre plusieurs minutes, mais il s'agit d'une transition progressive et progressive des ressources de stockage et de réseau. Le moment où les transferts de contrôle se produisent beaucoup plus rapidement que le temps nécessaire à l'exécution de la commande complète.

La principale différence entre le basculement/rétablissement et le basculement/rétablissement réside dans l'effet sur la connectivité FC SAN. Avec le Takeover/Giveback local, un hôte subit la perte de tous les chemins FC vers le nœud local et s'appuie sur son MPIO natif pour le basculer vers des chemins alternatifs disponibles. Les ports ne sont pas déplacés. Avec le basculement et le rétablissement, les ports cibles FC virtuels des contrôleurs passent à l'autre site. Ils cessent d'exister sur le SAN pendant un instant, puis réapparaissent sur un autre contrôleur.

SyncMirror expire

SyncMirror est une technologie de mise en miroir ONTAP qui offre une protection contre les défaillances de tiroirs. Lorsque les tiroirs sont séparés sur une distance, les données sont protégées à distance.

SyncMirror ne fournit pas de mise en miroir synchrone universelle. Le résultat est une meilleure disponibilité. Certains systèmes de stockage utilisent une mise en miroir totale ou nulle constante, parfois appelée mode domino. Cette forme de mise en miroir est limitée dans l'application car toutes les activités d'écriture doivent cesser en cas de perte de la connexion au site distant. Sinon, une écriture existerait sur un site, mais pas sur l'autre. Généralement, ces environnements sont configurés pour mettre les LUN hors ligne en cas de perte de la connectivité site à site pendant plus d'une courte période (par exemple, 30 secondes).

Ce comportement est souhaitable pour un petit sous-ensemble d'environnements. Cependant, la plupart des applications nécessitent une solution capable de garantir une réplication synchrone dans des conditions normales de fonctionnement, mais avec la possibilité de suspendre la réplication. Une perte complète de la connectivité site à site est souvent considérée comme une situation proche d'une catastrophe. Généralement, ces environnements sont maintenus en ligne et donnent accès aux données jusqu'à ce que la connectivité soit réparée ou qu'une décision officielle soit prise de fermer l'environnement pour protéger les données. Il n'est pas rare d'avoir besoin d'arrêter automatiquement l'application uniquement en raison d'une défaillance de réplication à distance.

SyncMirror prend en charge les exigences de mise en miroir synchrone avec la flexibilité d'un délai d'expiration. Si la connectivité à la télécommande et/ou au plex est perdue, une minuterie de 30 secondes

commence à s'arrêter. Lorsque le compteur atteint 0, le traitement des E/S d'écriture reprend en utilisant les données locales. La copie distante des données est utilisable, mais elle est figée à temps jusqu'à ce que la connectivité soit rétablie. La resynchronisation exploite des snapshots au niveau de l'agrégat pour rétablir le système en mode synchrone aussi rapidement que possible.

Notamment, dans de nombreux cas, ce type de réplication universelle en mode domino tout ou rien est mieux implémenté au niveau de la couche applicative. Par exemple, Oracle DataGuard inclut le mode de protection maximum, ce qui garantit la réplication à long terme en toutes circonstances. Si la liaison de réplication échoue pendant une période dépassant un délai configurable, les bases de données s'arrêtent.

Basculement automatique sans surveillance avec Fabric Attached MetroCluster

Le basculement automatique sans surveillance (AUSO) est une fonctionnalité MetroCluster intégrée au fabric qui offre une forme de haute disponibilité intersite. Comme évoqué précédemment, MetroCluster est disponible en deux types : un contrôleur unique sur chaque site ou une paire haute disponibilité sur chaque site. L'avantage principal de l'option haute disponibilité est que l'arrêt planifié ou non planifié du contrôleur permet toujours une E/S locale. L'avantage de l'option à nœud unique est de réduire les coûts, la complexité et l'infrastructure.

La principale valeur d'AUSO est d'améliorer les fonctionnalités haute disponibilité des systèmes MetroCluster connectés à la structure. Chaque site surveille l'état de santé du site opposé et, si aucun nœud n'est encore utilisé pour transmettre des données, l'AUSO assure un basculement rapide. Cette approche est particulièrement utile dans les configurations MetroCluster avec un seul nœud par site, car elle rapproche la configuration d'une paire haute disponibilité en termes de disponibilité.

AUSO ne peut pas offrir de surveillance complète au niveau d'une paire HA. Une paire haute disponibilité peut offrir une haute disponibilité, car elle inclut deux câbles physiques redondants pour une communication nœud à nœud directe. En outre, les deux nœuds d'une paire haute disponibilité ont accès au même ensemble de disques sur des boucles redondantes, ce qui permet à un nœud de suivre l'état d'un autre nœud sur une autre route.

Il existe des clusters MetroCluster sur plusieurs sites pour lesquels la communication nœud à nœud et l'accès au disque reposent sur la connectivité réseau site à site. La capacité à surveiller le pouls du reste du cluster est limitée. AUSO doit faire la distinction entre une situation où l'autre site est en fait hors service plutôt qu'indisponible en raison d'un problème de réseau.

Par conséquent, un contrôleur d'une paire haute disponibilité peut demander un basculement s'il détecte une panne de contrôleur qui s'est produite pour une raison spécifique, par exemple une situation critique du système. Elle peut également déclencher un basculement en cas de perte complète de la connectivité, parfois appelée « perte de pulsation ».

Un système MetroCluster ne peut effectuer un basculement automatique en toute sécurité que lorsqu'une panne spécifique est détectée sur le site d'origine. En outre, le contrôleur qui devient propriétaire du système de stockage doit être en mesure de garantir la synchronisation des données du disque et de la NVRAM. Le contrôleur ne peut pas garantir la sécurité d'un basculement simplement parce qu'il a perdu le contact avec le site source, qui pourrait toujours être opérationnel. Pour plus d'informations sur les options d'automatisation d'un basculement, reportez-vous aux informations sur la solution MetroCluster Tiebreaker (MCTB) dans la section suivante.

Disjoncteur d'attache MetroCluster avec MetroCluster FAS

Le "NetApp MetroCluster Tiebreaker" Le logiciel peut s'exécuter sur un troisième site afin de contrôler l'état de santé de votre environnement MetroCluster, d'envoyer des notifications et de forcer un basculement en cas d'incident. Une description complète du disjoncteur d'attache se trouve sur le "Site de support NetApp", Mais le but principal du Tiebreaker de MetroCluster est de détecter la perte de site. Il doit également faire la distinction

entre la perte du site et une perte de connectivité. Par exemple, le basculement ne doit pas se produire car le disjoncteur d'attache n'a pas pu atteindre le site principal. C'est pourquoi le disjoncteur d'attache surveille également la capacité du site distant à contacter le site principal.

Le basculement automatique avec AUSO est également compatible avec le MCTB. AUSO réagit très rapidement car il est conçu pour détecter des événements de défaillance spécifiques, puis n'invoque le basculement que lorsque les plexes NVRAM et SyncMirror sont synchronisés.

En revanche, le disjoncteur principal est situé à distance et doit donc attendre qu'une minuterie s'écoule avant de déclarer un site mort. Le disjoncteur d'attache détecte finalement le type de défaillance de contrôleur couverte par l'AUSO, mais en général, l'AUSO a déjà commencé le basculement et éventuellement terminé le basculement avant que le disjoncteur d'attache n'agisse. La deuxième commande de basculement qui en résulte provient du Tiebreaker serait rejetée.

*Attention : *le logiciel MCTB ne vérifie pas que la mémoire NVRAM était et/ou que les plexes sont synchronisés lorsque vous forcez un basculement. Le basculement automatique, s'il est configuré, doit être désactivé pendant les opérations de maintenance qui entraînent une perte de synchronisation des plexes NVRAM ou SyncMirror.

En outre, le MCTB peut ne pas traiter un désastre roulant qui conduit à la séquence d'événements suivante :

1. La connectivité entre les sites est interrompue pendant plus de 30 secondes.
2. La réplication SyncMirror est obsolète et les opérations se poursuivent sur le site principal, ce qui ne permet pas au réplica distant d'être obsolète.
3. Le site primaire est perdu. Le résultat est la présence de modifications non répliquées sur le site primaire. Un basculement peut alors se révéler indésirable pour plusieurs raisons, notamment :
    - Certaines données critiques peuvent être présentes sur le site primaire et peuvent être récupérées à terme. Un basculement qui a permis à l'application de continuer à fonctionner aurait pour effet de supprimer ces données stratégiques.
    - Des données peuvent être mises en cache pour une application sur le site survivant qui utilisait des ressources de stockage sur le site principal au moment de la perte du site. Le basculement introduit une version obsolète des données qui ne correspond pas au cache.
    - Des données peuvent être mises en cache sur un système d'exploitation du site survivant qui utilisait des ressources de stockage sur le site principal au moment de la perte du site. Le basculement introduit une version obsolète des données qui ne correspond pas au cache. L'option la plus sûre est de configurer le Tiebreaker pour envoyer une alerte s'il détecte une défaillance du site et demander à une personne de décider si elle doit forcer un basculement. Il peut être nécessaire d'abord d'arrêter les applications et/ou les systèmes d'exploitation pour effacer les données en cache. En outre, les paramètres NVFAIL peuvent être utilisés pour renforcer la protection et rationaliser le processus de basculement.

Mediator ONTAP avec MetroCluster IP

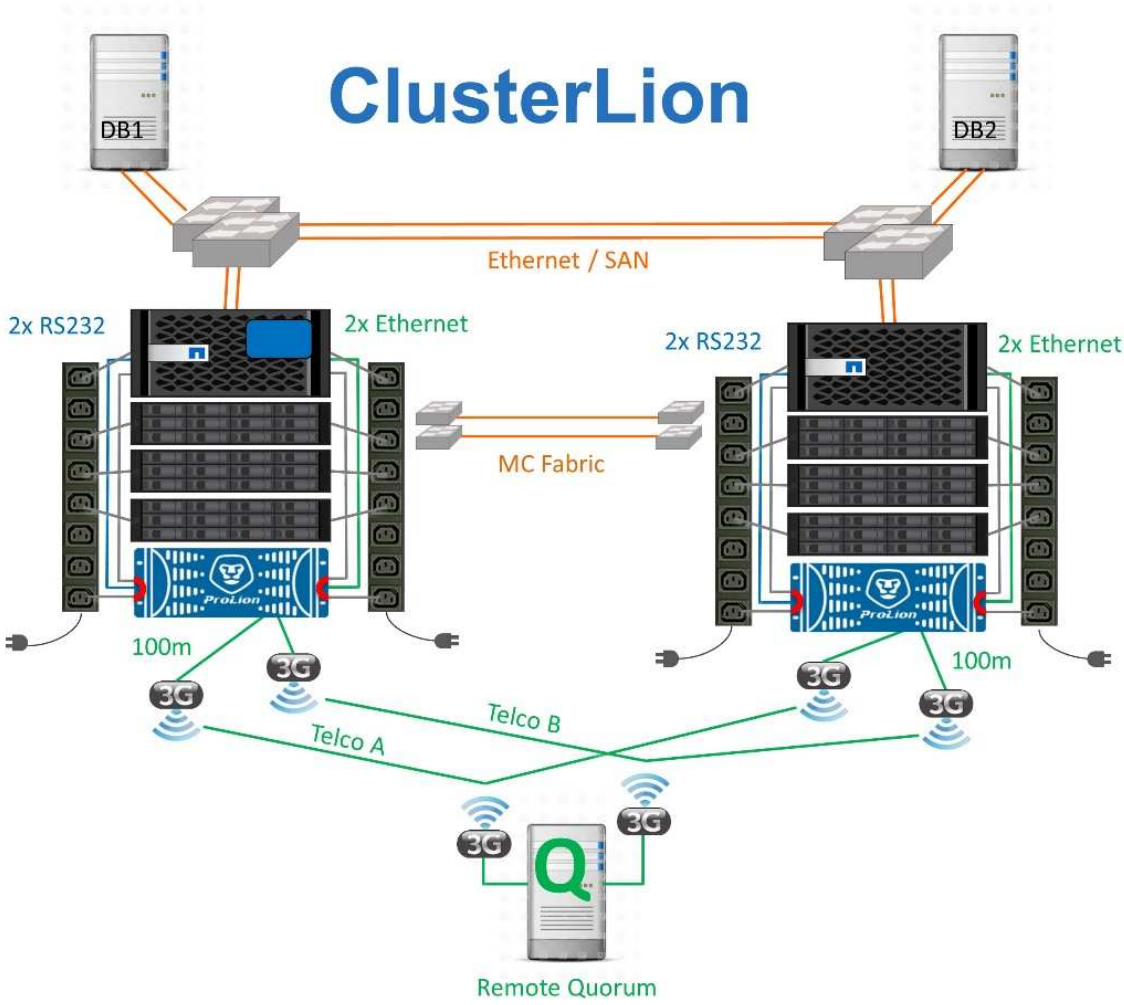
Le médiateur ONTAP est utilisé avec MetroCluster IP et certaines autres solutions ONTAP. Il fonctionne comme un service disjoncteur d'attache classique, tout comme le logiciel disjoncteur d'attache MetroCluster mentionné ci-dessus, mais comprend également une fonctionnalité essentielle, qui effectue un basculement automatique sans surveillance.

Un MetroCluster FAS dispose d'un accès direct aux dispositifs de stockage sur le site opposé. Cela permet à un contrôleur MetroCluster de surveiller l'intégrité des autres contrôleurs en lisant les données de pulsation à partir des disques. Cela permet à un contrôleur de reconnaître la défaillance d'un autre contrôleur et d'effectuer un basculement.

En revanche, l'architecture IP MetroCluster achemine toutes les E/S exclusivement via la connexion contrôleur-contrôleur ; il n'y a pas d'accès direct aux dispositifs de stockage sur le site distant. Cela limite la capacité d'un contrôleur à détecter les défaillances et à effectuer un basculement. Le Mediator ONTAP est donc requis comme dispositif Tiebreaker pour détecter la perte du site et effectuer automatiquement un basculement.

Troisième site virtuel avec ClusterLion

ClusterLion est un dispositif de surveillance MetroCluster avancé qui fonctionne comme un troisième site virtuel. Cette approche permet de déployer MetroCluster en toute sécurité dans une configuration à deux sites avec une fonctionnalité de basculement entièrement automatisée. De plus, ClusterLion peut effectuer un moniteur de niveau réseau supplémentaire et exécuter des opérations de post-basculement. La documentation complète est disponible auprès de ProLion.



- Les appliances ClusterLion contrôlent l'état des contrôleurs à l'aide de câbles série et Ethernet directement connectés.
- Les deux appareils sont connectés l'un à l'autre à l'aide de connexions 3G sans fil redondantes.
- L'alimentation vers le contrôleur ONTAP est acheminée via des relais internes. En cas de panne de site, ClusterLion, qui contient un système UPS interne, coupe les connexions d'alimentation avant d'appeler un basculement. Ce processus permet de s'assurer qu'aucune condition de split-brain ne se produit.
- ClusterLion effectue un basculement dans le délai d'attente SyncMirror de 30 secondes ou pas du tout.

- ClusterLion n’effectue pas de basculement à moins que les États des plexes NVRAM et SyncMirror ne soient synchronisés.
- Étant donné que ClusterLion effectue un basculement uniquement si MetroCluster est entièrement synchronisé, NVFAIL n’est pas nécessaire. Cette configuration permet aux environnements couvrant l’ensemble des sites, tels qu’un RAC Oracle étendu, de rester en ligne, même pendant un basculement non planifié.
- Il inclut les protocoles Fabric-Attached MetroCluster et MetroCluster IP

Les bases de données Oracle avec SyncMirror

Le socle de la protection des données Oracle avec un système MetroCluster est SyncMirror, une technologie de mise en miroir synchrone scale-out aux performances maximales.

Protection des données avec SyncMirror

Au niveau le plus simple, la réplication synchrone implique que toute modification doit être apportée des deux côtés du stockage en miroir avant d’être reconnue. Par exemple, si une base de données écrit un journal ou si un invité VMware est en cours de correction, une écriture ne doit jamais être perdue. Au niveau du protocole, le système de stockage ne doit pas accuser réception de l’écriture tant qu’il n’a pas été validé sur un support non volatile des deux sites. Ce n’est qu’à cette condition qu’il est possible de continuer sans risque de perte de données.

L’utilisation d’une technologie de réplication synchrone est la première étape de la conception et de la gestion d’une solution de réplication synchrone. Il est important de comprendre ce qui pourrait se passer lors de divers scénarios de défaillance planifiés ou non. Les solutions de réplication synchrone offrent toutes des fonctionnalités différentes. Si vous avez besoin d’une solution avec un objectif de point de récupération de zéro, c’est-à-dire sans perte de données, tous les scénarios de défaillance doivent être pris en compte. En particulier, quel est le résultat escompté lorsque la réplication est impossible en raison d’une perte de connectivité entre les sites ?

Disponibilité des données SyncMirror

La réplication MetroCluster repose sur la technologie NetApp SyncMirror, conçue pour basculer efficacement en mode synchrone et en sortir. Cette fonctionnalité répond aux exigences des clients qui demandent une réplication synchrone, mais qui ont également besoin d’une haute disponibilité pour leurs services de données. Par exemple, si la connectivité à un site distant est coupée, il est généralement préférable que le système de stockage continue de fonctionner dans un état non répliqué.

De nombreuses solutions de réplication synchrone ne peuvent fonctionner qu’en mode synchrone. Ce type de réplication « tout ou rien » est parfois appelé mode domino. Ces systèmes de stockage cessent d’accéder aux données au lieu d’interrompre la synchronisation des copies locales et distantes des données. Si la réplication est forcée, la resynchronisation peut prendre beaucoup de temps et laisser un client exposé à des pertes de données complètes pendant la période de rétablissement de la mise en miroir.

Non seulement SyncMirror peut basculer en mode synchrone sans interruption si le site distant est inaccessible, mais il peut également rapidement resynchroniser vers un état RPO = 0 une fois la connectivité restaurée. La copie obsolète des données sur le site distant peut également être conservée dans un état utilisable lors de la resynchronisation, garantissant la présence à tout moment de copies locales et distantes des données.

Si le mode domino est requis, NetApp propose SnapMirror synchrone (SM-S). Des options au niveau de

l'application existent également, telles qu'Oracle DataGuard ou des délais d'expiration étendus pour la mise en miroir des disques côté hôte. Pour plus d'informations et d'options, consultez votre équipe de compte NetApp ou partenaire.

Basculement de base de données Oracle avec MetroCluster

```
 Metrocluster is an ONTAP feature that can protect your Oracle databases
with RPO=0 synchronous mirroring across sites, and it scales up to support
hundreds of databases on a single MetroCluster system. It's also simple to
use. The use of MetroCluster does not necessarily add to or change any
best practices for operating a enterprise applications and databases.
Les bonnes pratiques habituelles s'appliquent toujours. Si vos besoins
requièrent uniquement une protection des données avec un objectif de point
de récupération de 0, MetroCluster répond à ce besoin. Cependant, la
plupart des clients utilisent MetroCluster non seulement pour la
protection des données avec un objectif de point de récupération de 0,
mais aussi pour améliorer l'objectif de délai de restauration en cas
d'incident et fournir un basculement transparent dans le cadre des
activités de maintenance du site.
```

Basculement avec un système d'exploitation préconfiguré

SyncMirror livre une copie synchrone des données au niveau du site de reprise d'activité. La mise à disposition des données requiert un système d'exploitation et les applications associées. L'automatisation de base peut considérablement améliorer le délai de basculement de l'environnement global. Les produits Clusterware tels qu'Oracle RAC, Veritas Cluster Server (VCS) ou VMware HA sont souvent utilisés pour créer un cluster sur les sites et, dans la plupart des cas, le processus de basculement peut être piloté avec de simples scripts.

En cas de perte des nœuds principaux, le cluster (ou les scripts) est configuré de manière à mettre les applications en ligne sur le site secondaire. Une option consiste à créer des serveurs de secours préconfigurés pour les ressources NFS ou SAN qui constituent l'application. En cas de défaillance du site principal, le logiciel de mise en cluster ou l'alternative scriptée effectue une séquence d'actions similaires à celles décrites ci-dessous :

1. Forçage du basculement MetroCluster
2. Découverte de LUN FC (SAN uniquement)
3. Montage de systèmes de fichiers
4. Démarrage de l'application

Cette approche doit avant tout se passer d'un système d'exploitation en cours d'exécution sur le site distant. Il doit être préconfiguré avec des binaires d'application, ce qui signifie également que des tâches telles que l'application de correctifs doivent être effectuées sur les sites principal et de secours. Les binaires d'application peuvent également être mis en miroir vers le site distant et montés en cas d'incident.

La procédure d'activation réelle est simple. Les commandes telles que la découverte de LUN ne nécessitent que quelques commandes par port FC. Le montage du système de fichiers n'est rien de plus qu'un `mount` Et les bases de données et ASM peuvent être démarrés et arrêtés sur l'interface de ligne de commande à l'aide d'une seule commande. Si les volumes et les systèmes de fichiers ne sont pas utilisés sur le site de reprise

d'activité avant le basculement, il n'est pas nécessaire de les définir `dr-force- nvfail` sur les volumes.

Basculement avec un système d'exploitation virtualisé

Le basculement des environnements de base de données peut être étendu pour inclure le système d'exploitation lui-même. En théorie, ce basculement peut être effectué avec des LUN de démarrage, mais le plus souvent avec un système d'exploitation virtualisé. La procédure est similaire aux étapes suivantes :

1. Forçage du basculement MetroCluster
2. Montage des datastores hébergeant les machines virtuelles du serveur de base de données
3. Démarrage des machines virtuelles
4. Démarrage manuel des bases de données ou configuration des machines virtuelles pour démarrer automatiquement les bases de données

Par exemple, un cluster ESX peut couvrir des sites. En cas d'incident, les machines virtuelles peuvent être mises en ligne sur le site de reprise après incident après le basculement. Tant que les datastores hébergeant les serveurs de base de données virtualisés ne sont pas utilisés au moment de l'incident, il n'est pas nécessaire de les définir `dr-force- nvfail` sur les volumes associés.

Bases de données Oracle, MetroCluster et NVFAIL

NVFAIL est une fonctionnalité d'intégrité générale des données de ONTAP conçue pour optimiser la protection de l'intégrité des données avec les bases de données.

Cette section décrit en détail les fonctionnalités de base de ONTAP NVFAIL et aborde également les sujets spécifiques à MetroCluster.

Avec MetroCluster, une écriture n'est pas confirmée tant qu'elle n'a pas été connectée à la NVRAM et à la NVRAM locales sur au moins un autre contrôleur. Cette approche évite toute panne matérielle ou de courant qui entraîne une perte des E/S à la volée En cas de panne de la mémoire NVRAM locale ou de la connectivité aux autres nœuds, les données ne seront plus mises en miroir.

Si la mémoire NVRAM locale signale une erreur, le nœud s'arrête. Cet arrêt entraîne le basculement vers un contrôleur partenaire lorsque des paires haute disponibilité sont utilisées. Avec MetroCluster, le comportement dépend de la configuration globale choisie, mais il peut entraîner un basculement automatique vers la note distante. Dans tous les cas, aucune donnée n'est perdue parce que le contrôleur qui connaît la défaillance n'a pas acquitté l'opération d'écriture.

Une défaillance de connectivité site à site qui bloque la réplication NVRAM sur des nœuds distants est une situation plus compliquée. Les écritures ne sont plus répliquées sur les nœuds distants, ce qui crée un risque de perte de données en cas d'erreur catastrophique sur un contrôleur. Plus important encore, une tentative de basculement vers un autre nœud dans ces conditions entraîne une perte de données.

Le facteur de contrôle est de savoir si la NVRAM est synchronisée. Si la mémoire NVRAM est synchronisée, le basculement nœud à nœud peut se poursuivre sans risque de perte de données. Dans une configuration MetroCluster, si la mémoire NVRAM et les plexes d'agrégats sous-jacents sont synchronisés, vous pouvez effectuer le basculement sans risque de perte de données.

ONTAP n'autorise pas le basculement ou le basculement lorsque les données ne sont pas synchronisées, sauf si le basculement ou le basculement est forcé. Le fait de forcer une modification des conditions de cette manière reconnaît que les données peuvent être laissées pour compte dans le contrôleur d'origine et que la

perte de données est acceptable.

Les bases de données sont particulièrement vulnérables à la corruption si un basculement ou un basculement est forcé, car les bases de données conservent des caches internes de données plus volumineux sur disque. En cas de basculement forcé ou de basculement forcé, les modifications précédemment reconnues sont effectivement supprimées. Le contenu de la baie de stockage recule dans le temps et l'état du cache de la base de données ne reflète plus l'état des données sur le disque.

Afin de protéger les applications de cette situation, ONTAP permet de configurer les volumes pour une protection spéciale contre les défaillances de mémoire NVRAM. Lorsqu'il est déclenché, ce mécanisme de protection entraîne l'entrée d'un volume dans un état appelé NVFAIL. Cet état entraîne des erreurs d'E/S qui entraînent l'arrêt d'une application et n'utilisent donc pas de données obsolètes. Les données ne doivent pas être perdues car des écritures reconnues sont toujours présentes sur le système de stockage et, avec les bases de données, toutes les données de transaction validées doivent être présentes dans les journaux.

Les étapes suivantes habituelles sont qu'un administrateur arrête complètement les hôtes avant de remettre manuellement en ligne les LUN et les volumes. Bien que ces étapes puissent impliquer un certain travail, cette approche est le moyen le plus sûr d'assurer l'intégrité des données. Toutes les données n'ont pas besoin de cette protection. C'est pourquoi NVFAIL peut être configuré volume par volume.

NVFAIL forcé manuellement

Pour forcer un basculement avec un cluster d'applications (y compris VMware, Oracle RAC et autres) distribué sur plusieurs sites, il faut spécifier la méthode la plus sûre `-force-nvfail-all` en ligne de commande. Cette option est disponible en tant que mesure d'urgence pour s'assurer que toutes les données mises en cache sont vidées. Si un hôte utilise des ressources de stockage initialement situées sur le site sinistré, il reçoit des erreurs d'E/S ou un descripteur de fichier obsolète (`ESTALE`) erreur. Les bases de données Oracle planent et les systèmes de fichiers passent entièrement hors ligne ou en mode lecture seule.

Une fois le basculement terminé, le `in-nvfailed-state` L'indicateur doit être effacé et les LUN doivent être mis en ligne. Une fois cette activité terminée, la base de données peut être redémarrée. Ces tâches peuvent être automatisées afin de réduire le RTO.

dr-force-nvfail

En tant que mesure de sécurité générale, réglez le `dr-force-nvfail` drapeau sur tous les volumes accessibles depuis un site distant pendant les opérations normales, ce qui signifie qu'il s'agit d'activités utilisées avant le basculement. Le résultat de ce paramètre est que les volumes distants sélectionnés deviennent indisponibles lorsqu'ils entrent `in-nvfailed-state` lors d'un basculement. Une fois le basculement terminé, le `in-nvfailed-state` L'indicateur doit être effacé et les LUN doivent être mis en ligne. Une fois ces activités terminées, les applications peuvent être redémarrées. Ces tâches peuvent être automatisées afin de réduire le RTO.

Le résultat est similaire à l'utilisation du `-force-nvfail-all` indicateur pour commutateurs manuels. Toutefois, le nombre de volumes affectés peut être limité aux volumes qui doivent être protégés contre les applications ou les systèmes d'exploitation dotés de caches obsolètes.

Il existe deux exigences critiques pour un environnement qui n'utilise pas `dr-force-nvfail` sur les volumes d'application :

- Un basculement forcé ne doit pas se produire plus de 30 secondes après la perte du site principal.
- Le basculement ne doit pas avoir lieu pendant les tâches de maintenance ou tout autre mode dans lequel les plexes SyncMirror ou la réplication NVRAM sont désynchronisés. Le premier critère peut être atteint à l'aide d'un logiciel disjoncteur d'attache configuré pour effectuer un basculement dans les 30 secondes qui

suivent la défaillance d'un site. Cela ne signifie pas que le basculement doit être effectué dans les 30 secondes qui suivent la détection d'une défaillance de site. Cela signifie qu'il n'est plus sûr de forcer un basculement si 30 secondes se sont écoulées depuis qu'un site a été confirmé opérationnel.

Le deuxième critère peut être partiellement respecté en désactivant toutes les fonctionnalités de basculement automatisé lorsque la configuration MetroCluster est désynchronisée. Il est préférable d'opter pour une solution disjoncteur d'attache capable de surveiller l'état de santé de la réplication NVRAM et des plexes SyncMirror. Si le cluster n'est pas entièrement synchronisé, le disjoncteur d'attache ne doit pas déclencher de basculement.

Le logiciel MCTB de NetApp ne peut pas contrôler l'état de la synchronisation. Il doit donc être désactivé lorsque MetroCluster n'est pas synchronisé pour quelque raison que ce soit. ClusterLion inclut des fonctionnalités de surveillance NVRAM et plex et peut être configuré pour ne pas déclencher le basculement à moins que le système MetroCluster ne soit entièrement synchronisé.

Instance unique Oracle sur MetroCluster

Comme indiqué précédemment, la présence d'un système MetroCluster n'ajoute pas nécessairement aux meilleures pratiques d'exploitation d'une base de données ou ne les modifie pas nécessairement. La majorité des bases de données qui s'exécutent actuellement sur les systèmes MetroCluster client sont à instance unique et suivent les recommandations de la documentation Oracle sur ONTAP.

Basculement avec un système d'exploitation préconfiguré

SyncMirror livre une copie synchrone des données au niveau du site de reprise d'activité. La mise à disposition des données requiert un système d'exploitation et les applications associées. L'automatisation de base peut considérablement améliorer le délai de basculement de l'environnement global. Les produits Clusterware tels que Veritas Cluster Server (VCS) sont souvent utilisés pour créer un cluster sur les sites et, dans la plupart des cas, le processus de basculement peut être piloté par des scripts simples.

En cas de perte des nœuds principaux, le cluster (ou les scripts) est configuré de manière à mettre les bases de données en ligne sur le site secondaire. Une option consiste à créer des serveurs de secours préconfigurés pour les ressources NFS ou SAN qui constituent la base de données. En cas de défaillance du site principal, le logiciel de mise en cluster ou l'alternative scriptée effectue une séquence d'actions similaires à celles décrites ci-dessous :

1. Forçage du basculement MetroCluster
2. Découverte de LUN FC (SAN uniquement)
3. Montage de systèmes de fichiers et/ou montage de groupes de disques ASM
4. Démarrage de la base de données

Cette approche doit avant tout se passer d'un système d'exploitation en cours d'exécution sur le site distant. Elles doivent être préconfigurées avec des binaires Oracle, ce qui signifie également que des tâches telles que l'application de correctifs Oracle doivent être effectuées sur les sites principal et de secours. Les binaires Oracle peuvent également être mis en miroir vers le site distant et montés en cas d'incident.

La procédure d'activation réelle est simple. Les commandes telles que la découverte de LUN ne nécessitent que quelques commandes par port FC. Le montage du système de fichiers n'est rien de plus qu'un `mount` Et les bases de données et ASM peuvent être démarrés et arrêtés sur l'interface de ligne de commande à l'aide d'une seule commande. Si les volumes et les systèmes de fichiers ne sont pas utilisés sur le site de reprise

d'activité avant le basculement, il n'est pas nécessaire de les définir `dr-force- nvfail` sur les volumes.

Basculement avec un système d'exploitation virtualisé

Le basculement des environnements de base de données peut être étendu pour inclure le système d'exploitation lui-même. En théorie, ce basculement peut être effectué avec des LUN de démarrage, mais le plus souvent avec un système d'exploitation virtualisé. La procédure est similaire aux étapes suivantes :

1. Forçage du basculement MetroCluster
2. Montage des datastores hébergeant les machines virtuelles du serveur de base de données
3. Démarrage des machines virtuelles
4. Démarrage manuel des bases de données ou configuration des machines virtuelles pour démarrer automatiquement les bases de données par exemple, un cluster ESX peut couvrir des sites. En cas d'incident, les machines virtuelles peuvent être mises en ligne sur le site de reprise après incident après le basculement. Tant que les datastores hébergeant les serveurs de base de données virtualisés ne sont pas utilisés au moment de l'incident, il n'est pas nécessaire de les définir `dr-force- nvfail` sur les volumes associés.

Oracle RAC étendu sur MetroCluster

De nombreux clients optimisent leur RTO en étendant un cluster Oracle RAC sur plusieurs sites, offrant une configuration entièrement active/active. La conception globale devient plus complexe car elle doit inclure la gestion du quorum d'Oracle RAC. En outre, l'accès aux données se fait depuis les deux sites, ce qui signifie qu'un basculement forcé peut entraîner l'utilisation d'une copie obsolète des données.

Bien qu'une copie des données soit présente sur les deux sites, seul le contrôleur qui possède actuellement un agrégat peut assurer le service des données. Par conséquent, avec les clusters RAC étendus, les nœuds distants doivent effectuer des E/S sur une connexion site à site. Il en résulte une latence d'E/S supplémentaire, mais cette latence n'est généralement pas problématique. Le réseau d'interconnexion RAC doit également être étendu entre les sites, ce qui signifie qu'un réseau haut débit à faible latence est requis de toute façon. Si la latence supplémentaire pose problème, le cluster peut être exploité de manière actif-passif. Les opérations exigeantes en E/S devront ensuite être dirigées vers les nœuds RAC locaux vers le contrôleur propriétaire des agrégats. Les nœuds distants effectuent alors des opérations d'E/S plus légères ou sont utilisés uniquement comme serveurs de secours.

Si un RAC étendu actif-actif est requis, la mise en miroir ASM doit être prise en compte à la place de MetroCluster. La mise en miroir ASM permet de privilégier une réplique spécifique des données. Par conséquent, un cluster RAC étendu peut être intégré dans lequel toutes les lectures se produisent localement. Les E/S de lecture ne traversent jamais les sites, ce qui assure la latence la plus faible possible. Toute activité d'écriture doit toujours transiter la connexion intersite, mais ce trafic est inévitable avec toute solution de mise en miroir synchrone.

> Si des LUN de démarrage, y compris des disques de démarrage virtualisés, sont utilisés avec Oracle RAC, le `misscount` il peut être nécessaire de modifier le paramètre. Pour plus d'informations sur les paramètres de délai d'expiration du RAC, reportez-vous à la section "Oracle RAC avec ONTAP".

Configuration à deux sites

Une configuration RAC étendue sur deux sites peut fournir des services de base de données actif-actif qui peuvent survivre à de nombreux scénarios d'incident, mais pas à tous, sans interruption.

Fichiers de vote RAC

La gestion du quorum doit être prise en compte lors du déploiement du RAC étendu sur MetroCluster. Oracle RAC dispose de deux mécanismes pour gérer le quorum : le battement de cœur du disque et le battement de cœur du réseau. La pulsation du disque surveille l'accès au stockage à l'aide des fichiers de vote. Dans le cas d'une configuration RAC à site unique, une ressource de vote unique suffit tant que le système de stockage sous-jacent offre des fonctionnalités haute disponibilité.

Dans les versions précédentes d'Oracle, les fichiers de vote étaient placés sur des périphériques de stockage physiques, mais dans les versions actuelles d'Oracle, les fichiers de vote sont stockés dans des groupes de disques ASM.

Oracle RAC est pris en charge par NFS. Pendant le processus d'installation de la grille, un ensemble de processus ASM est créé pour présenter l'emplacement NFS utilisé pour les fichiers de grille en tant que groupe de disques ASM. Le processus est presque transparent pour l'utilisateur final et ne nécessite aucune gestion ASM continue une fois l'installation terminée.

Dans une configuration à deux sites, il est tout d'abord nécessaire de s'assurer que chaque site peut toujours accéder à plus de la moitié des fichiers de vote, ce qui garantit un processus de reprise après incident sans interruption. Cette tâche était simple avant que les fichiers de vote ne soient stockés dans des groupes de disques ASM, mais aujourd'hui, les administrateurs doivent comprendre les principes de base de la redondance ASM.

Les groupes de disques ASM disposent de trois options de redondance `external`, `normal`, et `high`. En d'autres termes, sans miroir, avec miroir et miroir à 3 voies. Une option plus récente appelée `Flex` est également disponible, mais rarement utilisé. Le niveau de redondance et le placement des périphériques redondants contrôlent ce qui se passe dans les scénarios de panne. Par exemple :

- Placer les fichiers de vote sur un `diskgroup` avec `external` la redondance des ressources garantit la suppression d'un site en cas de perte de la connectivité intersite.
- Placer les fichiers de vote sur un `diskgroup` avec `normal` La redondance avec un seul disque ASM par site garantit la suppression des nœuds sur les deux sites en cas de perte de la connectivité intersite, car aucun des sites ne possède un quorum majoritaire.
- Placer les fichiers de vote sur un `diskgroup` avec `high` la redondance avec deux disques sur un site et un seul disque sur l'autre site permet des opérations actif-actif lorsque les deux sites sont opérationnels et mutuellement accessibles. Toutefois, si le site à disque unique est isolé du réseau, ce site est supprimé.

Pulsation du réseau RAC

Le signal de présence du réseau RAC Oracle surveille l'accessibilité des nœuds sur l'interconnexion de cluster. Pour rester dans le cluster, un nœud doit pouvoir contacter plus de la moitié des autres nœuds. Dans une architecture à deux sites, cette exigence crée les choix suivants pour le nombre de nœuds RAC :

- Le placement d'un nombre égal de nœuds par site entraîne la suppression sur un site en cas de perte de la connectivité réseau.
- Le placement de N nœuds sur un site et de N+1 nœuds sur le site opposé garantit que la perte de la

connectivité intersite entraîne le site avec le plus grand nombre de nœuds restants dans le quorum du réseau et le site avec moins de nœuds supprimés.

Avant Oracle 12cR2, il était impossible de contrôler quel côté devait être expulsé en cas de perte du site. Lorsque chaque site a un nombre égal de nœuds, l'exclusion est contrôlée par le nœud maître, qui est en général le premier nœud RAC à démarrer.

Oracle 12cR2 introduit la fonctionnalité de pondération des nœuds. L'administrateur peut ainsi mieux contrôler la manière dont Oracle résout les problèmes de partage du cerveau. À titre d'exemple simple, la commande suivante définit les préférences pour un nœud particulier dans un RAC :

```
[root@host-a ~]# /grid/bin/crsctl set server css_critical yes
CRS-4416: Server attribute 'CSS_CRITICAL' successfully changed. Restart
Oracle High Availability Services for new value to take effect.
```

Après le redémarrage d'Oracle High-Availability Services, la configuration se présente comme suit :

```
[root@host-a lib]# /grid/bin/crsctl status server -f | egrep
'^NAME|CSS_CRITICAL='
NAME=host-a
CSS_CRITICAL=yes
NAME=host-b
CSS_CRITICAL=no
```

Nœud `host-a` est maintenant désigné comme serveur critique. Si les deux nœuds RAC sont isolés, `host-a` survit, et `host-b` est supprimé.

Pour plus d'informations, consultez le livre blanc Oracle « Oracle Clusterware 12c Release 2 Technical Overview. »

Pour les versions d'Oracle RAC antérieures à 12cR2, le nœud maître peut être identifié en vérifiant les journaux CRS comme suit :

```
[root@host-a ~]# /grid/bin/crsctl status server -f | egrep
'^NAME|CSS_CRITICAL='
NAME=host-a
CSS_CRITICAL=yes
NAME=host-b
CSS_CRITICAL=no
 [root@host-a ~]# grep -i 'master node' /grid/diag/crs/host-
a/crs/trace/crsd.trc
2017-05-04 04:46:12.261525 :   CRSSE:2130671360: {1:16377:2} Master Change
Event; New Master Node ID:1 This Node's ID:1
2017-05-04 05:01:24.979716 :   CRSSE:2031576832: {1:13237:2} Master Change
Event; New Master Node ID:2 This Node's ID:1
2017-05-04 05:11:22.995707 :   CRSSE:2031576832: {1:13237:221} Master
Change Event; New Master Node ID:1 This Node's ID:1
2017-05-04 05:28:25.797860 :   CRSSE:3336529664: {1:8557:2} Master Change
Event; New Master Node ID:2 This Node's ID:1
```

Ce journal indique que le nœud maître est `2` et le nœud `host-a` A un ID de `1`. Ce fait signifie que `host-a` n'est pas le nœud maître. L'identité du nœud maître peut être confirmée avec la commande `olsnodes -n`.

```
[root@host-a ~]# /grid/bin/olsnodes -n
host-a  1
host-b  2
```

Le nœud ayant l'ID de `2` est `host-b`, qui est le nœud maître. Dans une configuration avec un nombre égal de nœuds sur chaque site, le site avec `host-b` est le site qui survit si les deux ensembles perdent la connectivité réseau pour quelque raison que ce soit.

Il est possible que l'entrée de journal qui identifie le nœud maître puisse sortir du système. Dans ce cas, les horodatages des sauvegardes du registre des clusters Oracle (OCR) peuvent être utilisés.

```
[root@host-a ~]#  /grid/bin/ocrconfig -showbackup
host-b     2017/05/05 05:39:53     /grid/cdata/host-cluster/backup00.ocr
0
host-b     2017/05/05 01:39:53     /grid/cdata/host-cluster/backup01.ocr
0
host-b     2017/05/04 21:39:52     /grid/cdata/host-cluster/backup02.ocr
0
host-a     2017/05/04 02:05:36     /grid/cdata/host-cluster/day.ocr     0
host-a     2017/04/22 02:05:17     /grid/cdata/host-cluster/week.ocr     0
```

Cet exemple montre que le nœud maître est `host-b`. Il indique également un changement dans le nœud maître de `host-a` à `host-b` Quelque part entre 2:05 et 21:39 le 4 mai. Cette méthode d'identification du nœud maître n'est sûre que si les journaux CRS ont également été vérifiés car il est possible que le nœud

maître ait changé depuis la sauvegarde OCR précédente. Si ce changement s'est produit, il doit être visible dans les journaux OCR.

La plupart des clients choisissent un seul groupe de disques de vote qui dessert l'ensemble de l'environnement et un nombre égal de nœuds RAC sur chaque site. Le groupe de disques doit être placé sur le site qui contient la base de données. En conséquence, une perte de connectivité entraîne la suppression du site distant. Le site distant n'aurait plus le quorum, ni l'accès aux fichiers de base de données, mais le site local continue à fonctionner normalement. Une fois la connectivité rétablie, l'instance distante peut être de nouveau mise en ligne.

En cas d'incident, un basculement est nécessaire pour mettre en ligne les fichiers de base de données et le groupe de disques de vote sur le site survivant. Si l'incident permet à AUSO de déclencher le basculement, NVFAIL n'est pas déclenché, car le cluster est connu pour être synchronisé et les ressources de stockage sont normalement mises en ligne. L'AUSO est une opération très rapide et doit se terminer avant le `disktimeout` la période expire.

Comme il n'y a que deux sites, il n'est pas possible d'utiliser n'importe quel type de logiciel automatisé externe de rupture de tieBreaking, ce qui signifie que le basculement forcé doit être une opération manuelle.

Configurations à trois sites

Un cluster RAC étendu est beaucoup plus facile à concevoir avec trois sites. Les deux sites hébergeant chaque moitié du système MetroCluster prennent également en charge les workloads de la base de données, tandis que le troisième sert de disjoncteur pour la base de données et le système MetroCluster. La configuration Oracle Tiebreaker peut être aussi simple que le placement d'un membre du groupe de disques ASM utilisé pour le vote sur un troisième site, et peut également inclure une instance opérationnelle sur le troisième site pour s'assurer qu'il y a un nombre impair de nœuds dans le cluster RAC.

Consultez la documentation Oracle sur « quorum failure group » pour obtenir des informations importantes sur l'utilisation de NFS dans une configuration RAC étendue. En résumé, il peut être nécessaire de modifier les options de montage NFS pour inclure l'option logicielle permettant de s'assurer que la perte de connectivité au troisième site hébergeant les ressources quorum n'affecte pas les serveurs Oracle ou les processus RAC Oracle principaux.

Informations sur le copyright

Informations sur les marques commerciales