Reprise sur incident Oracle

Enterprise applications

NetApp
February 10, 2026

Sommaire

Reprise sur incident Oracle

Présentation

La reprise d'activité consiste à restaurer les services de données après une catastrophe, par exemple un incendie qui détruit un système de stockage, voire un site entier.

Cette documentation remplace les rapports techniques *TR-4591 : Oracle Data protection* et *TR-4592 : Oracle on MetroCluster.*

La reprise après incident peut être effectuée par une simple réplication des données à l'aide de SnapMirror, bien sûr, lorsque de nombreux clients mettent à jour les réplicas en miroir toutes les heures.

Pour la plupart des clients, la reprise après incident ne suffit pas à posséder une copie distante des données. Il est donc nécessaire de pouvoir les exploiter rapidement. NetApp propose deux technologies pour répondre à ce besoin : MetroCluster et SnapMirror Active Sync

MetroCluster fait référence à ONTAP dans une configuration matérielle qui inclut un stockage en miroir synchrone de faible niveau et de nombreuses fonctionnalités supplémentaires. Les solutions intégrées telles que MetroCluster simplifient les bases de données, les applications et les infrastructures de virtualisation complexes et évolutives. Elle remplace plusieurs produits et stratégies externes de protection des données par une seule baie de stockage centrale simple. Elle offre également des fonctionnalités intégrées de sauvegarde, de restauration, de reprise après incident et de haute disponibilité au sein d'un seul système de stockage en cluster.

La synchronisation active SnapMirror (SM-AS) est basée sur la synchronisation SnapMirror synchrone. Avec MetroCluster, chaque contrôleur ONTAP est responsable de la réplication des données de son disque vers un emplacement distant. Avec la synchronisation active SnapMirror, deux systèmes ONTAP différents conservent des copies indépendantes de vos données LUN, mais fonctionnent ensemble pour présenter une seule instance de ce LUN. Du point de vue de l'hôte, il s'agit d'une entité LUN unique.

Comparaison SM-AS et MCC

Si les solutions SM-AS et MetroCluster sont similaires en termes de fonctionnalité globale, elles présentent d'importantes différences dans la mise en œuvre de la réplication avec un objectif de point de récupération de 0 et sa gestion. Les modes asynchrone et synchrone de SnapMirror peuvent également être utilisés dans le cadre d'un plan de reprise d'activité, mais ils ne sont pas conçus pour être utilisés en tant que technologies de réplication haute disponibilité.

- Une configuration MetroCluster ressemble davantage à un cluster intégré avec des nœuds distribués sur plusieurs sites. SM-AS se comporte comme deux clusters indépendants qui coopèrent pour fournir des LUN répliquées synchrones avec RPO=0 sélectionnés.
- Les données d'une configuration MetroCluster ne sont accessibles qu'à partir d'un site particulier à la fois. Une deuxième copie des données est présente sur le site opposé, mais les données sont passives. Il est impossible d'y accéder sans un basculement du système de stockage.
- La mise en miroir des systèmes MetroCluster et SM-AS effectue des opérations à différents niveaux. La mise en miroir MetroCluster s'effectue au niveau de la couche RAID. Les données de bas niveau sont stockées dans un format miroir à l'aide de SyncMirror. L'utilisation de la mise en miroir est pratiquement invisible au niveau des couches LUN, volume et protocole.
- En revanche, la mise en miroir SM-AS se produit au niveau de la couche de protocole. Les deux clusters sont globalement indépendants. Une fois les deux copies de données synchronisées, les deux clusters

n’ont besoin que de mettre en miroir les écritures. Lorsqu’une écriture a lieu sur un cluster, elle est répliquée sur l’autre. L’écriture est uniquement validée par l’hôte lorsque l’écriture est terminée sur les deux sites. En dehors de ce comportement de fractionnement de protocole, les deux clusters sont des clusters ONTAP normaux.

- Le rôle principal de MetroCluster est la réplication à grande échelle. Vous pouvez répliquer une baie complète avec un objectif de point de récupération RPO=0 et un objectif de durée de restauration proche de zéro. Le processus de basculement est ainsi simplifié, car il n’y a qu’une seule « chose » à basculer et il offre une excellente évolutivité en termes de capacité et d’IOPS.
- L’une des principales utilisations de SM-AS est la réplication granulaire. Parfois, vous ne souhaitez pas répliquer toutes les données en tant qu’unité unique ou vous devez pouvoir basculer sélectivement sur certains workloads.
- Autre cas d’utilisation clé de la solution SM-as pour les opérations actives/actives : vous souhaitez que des copies de données entièrement exploitables soient disponibles sur deux clusters différents situés à deux emplacements différents avec des performances identiques et, si vous le souhaitez, vous n’avez pas besoin d’étendre le SAN sur plusieurs sites. Vos applications peuvent déjà s’exécuter sur les deux sites, ce qui réduit le RTO global pendant les opérations de basculement.

MetroCluster

Reprise d’activité avec MetroCluster

MetroCluster est une fonctionnalité ONTAP qui protège vos bases de données Oracle avec une mise en miroir synchrone RPO=0 sur tous les sites. Elle peut évoluer jusqu’à prendre en charge des centaines de bases de données sur un seul système MetroCluster.

Il est également simple à utiliser. L’utilisation de MetroCluster n’ajoute pas nécessairement à ou ne modifie pas nécessairement les meilleures ractiques pour l’exploitation des applications et bases de données d’entreprise.

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.

Architecture physique

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

Le terme « connectivité » fait référence à la connexion au 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.

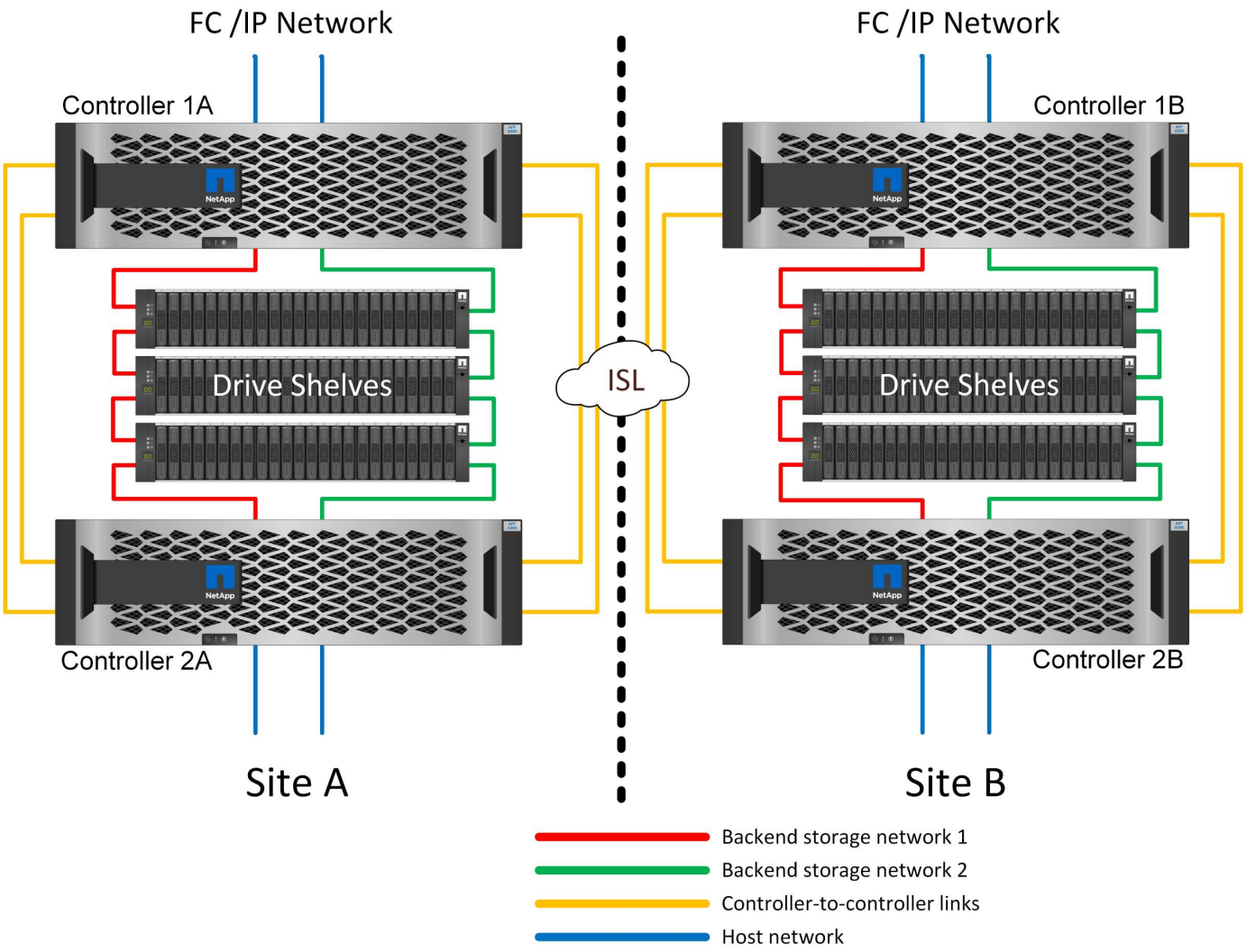
IP MetroCluster

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.

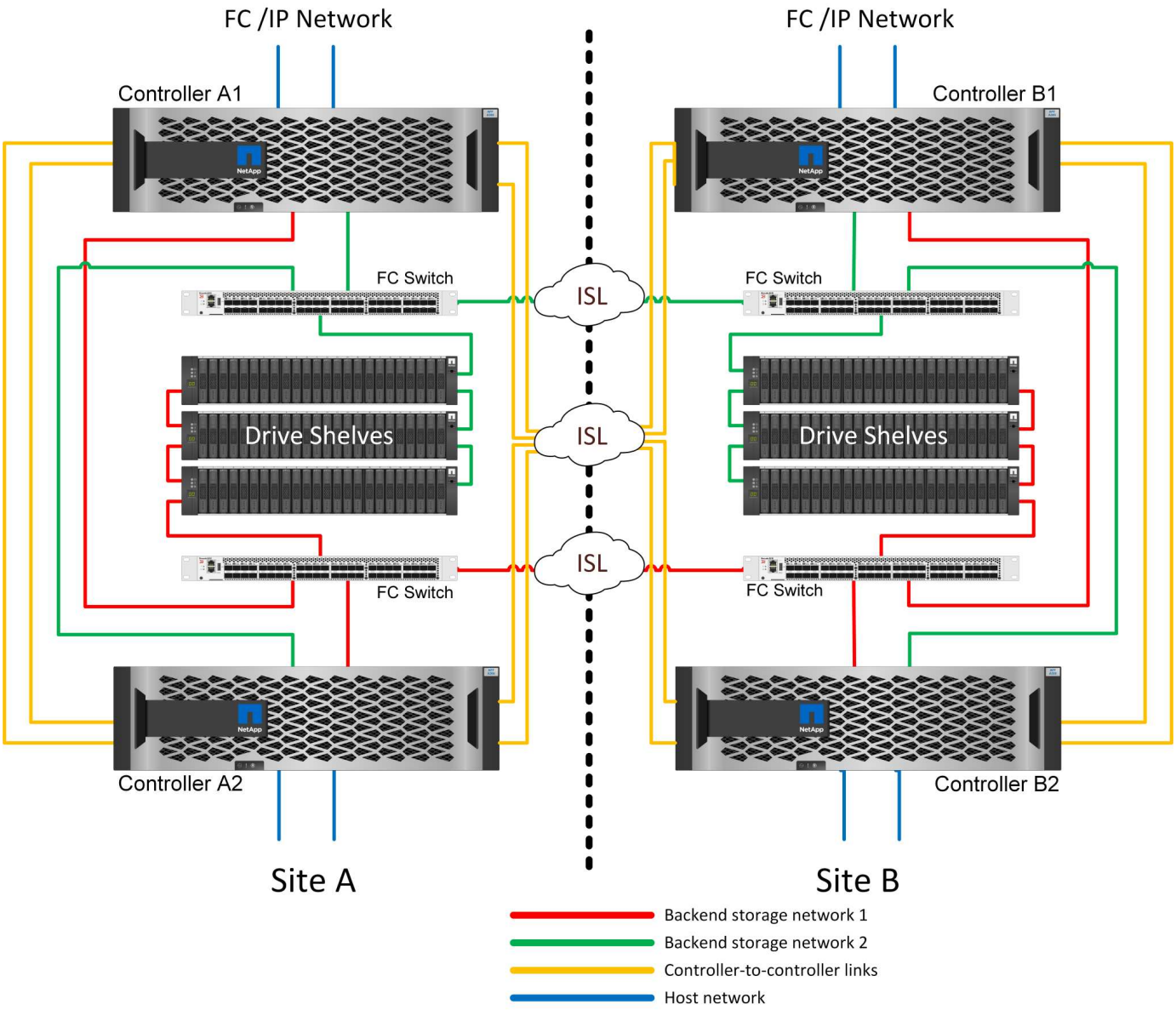
L'architecture est également plus simple, car les contrôleurs disposent des seules connexions entre les sites. Dans les MetroCluster FC, un contrôleur écrit directement sur les disques du site opposé et requiert ainsi des connexions SAN, des commutateurs et des ponts supplémentaires. En revanche, un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



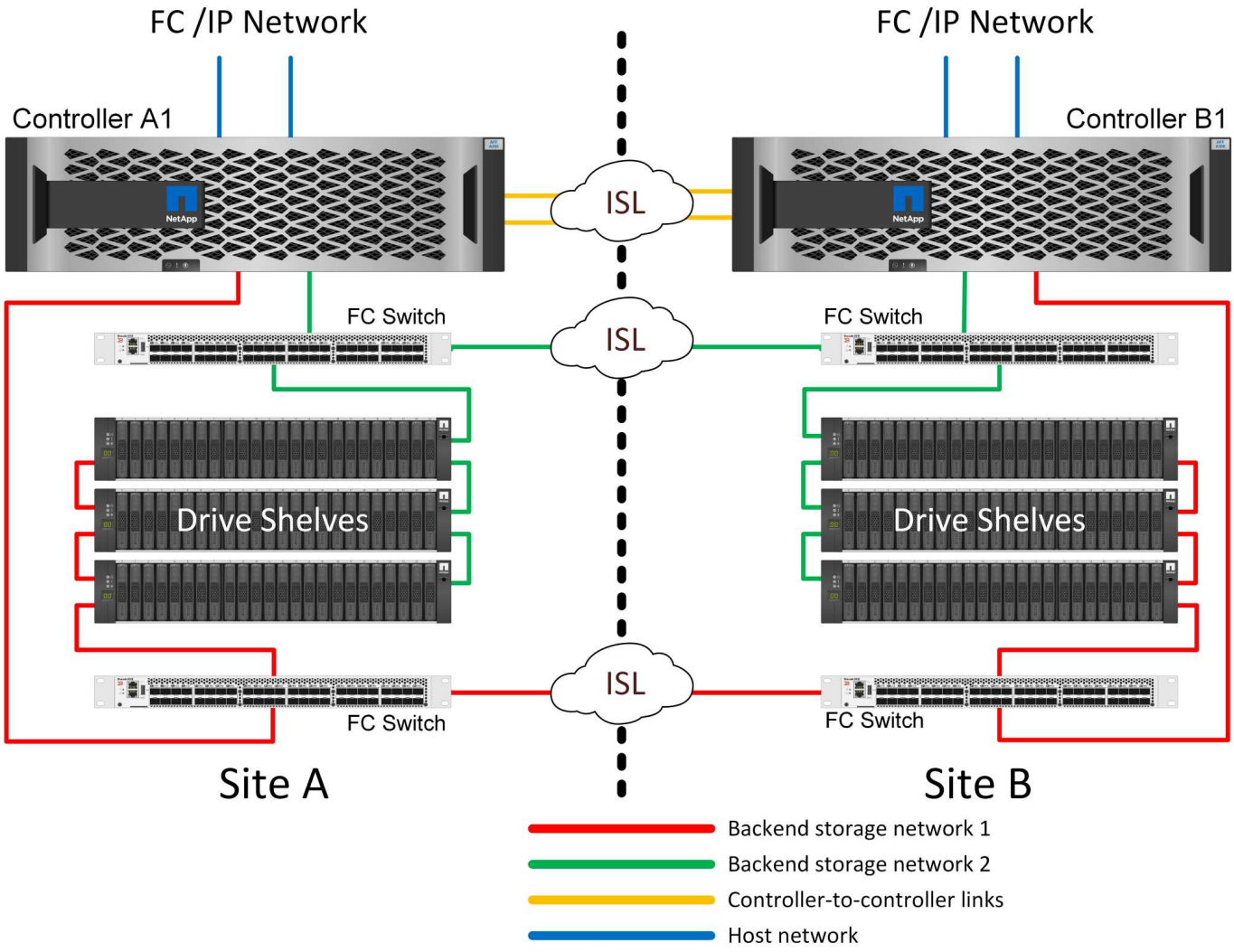
Certaines infrastructures multisites ne sont pas conçues pour les opérations en mode actif-actif. Elles sont plutôt utilisées comme site principal et site de reprise après incident. Dans ce cas, il est généralement préférable d'utiliser une option MetroCluster à paire HA pour les raisons suivantes :

- 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

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

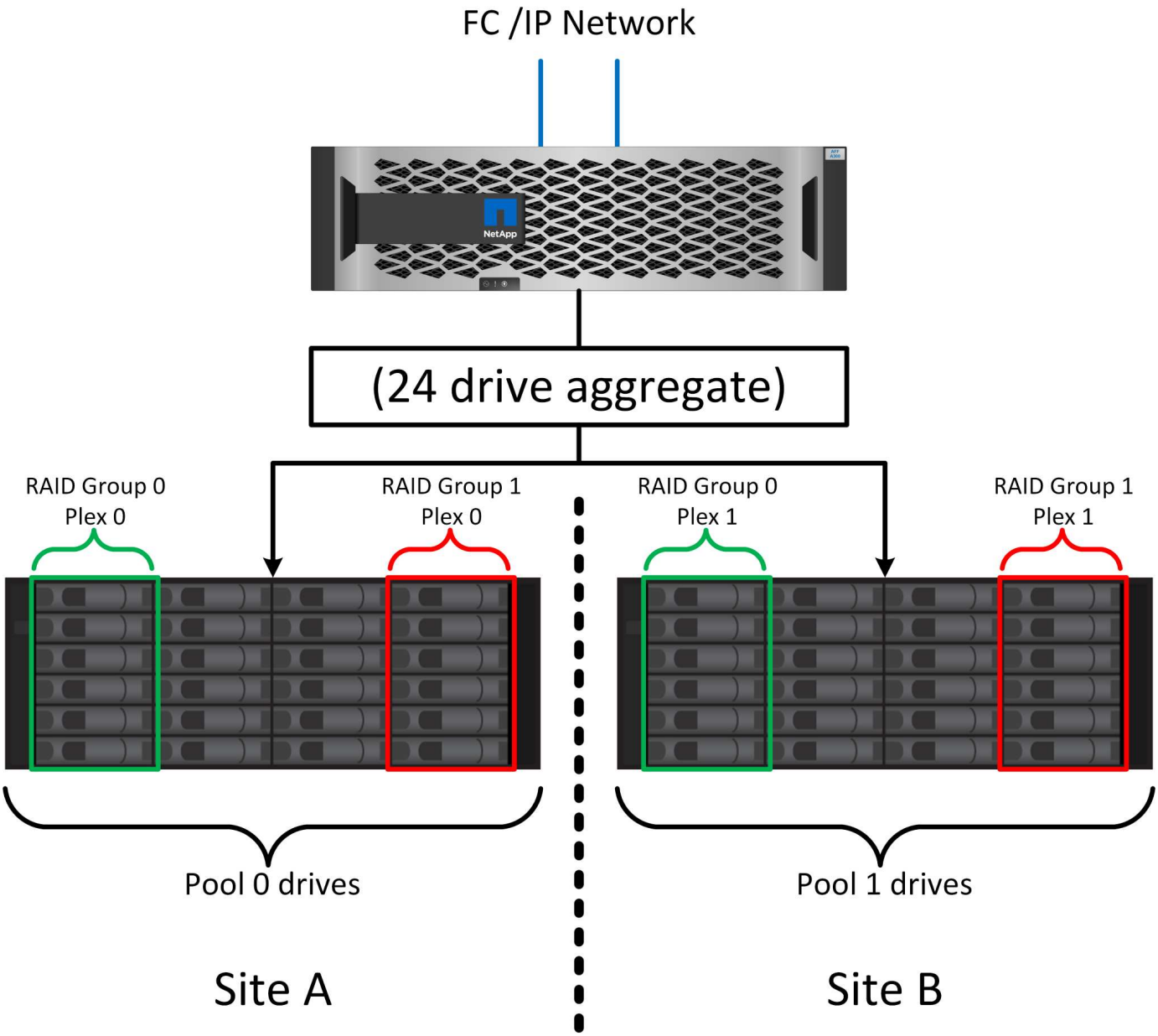
MetroCluster étend la protection des données NVRAM de plusieurs manières :

- 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

"NetApp MetroCluster Tiebreaker"Exécuté sur un troisième site, le logiciel peut contrôler l'état de santé de l'environnement MetroCluster, envoyer des notifications et forcer un basculement en cas d'incident. Une description complète du Tiebreaker se trouve sur le "Site de support NetApp", mais le but principal du Tiebreaker MetroCluster est de détecter la perte du 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.

Le logiciel MCTB ne vérifie pas que NVRAM était et/ou que les plexes sont synchronisés lorsqu'un basculement est forcé. 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

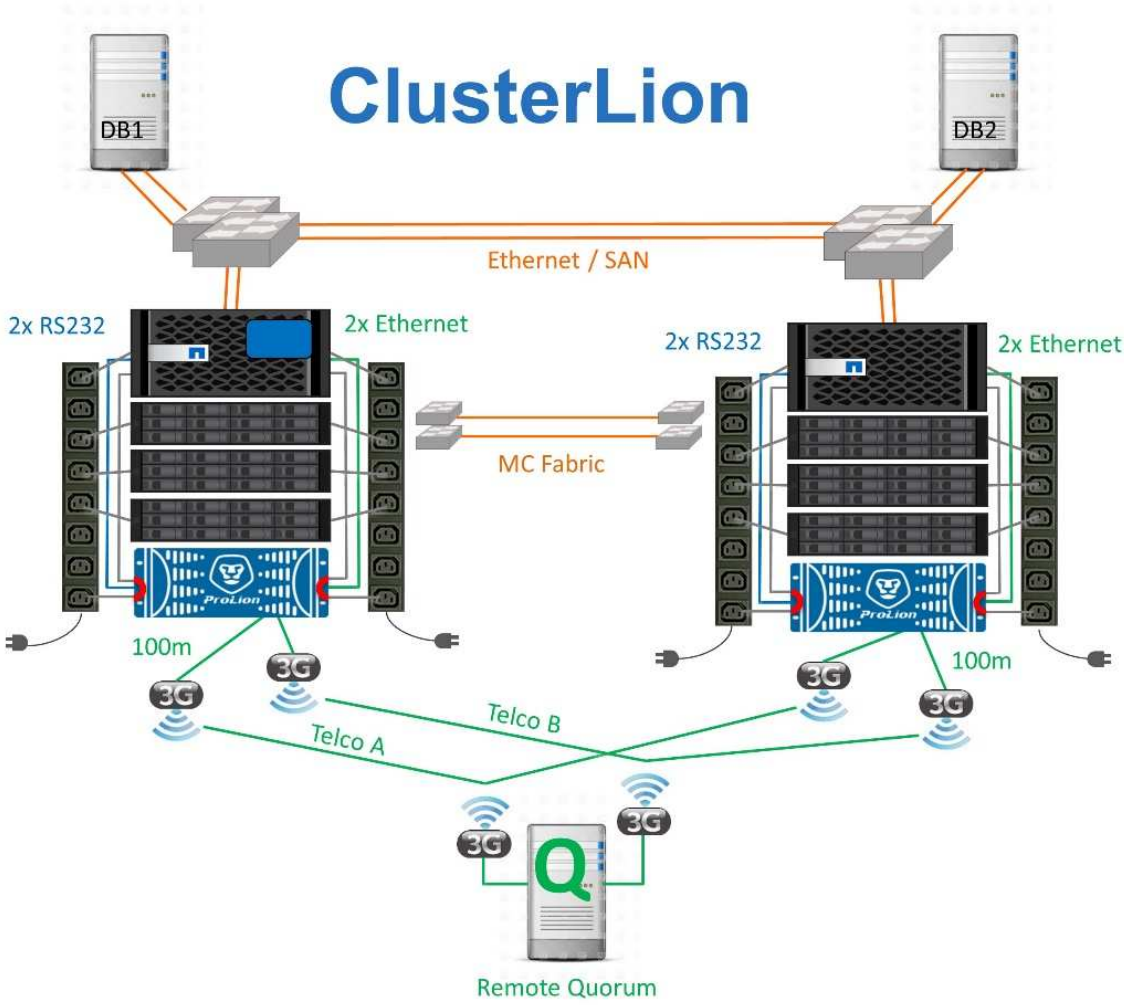
Le médiateur ONTAP est utilisé avec MetroCluster IP et certaines autres solutions ONTAP. Il fonctionne comme un service disjoncteur d'attache classique, tout comme le logiciel disjoncteur d'attache MetroCluster mentionné ci-dessus, mais comprend également une fonctionnalité essentielle, qui effectue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

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, comme Oracle DataGuard ou SQL Server Always On Availability Groups. La

mise en miroir des disques au niveau du système d'exploitation peut être optionnelle. Pour plus d'informations et d'options, consultez votre équipe de compte NetApp ou partenaire.

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

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.

RAC étendu Oracle

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 synchronisation active SnapMirror doit être considérée à la place de MetroCluster. La réplication SM-AS 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, il `misscount` 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.

Synchronisation active SnapMirror

Présentation

La synchronisation active SnapMirror vous permet de créer des environnements de base de données Oracle à ultra haute disponibilité où des LUN sont disponibles à partir de deux clusters de stockage différents.

Avec la synchronisation active SnapMirror, il n'y a pas de copie « principale » ni de copie « secondaire » des données. Chaque cluster peut fournir des E/S de lecture à partir de sa copie locale des données, et chaque cluster réplique une écriture vers son partenaire. Le résultat est un comportement d'E/S symétrique.

Entre autres options, vous pouvez exécuter Oracle RAC en tant que cluster étendu avec des instances opérationnelles sur les deux sites. Vous pouvez également créer des clusters de bases de données actif-passif RPO=0, dans lesquels les bases de données à instance unique peuvent être déplacées entre les sites en cas de panne sur le site. Ce processus peut également être automatisé via des produits tels que Pacemaker ou VMware HA. Toutes ces options reposent sur la réplication synchrone gérée par SnapMirror Active Sync.

Réplication synchrone

En fonctionnement normal, la synchronisation active SnapMirror fournit en permanence une réplique synchrone avec un objectif de point de récupération de 0, à une exception près. Si les données ne peuvent pas être répliquées, ONTAP exige de répliquer les données et de reprendre le traitement des E/S sur un site pendant que les LUN de l'autre site sont mises hors ligne.

Matériel de stockage

Contrairement à d'autres solutions de reprise après incident du stockage, la synchronisation active SnapMirror offre une flexibilité asymétrique de la plateforme. Le matériel de chaque site n'a pas besoin d'être identique. Cette fonctionnalité vous permet d'ajuster la taille du matériel utilisé pour prendre en charge la synchronisation active SnapMirror. Le système de stockage distant peut être identique au site principal s'il doit prendre en charge une charge de travail de production complète, mais si un incident entraîne une réduction des E/S, un système plus petit sur le site distant peut être plus économique.

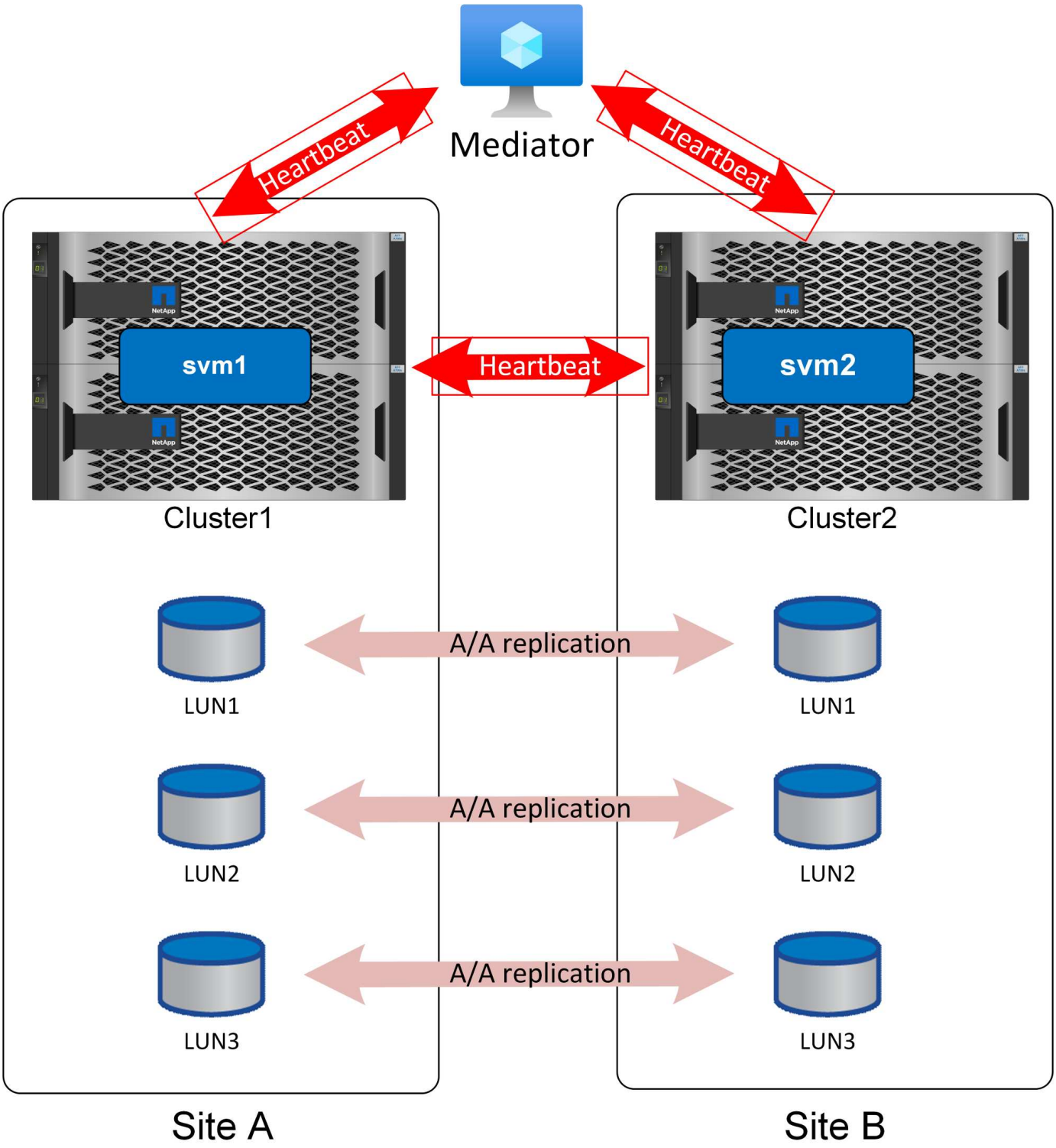
ONTAP Médiateur

Le médiateur ONTAP est une application logicielle téléchargée depuis la prise en charge de NetApp et généralement déployée sur une petite machine virtuelle. Le Mediator ONTAP n'est pas un Tiebreaker lorsqu'il est utilisé avec la synchronisation active SnapMirror. Il s'agit d'un canal de communication alternatif pour les deux clusters qui participent à la réplication SnapMirror active Sync. Les opérations automatisées sont dirigées par ONTAP sur la base des réponses reçues du partenaire via des relations directes et via le médiateur.

Médiateur de ONTAP

Le médiateur est requis pour automatiser le basculement en toute sécurité. Dans l'idéal, elle serait placée sur un site tiers indépendant, mais elle peut toujours fonctionner pour la plupart des besoins si elle est en colocation avec l'un des clusters participant à la réplication.

Le médiateur n'est pas vraiment un arbitre, même si c'est effectivement la fonction qu'il remplit. Le médiateur aide à déterminer l'état des nœuds du cluster et facilite le processus de basculement automatique en cas de panne d'un site. Le médiateur ne transfère aucune donnée, quelles que soient les circonstances.



Le principal défi lié au basculement automatisé est le problème des réseaux partagés, qui se pose en cas de perte de connectivité entre les deux sites. Que doit-on faire ? Vous ne voulez pas que deux sites différents se désignent comme les copies restantes des données, mais comment un seul site peut-il faire la différence entre la perte réelle du site opposé et l'incapacité à communiquer avec le site opposé ?

C'est là que le médiateur entre dans la photo. S'il est placé sur un troisième site, et chaque site a une connexion réseau distincte à ce site, alors vous avez un chemin supplémentaire pour chaque site pour valider l'état de santé de l'autre. Examinez à nouveau l'image ci-dessus et examinez les scénarios suivants.

- Que se passe-t-il si le médiateur échoue ou est inaccessible à partir d'un ou des deux sites ?
  - Les deux clusters peuvent toujours communiquer entre eux sur le même lien que celui utilisé pour les services de réplication.
  - Les données restent protégées avec un objectif de point de récupération de 0
- Que se passe-t-il si le site A tombe en panne ?
  - Le site B verra les deux canaux de communication tomber en panne.
  - Le site B prendra le contrôle des services de données, mais sans mise en miroir RPO=0
- Que se passe-t-il si le site B tombe en panne ?
  - Le site A verra les deux canaux de communication tomber en panne.
  - Le site A prend le relais des services de données, mais sans mise en miroir avec un objectif de point de récupération de 0

Il y a un autre scénario à prendre en compte : la perte du lien de réplication des données. En cas de perte de la liaison de réplication entre les sites, la mise en miroir avec un objectif de point de récupération de 0 sera évidemment impossible. Que devrait-on alors se passer ?

Ceci est contrôlé par le statut du site préféré. Dans une relation SM-AS, l'un des sites est secondaire à l'autre. Cela n'a aucun effet sur les opérations normales, et tout accès aux données est symétrique. Toutefois, si la réplication est interrompue, le nœud devra être rompu pour reprendre les opérations. Par conséquent, le site privilégié continuera les opérations sans mise en miroir et le site secondaire arrêtera le traitement des E/S jusqu'à ce que la communication de réplication soit restaurée.

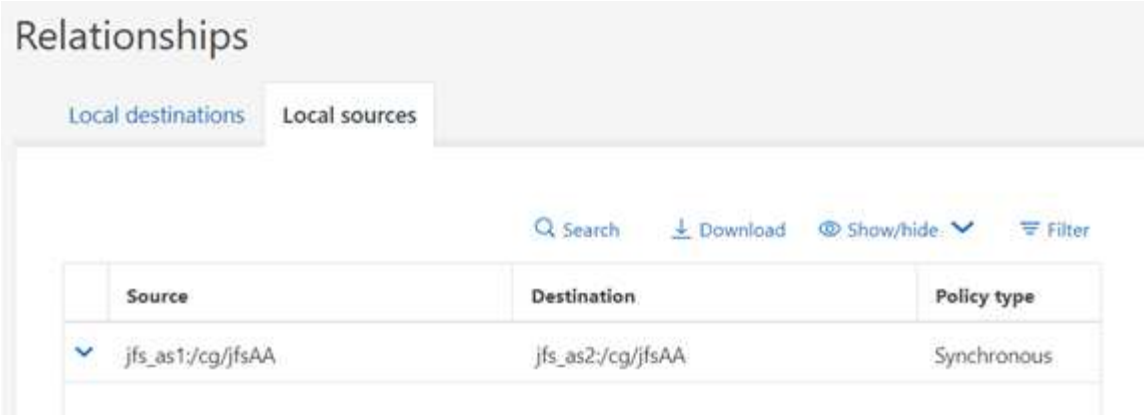
Site préféré de la synchronisation active SnapMirror

Le comportement de la synchronisation active SnapMirror est symétrique, avec une exception importante : la configuration du site préféré.

La synchronisation active SnapMirror considère un site comme la « source » et l'autre comme la « destination ». Cela implique une relation de réplication unidirectionnelle, mais cela ne s'applique pas au comportement d'E/S. La réplication est bidirectionnelle et symétrique. Les temps de réponse d'E/S sont identiques de part et d'autre du miroir.

La `source` désignation est le contrôle du site préféré. En cas de perte du lien de réplication, les chemins de LUN sur la copie source continueront à transmettre des données tandis que les chemins de LUN sur la copie de destination deviendront indisponibles jusqu'à ce que la réplication soit rétablie et que SnapMirror repasse à l'état synchrone. Les chemins reprennent alors le service des données.

La configuration source/destination peut être affichée via SystemManager :



Ou sur l'interface de ligne de commande :

```
Cluster2::> snapmirror show -destination-path jfs_as2:/cg/jfsAA

                           Source Path: jfs_as1:/cg/jfsAA
                        Destination Path: jfs_as2:/cg/jfsAA
                       Relationship Type: XDP
                 Relationship Group Type: consistencygroup
                     SnapMirror Schedule: -
                  SnapMirror Policy Type: automated-failover-duplex
                       SnapMirror Policy: AutomatedFailOverDuplex
                             Tries Limit: -
                           Throttle (KB/sec): -
                                 Mirror State: Snapmirrored
                          Relationship Status: InSync
```

La clé est que la source est le SVM sur le cluster1. Comme mentionné ci-dessus, les termes « source » et « destination » ne décrivent pas le flux des données répliquées. Les deux sites peuvent traiter une écriture et la répliquer sur le site opposé. En effet, les deux grappes sont des sources et des destinations. La désignation d'un cluster comme source contrôle simplement le cluster qui survit en tant que système de stockage en lecture/écriture en cas de perte du lien de réplication.

Topologie réseau

Accès uniforme

Un réseau d'accès uniforme signifie que les hôtes peuvent accéder aux chemins sur les deux sites (ou domaines de défaillance au sein du même site).

L'une des caractéristiques importantes de SM-AS est la capacité de configurer les systèmes de stockage pour savoir où se trouvent les hôtes. Lorsque vous mappez les LUN sur un hôte donné, vous pouvez indiquer si elles sont proximales ou non à un système de stockage donné.

Paramètres de proximité

La proximité fait référence à une configuration par cluster qui indique qu'un WWN d'hôte ou un ID d'initiateur iSCSI appartient à un hôte local. Il s'agit d'une deuxième étape facultative de configuration de l'accès aux

LUN.

La première étape correspond à la configuration habituelle du groupe initiateur. Chaque LUN doit être mappée sur un groupe initiateur qui contient les ID WWN/iSCSI des hôtes devant accéder à cette LUN. Cela contrôle quel hôte a *accès* à un LUN.

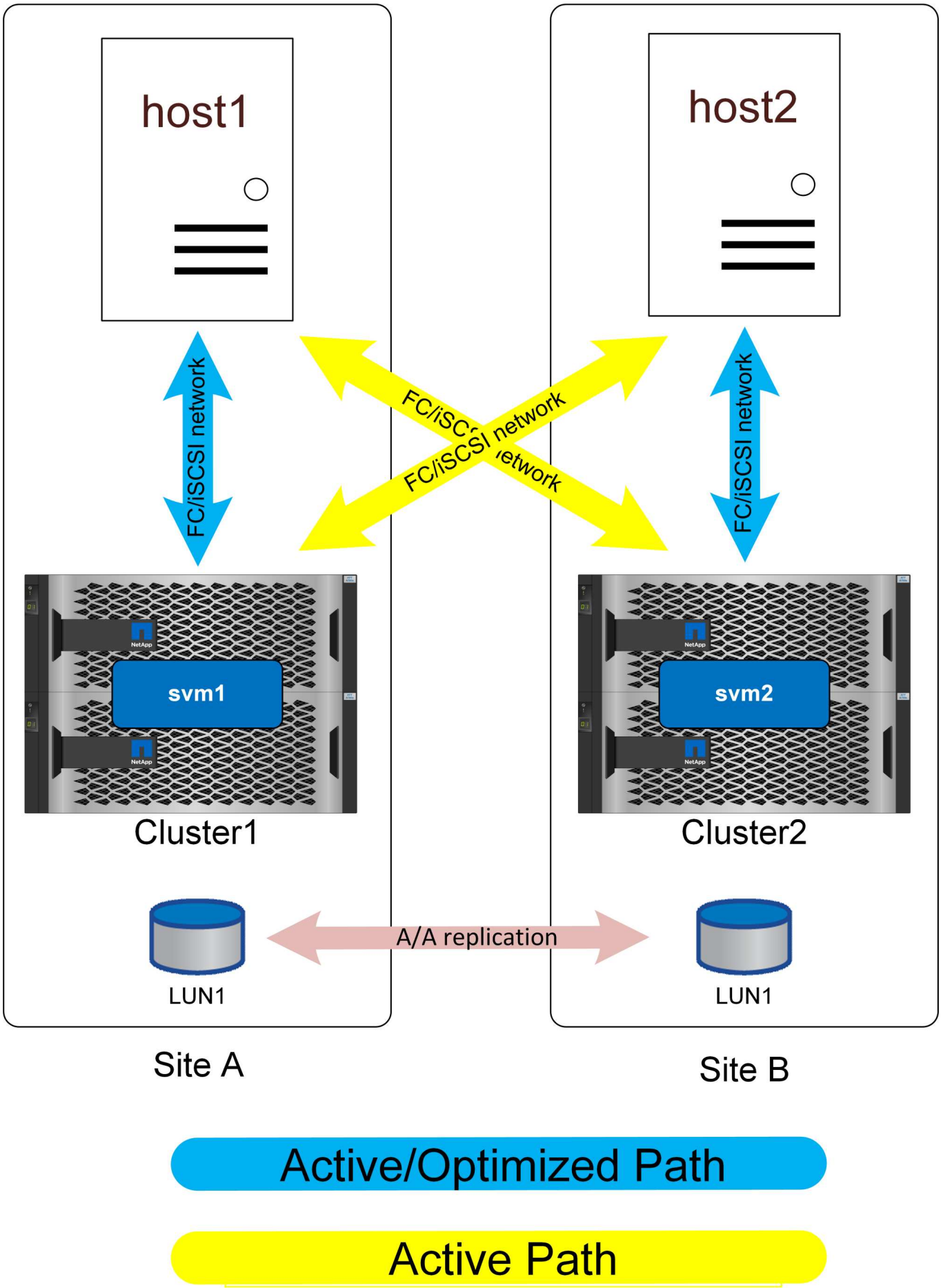
La deuxième étape facultative consiste à configurer la proximité de l'hôte. Cela ne contrôle pas l'accès, il contrôle *Priority*.

Par exemple, un hôte du site A peut être configuré pour accéder à une LUN protégée par la synchronisation active SnapMirror. Le SAN étant étendu entre les sites, les chemins d'accès sont disponibles pour cette LUN via le stockage sur le site A ou le stockage sur le site B.

Sans paramètres de proximité, cet hôte utilisera les deux systèmes de stockage de la même manière, car les deux systèmes de stockage annonceront des chemins actifs/optimisés. Si la latence SAN et/ou la bande passante entre les sites est limitée, il se peut que cela ne soit pas désirable, et vous pouvez vous assurer que, pendant le fonctionnement normal, chaque hôte utilise de préférence des chemins vers le système de stockage local. Cette configuration s'effectue en ajoutant l'ID WWN/iSCSI de l'hôte au cluster local en tant qu'hôte proximal. Cette opération peut être effectuée à partir de l'interface de ligne de commande ou de SystemManager.

AFF

Avec un système AFF, les chemins apparaissent comme indiqué ci-dessous lorsque la proximité de l'hôte a été configurée.


host1
host2
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
svm1
svm2
Cluster1
Cluster2
A/A replication
LUN1
LUN1
Site A
Site B
Active/Optimized Path
Active Path

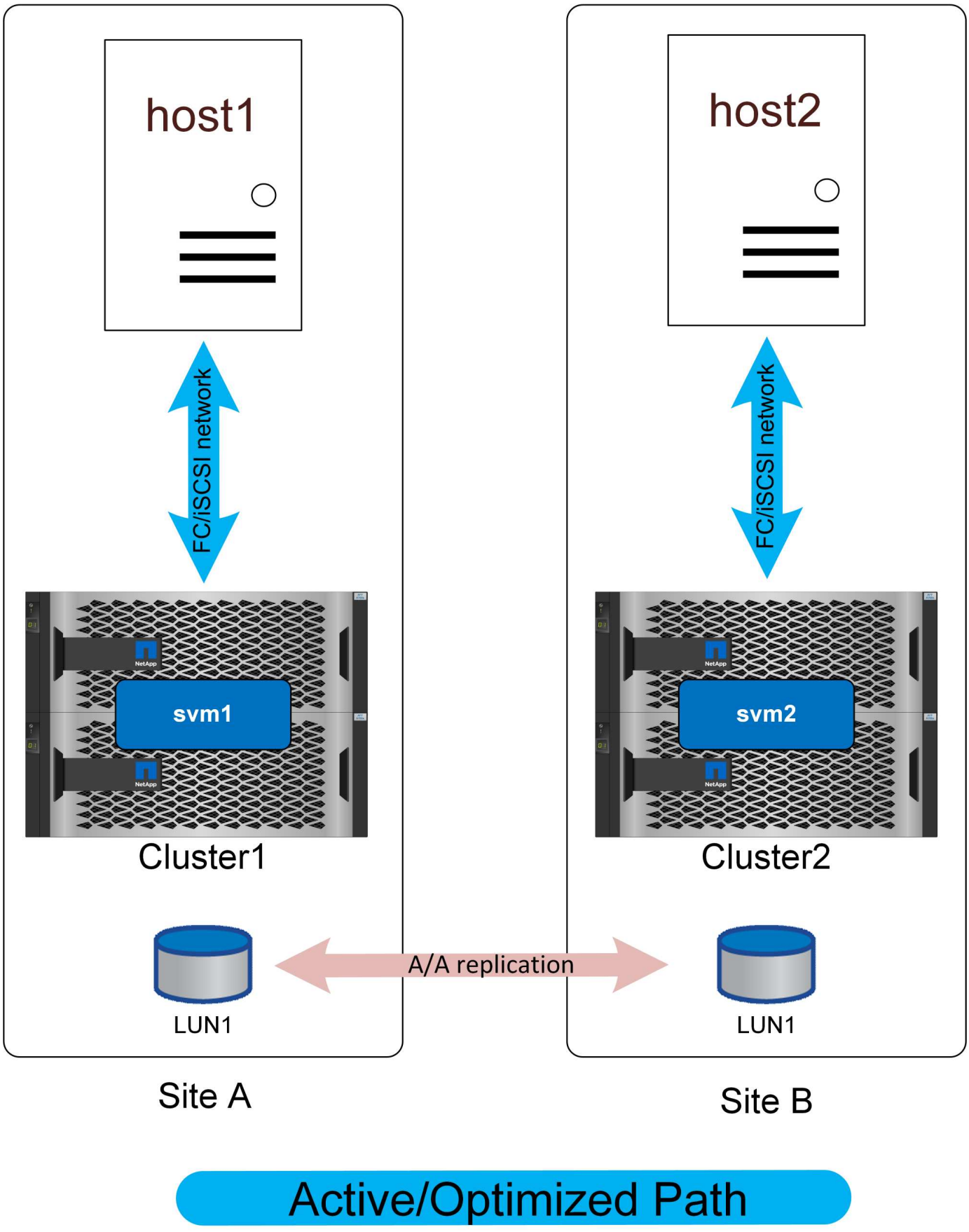
En fonctionnement normal, toutes les E/S sont des E/S locales. Les opérations de lecture et d'écriture sont gérées à partir de la baie de stockage locale. Bien entendu, les E/S en écriture devront également être répliquées par le contrôleur local sur le système distant avant d'être acquittées, mais toutes les E/S en lecture seront gérées localement et ne subiront pas de latence supplémentaire en traversant la liaison SAN entre les sites.

Le seul moment où les chemins non optimisés seront utilisés est la perte de tous les chemins actifs/optimisés. Par exemple, si l'ensemble de la baie sur le site A est hors tension, les hôtes du site A peuvent toujours accéder aux chemins d'accès à la baie sur le site B et donc rester opérationnels, même s'ils connaissent une latence plus élevée.

Il existe des chemins redondants à travers le cluster local qui ne sont pas illustrés sur ces schémas pour plus de simplicité. Les systèmes de stockage ONTAP étant dotés de la haute disponibilité, une panne du contrôleur ne devrait pas entraîner de panne sur le site. Il devrait simplement entraîner une modification dans laquelle les chemins locaux sont utilisés sur le site affecté.

ASA

Les systèmes NetApp ASA proposent des chemins d'accès multiples actif-actif sur tous les chemins d'accès à un cluster. Cela s'applique également aux configurations SM-AS.



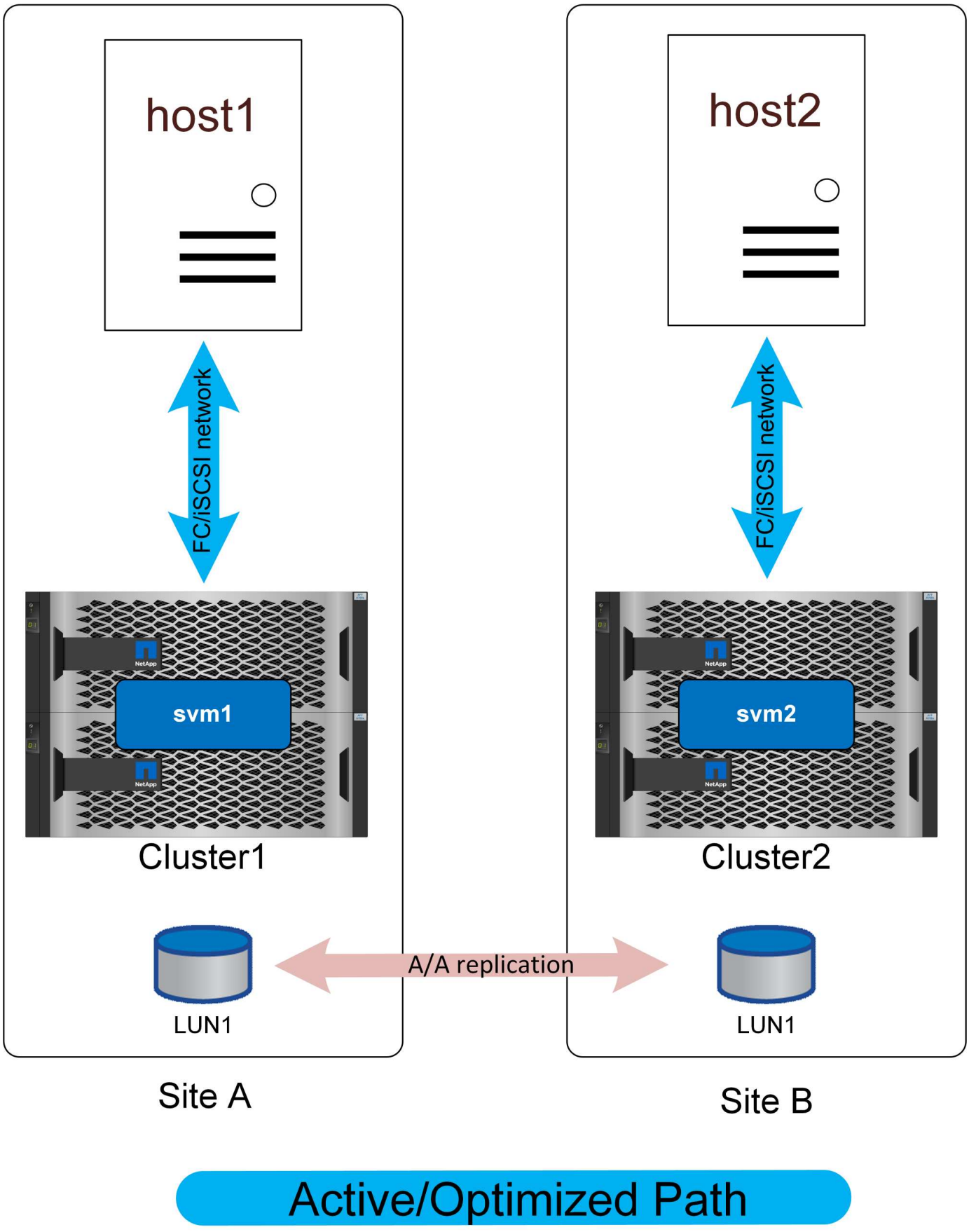
Une configuration ASA avec un accès non uniforme fonctionnera en grande partie comme avec AFF. Avec un accès uniforme, l'E/S traverserait le WAN. Cela peut être souhaitable ou non.

Si les deux sites étaient distants de 100 mètres avec une connectivité à fibre optique, il ne devrait pas y avoir de latence supplémentaire détectable traversant le WAN, mais si les sites étaient éloignés, les performances de lecture seraient affectées sur les deux sites. À l'inverse, avec AFF, ces chemins WAN seraient utilisés uniquement s'il n'existait aucun chemin local disponible et si les performances quotidiennes seraient meilleures, car toutes les E/S seraient des E/S locales. ASA avec un réseau d'accès non uniforme serait une option pour bénéficier des avantages de ASA en termes de coûts et de fonctionnalités sans engendrer de pénalités de latence entre les sites.

ASA avec SM-as dans une configuration à faible latence offre deux avantages intéressants. Tout d'abord, elle double *les performances de n'importe quel hôte, car les E/S peuvent être traitées par deux fois plus de contrôleurs en utilisant deux fois plus de chemins. Ensuite, dans un environnement à site unique, elle offre une disponibilité extrême, car l'intégralité du système de stockage peut être perdue sans interrompre l'accès aux hôtes.

Accès non uniforme

La mise en réseau à accès non uniforme signifie que chaque hôte n'a accès qu'aux ports du système de stockage local. Le SAN n'est pas étendu sur les sites (ou les domaines de défaillance au sein du même site).



Le principal avantage de cette approche est la simplicité du SAN : vous n'avez plus besoin d'étendre un SAN sur le réseau. Certains clients ne disposent pas d'une connectivité à faible latence suffisante entre les sites, ou

n’ont pas l’infrastructure nécessaire pour acheminer le trafic SAN FC sur un réseau intersite.

L’inconvénient de l’accès non uniforme est que certains scénarios de défaillance, notamment la perte du lien de réplication, entraînent la perte de l’accès au stockage par certains hôtes. En cas de perte de la connectivité du stockage local, les applications qui s’exécutent en tant qu’instances uniques, telles qu’une base de données non en cluster et qui ne s’exécute intrinsèquement que sur un hôte unique sur un montage donné, échouent. Les données seraient toujours protégées, mais le serveur de base de données n’aurait plus accès. Il doit être redémarré sur un site distant, de préférence par le biais d’un processus automatisé. Par exemple, VMware HA peut détecter une situation de tous les chemins d’accès sur un serveur et redémarrer une machine virtuelle sur un autre serveur sur lequel les chemins d’accès sont disponibles.

En revanche, une application en cluster telle qu’Oracle RAC peut fournir un service qui est disponible simultanément sur deux sites différents. La perte d’un site ne signifie pas la perte du service d’application dans son ensemble. Les instances restent disponibles et s’exécutent sur le site survivant.

Dans de nombreux cas, la surcharge liée à la latence supplémentaire qu’une application accède au système de stockage via une liaison site à site ne serait pas acceptable. Cela signifie que l’amélioration de la disponibilité des réseaux uniformes est minime, car la perte de stockage sur un site entraînerait la nécessité de fermer les services sur ce site défaillant.

> Il existe des chemins redondants à travers le cluster local qui ne sont pas illustrés sur ces schémas pour plus de simplicité. Les systèmes de stockage ONTAP étant dotés de la haute disponibilité, une panne du contrôleur ne devrait pas entraîner de panne sur le site. Il devrait simplement entraîner une modification dans laquelle les chemins locaux sont utilisés sur le site affecté.

Configurations Oracle

Présentation

L’utilisation de la synchronisation active SnapMirror n’ajoute pas nécessairement aux meilleures pratiques d’exploitation d’une base de données ou ne modifie pas nécessairement ces pratiques.

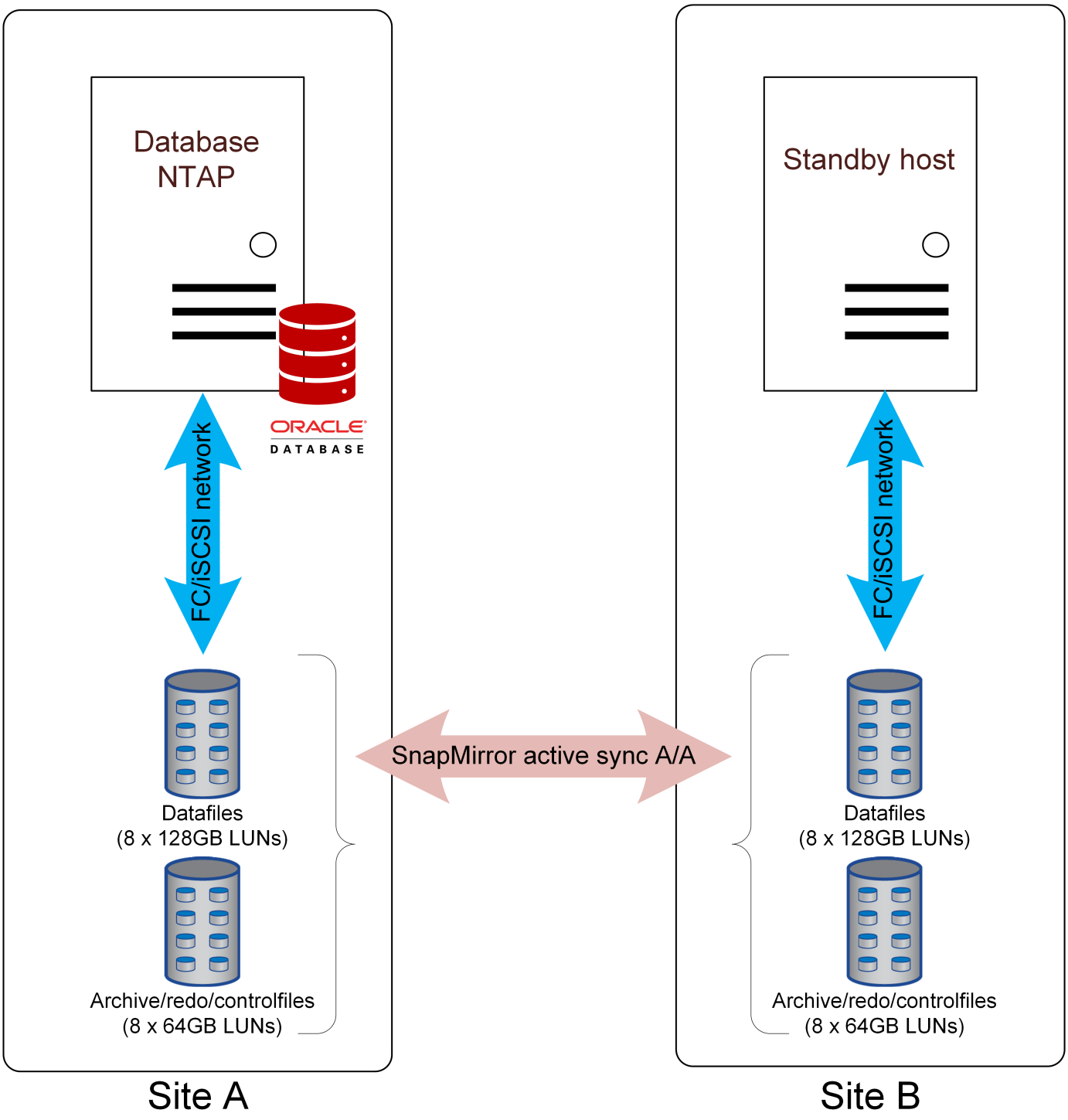
La meilleure architecture dépend des besoins de l’entreprise. Par exemple, si l’objectif est de bénéficier d’une protection RPO=0 contre la perte de données, mais que l’objectif RTO est assoupli, l’utilisation de bases de données Oracle Single instance et la réplication des LUN avec SM-AS peuvent suffire et être moins coûteuses d’un standard de licences Oracle. Toute panne du site distant n’interrompt pas les opérations, et la perte du site principal entraînerait la présence de LUN en ligne et prêtes à être utilisées sur le site survivant.

Si le RTO était plus strict, l’automatisation actif-passif de base via des scripts ou des clusters comme Pacemaker ou Ansible améliorerait le délai de basculement. Par exemple, VMware HA peut être configuré pour détecter une panne de VM sur le site principal et activer cette dernière sur le site distant.

Enfin, pour un basculement extrêmement rapide, Oracle RAC peut être déployé sur plusieurs sites. L’objectif de délai de restauration serait essentiellement égal à zéro, car la base de données serait en ligne et disponible à tout moment sur les deux sites.

Instance unique Oracle

Les exemples décrits ci-dessous illustrent certaines des nombreuses options de déploiement des bases de données Oracle Single instance avec la réplication SnapMirror Active Sync.



Basculement avec un système d'exploitation préconfiguré

La synchronisation active SnapMirror fournit une copie synchrone des données au niveau du site de reprise d'activité. Toutefois, 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 Pacemaker 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) mettra les bases de données en ligne sur le site secondaire. Une option consiste à créer des serveurs de secours préconfigurés pour les ressources 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. Détection d'une défaillance du site principal
2. Effectuez la détection des LUN FC ou iSCSI
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'une `mount` commande 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.

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. Détection d'une défaillance du site principal
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.

Protection contre les défaillances du stockage

Le diagramme ci-dessus montre l'utilisation de "accès non uniforme", où le SAN n'est pas étendu entre les sites. Cela peut être plus simple à configurer et, dans certains cas, peut être la seule option étant donné les fonctionnalités SAN actuelles, mais cela signifie également que la défaillance du système de stockage principal entraînerait une panne de la base de données jusqu'à ce que l'application ait été ratée.

Pour une résilience supplémentaire, la solution pourrait être déployée avec "accès uniforme". Cela permettrait aux applications de continuer à fonctionner en utilisant les chemins annoncés à partir du site opposé.

RAC étendu Oracle

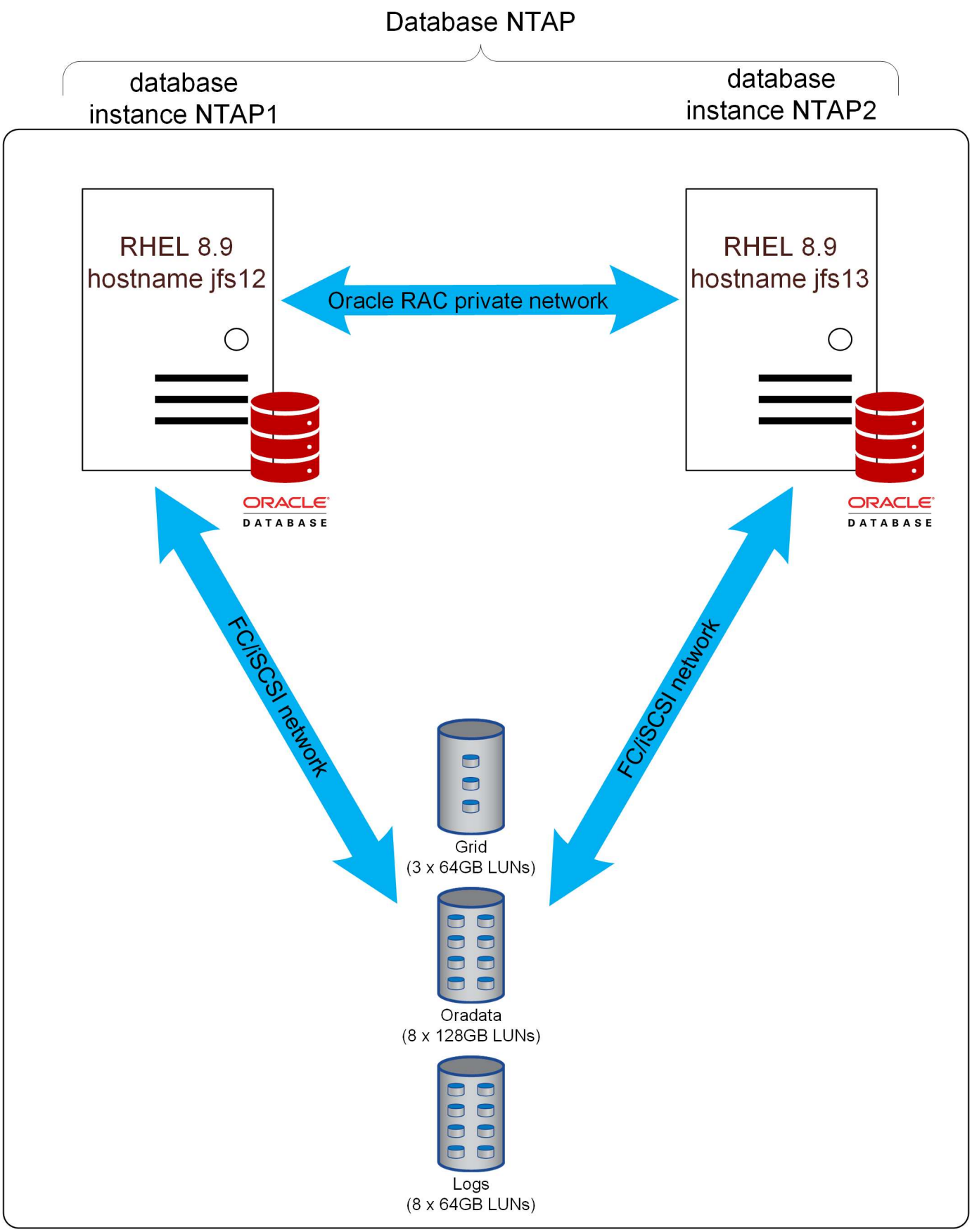
De nombreux clients optimisent leur RTO en étendant un cluster Oracle RAC sur plusieurs sites, offrant une configuration entièrement active/active. La conception globale devient plus complexe car elle doit inclure la gestion du quorum d'Oracle RAC.

La mise en cluster RAC étendue traditionnelle s'est appuyée sur la mise en miroir ASM pour assurer la protection des données. Cette approche fonctionne, mais elle implique également de nombreuses étapes manuelles de configuration et entraîne une surcharge de l'infrastructure réseau. À l'inverse, la réplication des

données peut être prise en charge par la synchronisation active SnapMirror, ce qui simplifie considérablement la solution. Les opérations telles que la synchronisation, la resynchronisation après les interruptions, les basculements et la gestion du quorum sont plus simples. En outre, le SAN n'a pas besoin d'être distribué entre les sites, ce qui simplifie la conception et la gestion du SAN.

La réplication

Pour comprendre la fonctionnalité RAC sur SnapMirror Active Sync, il est essentiel de considérer le stockage comme un ensemble unique de LUN hébergés sur un stockage en miroir. Par exemple :



Il n'y a pas de copie principale ou miroir. Pour schématiser, il n'y a qu'une seule copie de chaque LUN et cette LUN est disponible sur les chemins SAN situés sur deux systèmes de stockage différents. Du point de vue de l'hôte, il n'y a pas de basculement de stockage ; il y a des changements de chemin. Plusieurs défaillances

peuvent entraîner la perte de certains chemins vers la LUN, tandis que les autres chemins restent en ligne. La synchronisation active SnapMirror garantit la disponibilité des mêmes données sur tous les chemins opérationnels.

Configuration de stockage sous-jacente

Dans cet exemple de configuration, les disques ASM sont configurés de la même manière que dans n'importe quelle configuration RAC à site unique sur le stockage d'entreprise. Étant donné que le système de stockage assure la protection des données, la redondance ASM externe est utilisée.

Accès uniforme ou non informé

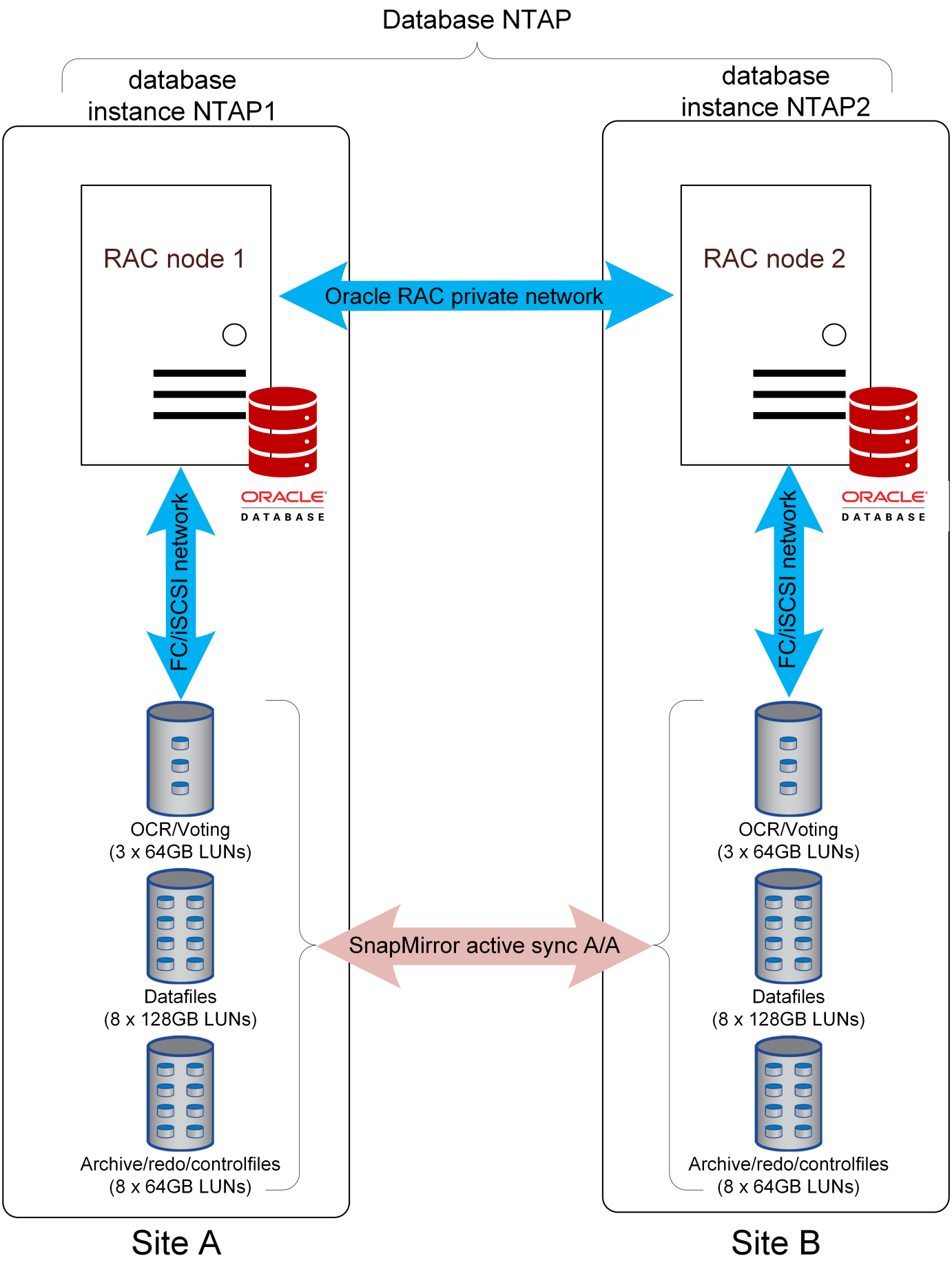
L'élément le plus important à prendre en compte avec Oracle RAC sur SnapMirror Active Sync est de savoir s'il faut utiliser un accès uniforme ou non.

Un accès uniforme signifie que chaque hôte peut voir les chemins sur les deux clusters. L'accès non uniforme signifie que les hôtes peuvent uniquement voir les chemins vers le cluster local.

Aucune de ces options n'est spécifiquement recommandée ou déconseillée. Certains clients ont facilement accès à la fibre noire pour connecter les sites, d'autres ne disposent pas d'une telle connectivité ou leur infrastructure SAN ne prend pas en charge l'ISL longue distance.

Accès non uniforme

L'accès non uniforme est plus simple à configurer du point de vue du SAN.


Database NTAP
database instance NTAP1
database instance NTAP2
RAC node 1
RAC node 2
Oracle RAC private network
ORACLE DATABASE
ORACLE DATABASE
FC/iSCSI network
FC/iSCSI network
OCR/Voting (3 x 64GB LUNs)
OCR/Voting (3 x 64GB LUNs)
SnapMirror active sync A/A
Datafiles (8 x 128GB LUNs)
Datafiles (8 x 128GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Site A
Site B

L’inconvénient principal de cette "accès non uniforme" approche est que la perte de la connectivité ONTAP site à site ou la perte d’un système de stockage entraînera la perte des instances de base de données sur un site. Cela n’est évidemment pas souhaitable, mais cela peut constituer un risque acceptable en échange d’une configuration SAN plus simple.

Accès uniforme

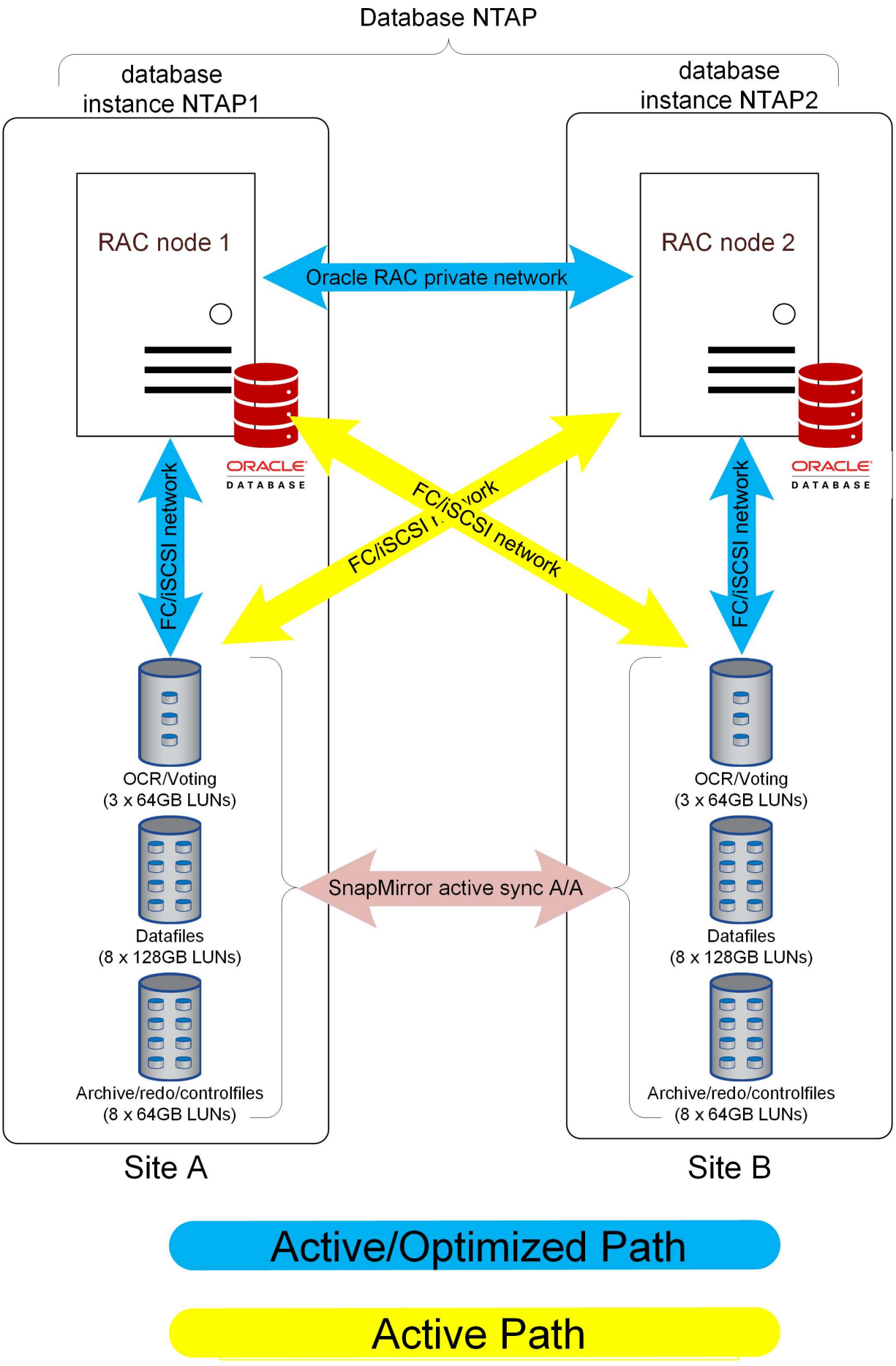
L’accès uniforme requiert l’extension du SAN sur les sites. Le principal avantage est que la perte d’un système de stockage n’entraîne pas la perte d’une instance de base de données. Au lieu de cela, cela entraînerait une modification des chemins d’accès multiples dans lesquels les chemins sont actuellement utilisés.

Il existe plusieurs façons de configurer l’accès non uniforme.

> Dans les schémas ci-dessous, il existe également des chemins actifs mais non optimisés qui seraient utilisés en cas de défaillances simples du contrôleur, mais ces chemins ne sont pas affichés dans l’intérêt de simplifier les diagrammes.

AFF avec paramètres de proximité

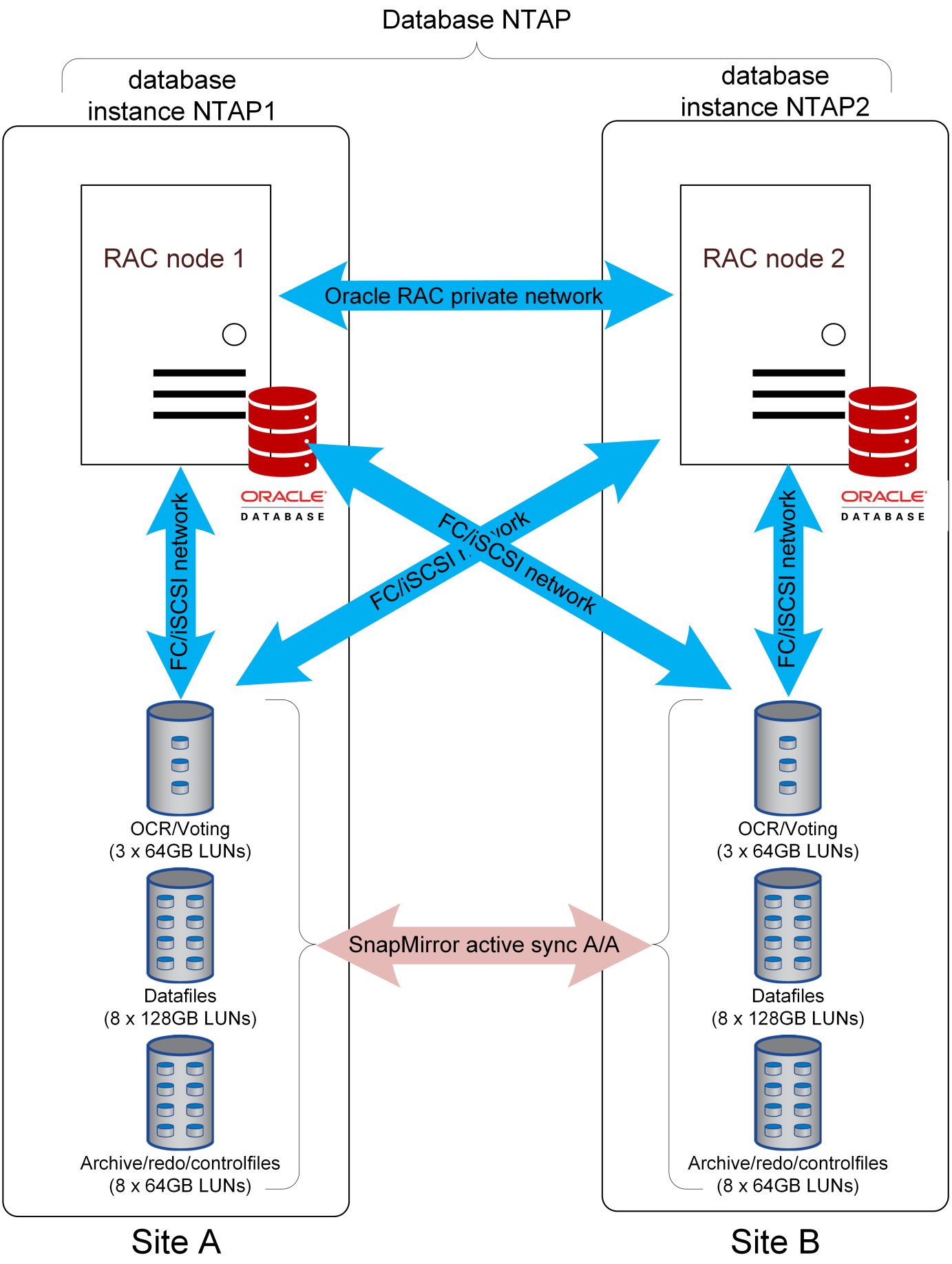
En cas de latence importante entre les sites, les systèmes AFF peuvent être configurés avec des paramètres de proximité des hôtes. Cela permet à chaque système de stockage d’identifier les hôtes locaux et distants, et d’attribuer les priorités de chemin en conséquence.


Database NTAP
database instance NTAP1
database instance NTAP2
RAC node 1
RAC node 2
Oracle RAC private network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
ORACLE DATABASE
ORACLE DATABASE
OCR/Voting (3 x 64GB LUNs)
OCR/Voting (3 x 64GB LUNs)
SnapMirror active sync A/A
Datafiles (8 x 128GB LUNs)
Datafiles (8 x 128GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Site A
Site B
Active/Optimized Path
Active Path

En fonctionnement normal, chaque instance Oracle utilisera de préférence les chemins locaux actifs/optimisés. Par conséquent, toutes les lectures seront traitées par la copie locale des blocs. La latence est ainsi la plus faible possible. Les E/S d'écriture sont envoyées de la même manière vers le contrôleur local. L'E/S doit toujours être répliquée avant d'être reconnue, ce qui entraîne toujours une latence supplémentaire en traversant le réseau site à site, mais cela ne peut pas être évité dans une solution de réplication synchrone.

ASA / AFF sans paramètres de proximité

S'il n'y a pas de latence significative entre les sites, les systèmes AFF peuvent être configurés sans paramètres de proximité des hôtes, ou ASA peut être utilisé.


Database NTAP
database instance NTAP1
database instance NTAP2
RAC node 1
RAC node 2
Oracle RAC private network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
ORACLE DATABASE
ORACLE DATABASE
OCR/Voting (3 x 64GB LUNs)
OCR/Voting (3 x 64GB LUNs)
SnapMirror active sync A/A
Datafiles (8 x 128GB LUNs)
Datafiles (8 x 128GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Site A
Site B

Chaque hôte pourra utiliser tous les chemins opérationnels sur les deux systèmes de stockage. Cela améliore considérablement les performances en permettant à chaque hôte d'exploiter le potentiel de performance de deux clusters, et non d'un seul.

Avec ASA, non seulement tous les chemins vers les deux clusters sont considérés comme actifs et optimisés, mais les chemins sur les contrôleurs partenaires sont également actifs. Il en résulte des chemins SAN entièrement actifs sur l'ensemble du cluster, à tout moment.

> Les systèmes ASA peuvent également être utilisés dans une configuration d'accès non uniforme. Étant donné qu'il n'existe aucun chemin entre les sites, les performances ne seraient pas améliorées par le franchissement de l'ISL par les E/S.

Disjoncteur d'attache RAC

Bien que le RAC étendu utilisant la synchronisation active SnapMirror soit une architecture symétrique par rapport aux E/S, il existe une exception qui est connectée à la gestion du split-brain.

Que se passe-t-il si le lien de réplication est perdu et qu'aucun des sites n'a le quorum ? Que doit-on faire ? Cette question s'applique à la fois au comportement d'Oracle RAC et de ONTAP. Si les modifications ne peuvent pas être répliquées sur tous les sites et que vous souhaitez reprendre les opérations, l'un des sites devra survivre et l'autre site devra être indisponible.

Le système "Médiateur de ONTAP" répond à cette exigence au niveau de la couche ONTAP. Il existe plusieurs options pour le tricover RAC.

Disjoncteurs Oracle

La meilleure méthode pour gérer les risques Oracle RAC split-brain consiste à utiliser un nombre impair de nœuds RAC, de préférence à l'aide d'un Tiebreaker 3rd site. Si un troisième site n'est pas disponible, l'instance Tiebreaker pourrait être placée sur un site des deux sites, ce qui la désignerait en fait un site de survivant préféré.

Oracle et CSS_Critical

Avec un nombre pair de nœuds, le comportement par défaut d'Oracle RAC est que l'un des nœuds du cluster sera considéré plus important que les autres nœuds. Le site avec ce nœud de priorité supérieure survivra à l'isolation du site tandis que les nœuds de l'autre site seront supprimés. La hiérarchisation est basée sur plusieurs facteurs, mais vous pouvez également contrôler ce comportement à l'aide du `css_critical` paramètre.

Dans l'"exemple"architecture, les noms d'hôte des nœuds RAC sont jfs12 et jfs13. Les paramètres actuels de `css_critical` sont les suivants :

```
[root@jfs12 ~]# /grid/bin/crsctl get server css_critical
CRS-5092: Current value of the server attribute CSS_CRITICAL is no.

[root@jfs13 trace]# /grid/bin/crsctl get server css_critical
CRS-5092: Current value of the server attribute CSS_CRITICAL is no.
```

Si vous voulez que le site avec jfs12 soit le site préféré, définissez cette valeur sur oui sur un site Un noeud et

redémarrez les services.

```
[root@jfs12 ~]# /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.

[root@jfs12 ~]# /grid/bin/crsctl stop crs
CRS-2791: Starting shutdown of Oracle High Availability Services-managed
resources on 'jfs12'
CRS-2673: Attempting to stop 'ora.crsd' on 'jfs12'
CRS-2790: Starting shutdown of Cluster Ready Services-managed resources on
server 'jfs12'
CRS-2673: Attempting to stop 'ora.ntap.ntappdb1.pdb' on 'jfs12'
…
CRS-2673: Attempting to stop 'ora.gipcd' on 'jfs12'
CRS-2677: Stop of 'ora.gipcd' on 'jfs12' succeeded
CRS-2793: Shutdown of Oracle High Availability Services-managed resources
on 'jfs12' has completed
CRS-4133: Oracle High Availability Services has been stopped.

[root@jfs12 ~]# /grid/bin/crsctl start crs
CRS-4123: Oracle High Availability Services has been started.
```

Scénarios d’échec

Présentation

La planification d’une architecture complète d’applications de synchronisation active SnapMirror nécessite de comprendre comment les SM-AS répondront dans divers scénarios de basculement planifiés et non planifiés.

Pour les exemples suivants, supposons que le site A est configuré comme le site préféré.

Perte de la connectivité de réplication

Si la réplication SM-AS est interrompue, l’E/S d’écriture ne peut pas être terminée, car un cluster ne peut pas répliquer les modifications sur le site opposé.

Site A (site préféré)

Le résultat de l’échec de la liaison de réplication sur le site préféré sera une pause d’environ 15 secondes dans le traitement des E/S d’écriture, car ONTAP relance les opérations d’écriture répliquées avant de déterminer que la liaison de réplication est véritablement inaccessible. Au bout de 15 secondes, le site A du système reprend le traitement des E/S de lecture et d’écriture. Les chemins SAN ne changent pas et les LUN restent en ligne.

Site B

Le site B n'étant pas le site privilégié de synchronisation active SnapMirror, ses chemins de LUN deviennent indisponibles au bout de 15 secondes environ.

Panne du système de stockage

Le résultat d'une défaillance du système de stockage est presque identique au résultat de la perte du lien de réplication. Le site survivant devrait subir une pause d'E/S d'environ 15 seconde. Une fois cette période de 15 secondes écoulée, l'E/S reprend sur ce site comme d'habitude.

Perte du médiateur

Le service médiateur ne contrôle pas directement les opérations de stockage. Il fonctionne comme un chemin de contrôle alternatif entre les clusters. Il existe principalement pour automatiser le basculement sans les risques associés à un scénario « split-brain ». En conditions normales de fonctionnement, chaque cluster réplique les modifications apportées à son partenaire et chaque cluster peut donc vérifier que le cluster partenaire est en ligne et qu'il transmet les données. Si le lien de réplication échoue, la réplication s'arrête.

La raison pour laquelle un médiateur est nécessaire pour un basculement automatisé sécurisé est parce qu'il serait autrement impossible à un cluster de stockage de déterminer si la perte de la communication bidirectionnelle était le résultat d'une panne du réseau ou d'une défaillance réelle du stockage.

Le médiateur fournit un chemin alternatif pour chaque cluster afin de vérifier l'état de santé de son partenaire. Les scénarios sont les suivants :

- Si un cluster peut contacter directement son partenaire, les services de réplication sont opérationnels. Aucune action requise.
- Si un site privilégié ne peut pas contacter son partenaire directement ou via le médiateur, il suppose que le partenaire est réellement indisponible ou a été isolé et a mis ses chemins LUN hors ligne. Le site préféré va ensuite publier l'état RPO=0 et continuer à traiter les E/S en lecture et en écriture.
- Si un site non préféré ne peut pas contacter directement son partenaire, mais peut le contacter via le médiateur, il mettra ses chemins hors ligne et attend le retour de la connexion de réplication.
- Si un site non privilégié ne peut pas contacter son partenaire directement ou via un médiateur opérationnel, il suppose que le partenaire est réellement indisponible ou a été isolé et a mis ses chemins LUN hors ligne. Le site non privilégié va ensuite publier l'état RPO=0 et continuer le traitement des E/S en lecture et en écriture. Il assumera le rôle de la source de réplication et deviendra le nouveau site préféré.

Si le médiateur n'est pas disponible :

- En cas de défaillance des services de réplication, quelle qu'en soit la raison, y compris la défaillance du site ou du système de stockage non privilégié, le site préféré libère l'état RPO=0 et reprend le traitement des E/S de lecture et d'écriture. Le site non préféré mettra ses chemins hors ligne.
- La défaillance du site préféré entraînera une panne, car le site non préféré ne pourra pas vérifier que le site opposé est réellement hors ligne et, par conséquent, il ne serait pas sûr que le site non préféré puisse reprendre ses services.

Restauration des services

Après résolution d'une panne, par exemple lors de la restauration de la connectivité site à site ou de la mise sous tension d'un système défaillant, les terminaux de synchronisation active SnapMirror détectent automatiquement la présence d'une relation de réplication défectueuse et la raverront à l'état RPO=0. Une fois la réplication synchrone rétablie, les chemins défaillants se reconnectent.

Dans de nombreux cas, les applications en cluster détectent automatiquement le retour des chemins défaillants, et ces applications sont également reconnectées. Dans d'autres cas, une analyse SAN au niveau de l'hôte peut être nécessaire ou les applications doivent être reconnectées manuellement. Cela dépend de l'application et de la façon dont elle est configurée et, en général, de telles tâches peuvent être facilement automatisées. La fonctionnalité ONTAP elle-même est dotée d'une fonctionnalité d'autorétablissement et ne nécessite aucune intervention de l'utilisateur pour reprendre les opérations de stockage avec un objectif de point de récupération de 0.

Basculement manuel

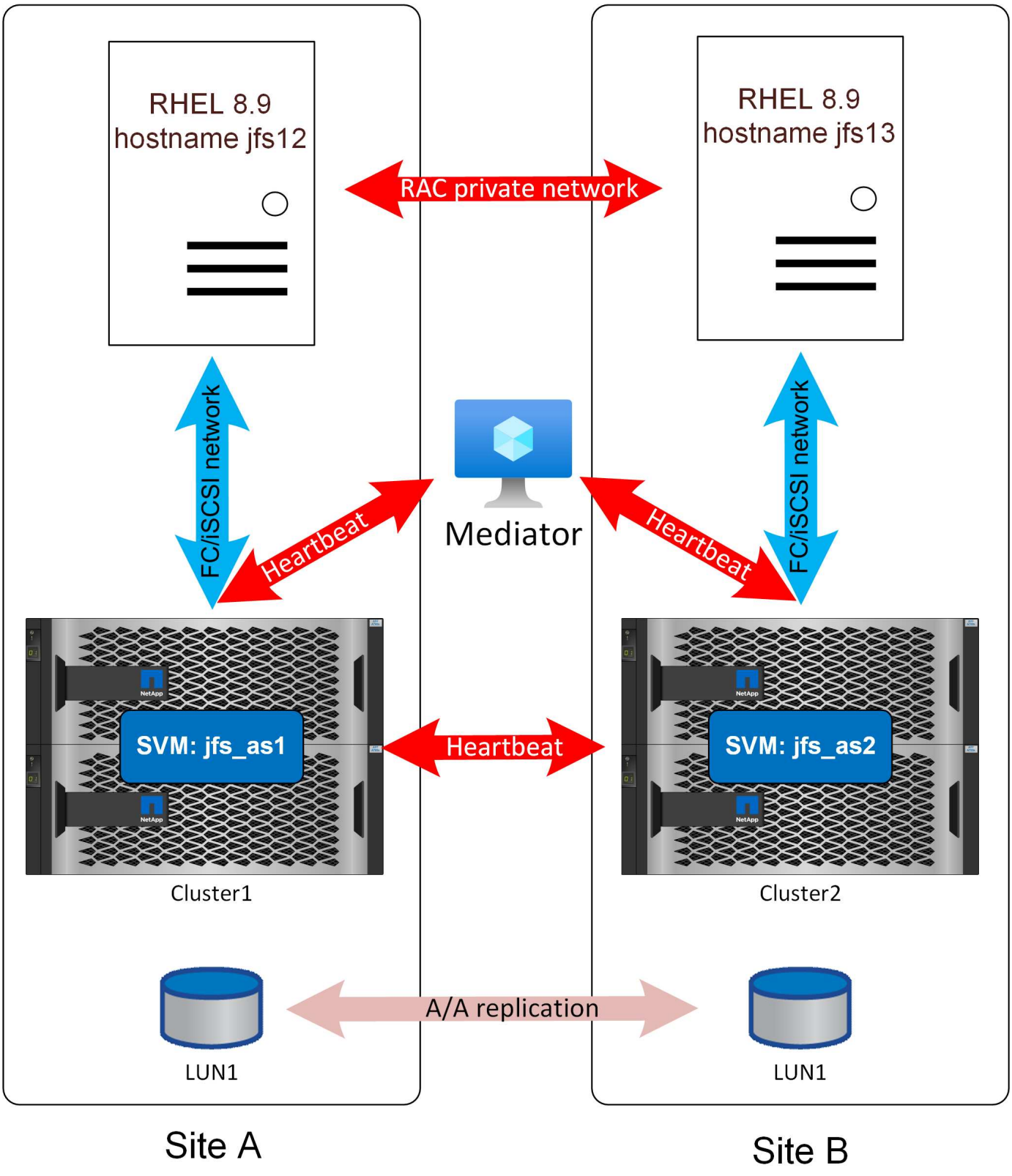
La modification du site préféré nécessite une opération simple. L'E/S s'interrompt pendant une ou deux secondes car l'autorité sur le comportement de réplication change entre les clusters, mais l'E/S n'est pas affectée.

Exemple d'architecture

Les exemples détaillés de défaillances présentés dans cette section sont basés sur l'architecture présentée ci-dessous.

> Il ne s'agit que de l'une des nombreuses options pour les bases de données Oracle sur la synchronisation active SnapMirror. Cette conception a été choisie parce qu'elle illustre certains des scénarios les plus complexes.

Dans cette conception, supposons que le site A est défini sur "site préféré".



Échec de l'interconnexion du RAC

La perte du lien de réplication RAC Oracle produira un résultat similaire à la perte de la connectivité SnapMirror, sauf que les délais d'expiration seront plus courts par défaut. Dans les paramètres par défaut, un nœud RAC Oracle attend 200 secondes après une

perte de connectivité du stockage avant d'être supprimé, mais il n'attend que 30 secondes après la perte du signal de détection du réseau RAC.

Les messages CRS sont similaires à ceux indiqués ci-dessous. Vous pouvez voir le délai d'expiration de 30 secondes. Comme css_Critical a été défini sur jfs12, situé sur le site A, ce sera le site pour survivre et jfs13 sur le site B sera supprimé.

```
2024-09-12 10:56:44.047 [ONMD(3528)]CRS-1611: Network communication with
node jfs13 (2) has been missing for 75% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 6.980 seconds
2024-09-12 10:56:48.048 [ONMD(3528)]CRS-1610: Network communication with
node jfs13 (2) has been missing for 90% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 2.980 seconds
2024-09-12 10:56:51.031 [ONMD(3528)]CRS-1607: Node jfs13 is being evicted
in cluster incarnation 621599354; details at (:CSSNM00007:) in
/gridbase/diag/crs/jfs12/crs/trace/onmd.trc.
2024-09-12 10:56:52.390 [CRSD(6668)]CRS-7503: The Oracle Grid
Infrastructure process 'crsd' observed communication issues between node
'jfs12' and node 'jfs13', interface list of local node 'jfs12' is
'192.168.30.1:33194;', interface list of remote node 'jfs13' is
'192.168.30.2:33621;'.
2024-09-12 10:56:55.683 [ONMD(3528)]CRS-1601: CSSD Reconfiguration
complete. Active nodes are jfs12 .
2024-09-12 10:56:55.722 [CRSD(6668)]CRS-5504: Node down event reported for
node 'jfs13'.
2024-09-12 10:56:57.222 [CRSD(6668)]CRS-2773: Server 'jfs13' has been
removed from pool 'Generic'.
2024-09-12 10:56:57.224 [CRSD(6668)]CRS-2773: Server 'jfs13' has been
removed from pool 'ora.NTAP'.
```

Échec de communication SnapMirror

Si la liaison de réplication SnapMirror active Sync, l'E/S d'écriture ne peut pas être terminée, car un cluster ne peut pas répliquer les modifications sur le site opposé.

Site A

Le site A qui présente une défaillance de liaison de réplication entraînera une pause d'environ 15 secondes dans le traitement des E/S d'écriture au fur et à mesure que ONTAP tente de répliquer des écritures avant de déterminer que la liaison de réplication est réellement inutilisable. Au bout de 15 secondes, le cluster ONTAP sur le site A reprend le traitement des E/S de lecture et d'écriture. Les chemins SAN ne changent pas et les LUN restent en ligne.

Site B

Le site B n'étant pas le site privilégié de synchronisation active SnapMirror, ses chemins de LUN deviennent indisponibles au bout de 15 secondes environ.

Le lien de réplication a été coupé à l'horodatage 15:19:44. Le premier avertissement d'Oracle RAC arrive 100 secondes plus tard lorsque le délai d'expiration de 200 secondes (contrôlé par le paramètre Oracle RAC disktimeout) approche.

```
2024-09-10 15:21:24.702 [ONMD(2792)]CRS-1615: No I/O has completed after
50% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 99340 milliseconds.
2024-09-10 15:22:14.706 [ONMD(2792)]CRS-1614: No I/O has completed after
75% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 49330 milliseconds.
2024-09-10 15:22:44.708 [ONMD(2792)]CRS-1613: No I/O has completed after
90% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 19330 milliseconds.
2024-09-10 15:23:04.710 [ONMD(2792)]CRS-1604: CSSD voting file is offline:
/dev/mapper/grid2; details at (:CSSNM00058:) in
/gridbase/diag/crs/jfs13/crs/trace/onmd.trc.
2024-09-10 15:23:04.710 [ONMD(2792)]CRS-1606: The number of voting files
available, 0, is less than the minimum number of voting files required, 1,
resulting in CSSD termination to ensure data integrity; details at
(:CSSNM00018:) in /gridbase/diag/crs/jfs13/crs/trace/onmd.trc
2024-09-10 15:23:04.716 [ONMD(2792)]CRS-1699: The CSS daemon is
terminating due to a fatal error from thread:
clssnmvDiskPingMonitorThread; Details at (:CSSSC00012:) in
/gridbase/diag/crs/jfs13/crs/trace/onmd.trc
2024-09-10 15:23:04.731 [OCSSD(2794)]CRS-1652: Starting clean up of CRSD
resources.
```

Une fois que le délai d'expiration du disque de vote de 200 secondes a été atteint, ce nœud RAC Oracle s'expulse automatiquement du cluster et redémarre.

Échec total de l'interconnectivité réseau

Si la liaison de réplication entre les sites est totalement perdue, la synchronisation active SnapMirror et la connectivité RAC Oracle seront interrompues.

La détection d'Oracle RAC à cerveau divisé dépend du pulsation du stockage Oracle RAC. Si la perte de la connectivité site à site entraîne la perte simultanée du signal de détection du réseau RAC et des services de réplication du stockage, les sites RAC ne pourront pas communiquer entre sites via l'interconnexion RAC ou les disques de vote RAC. Le résultat d'un ensemble de nœuds à numéro pair peut être l'exclusion des deux sites sous les paramètres par défaut. Le comportement exact dépend de la séquence des événements et de la synchronisation des sondages de pulsation du réseau RAC et du disque.

Le risque d'une panne sur deux sites peut être résolu de deux manières. Tout d'abord, une "disjoncteur d'attache" configuration peut être utilisée.

Si aucun site tiers n'est disponible, ce risque peut être résolu en ajustant le paramètre misscount sur le cluster RAC. Sous les valeurs par défaut, le délai d'expiration de la pulsation réseau du RAC est de 30 secondes. Il est généralement utilisé par RAC pour identifier les nœuds RAC défaillants et les supprimer du cluster. Il dispose également d'une connexion à la pulsation du disque de vote.

Si, par exemple, le conduit transportant le trafic intersite pour Oracle RAC et les services de réplication de stockage est coupé par une pelle rétro, le compte à rebours des erreurs de 30 secondes commence. Si le nœud du site RAC préféré ne peut pas rétablir le contact avec le site opposé dans les 30 secondes et qu'il ne peut pas utiliser les disques de vote pour confirmer que le site opposé est en panne dans la même fenêtre de 30 secondes, les nœuds du site préféré seront également supprimés. Il en résulte une interruption complète de la base de données.

Selon le moment où l'interrogation du compte erroné se produit, 30 secondes peuvent ne pas suffire à la temporisation de la synchronisation active SnapMirror et à permettre au stockage du site préféré de reprendre les services avant l'expiration de la fenêtre de 30 secondes. Cette fenêtre de 30 secondes peut être augmentée.

```
[root@jfs12 ~]# /grid/bin/crsctl set css misscount 100
CRS-4684: Successful set of parameter misscount to 100 for Cluster
Synchronization Services.
```

Cette valeur permet au système de stockage sur le site préféré de reprendre les opérations avant que le délai d'erreur n'expire. Le résultat sera alors la suppression uniquement des nœuds sur le site où les chemins de LUN ont été supprimés. Exemple ci-dessous :

```
2024-09-12 09:50:59.352 [ONMD(681360)]CRS-1612: Network communication with
node jfs13 (2) has been missing for 50% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 49.570 seconds
2024-09-12 09:51:10.082 [CRSD(682669)]CRS-7503: The Oracle Grid
Infrastructure process 'crsd' observed communication issues between node
'jfs12' and node 'jfs13', interface list of local node 'jfs12' is
'192.168.30.1:46039;', interface list of remote node 'jfs13' is
'192.168.30.2:42037;'.
2024-09-12 09:51:24.356 [ONMD(681360)]CRS-1611: Network communication with
node jfs13 (2) has been missing for 75% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 24.560 seconds
2024-09-12 09:51:39.359 [ONMD(681360)]CRS-1610: Network communication with
node jfs13 (2) has been missing for 90% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 9.560 seconds
2024-09-12 09:51:47.527 [OHASD(680884)]CRS-8011: reboot advisory message
from host: jfs13, component: cssagent, with time stamp: L-2024-09-12-
09:51:47.451
2024-09-12 09:51:47.527 [OHASD(680884)]CRS-8013: reboot advisory message
text: oracssdagent is about to reboot this node due to unknown reason as
it did not receive local heartbeats for 10470 ms amount of time
2024-09-12 09:51:48.925 [ONMD(681360)]CRS-1632: Node jfs13 is being
removed from the cluster in cluster incarnation 621596607
```

Le support Oracle déconseille fortement de modifier les paramètres misscount ou disktimeout pour résoudre les problèmes de configuration. Toutefois, la modification de ces paramètres peut s'avérer justifiée et inévitable dans de nombreux cas, notamment dans les configurations de démarrage SAN, de virtualisation et de réplication du stockage. Si, par exemple, vous avez rencontré des problèmes de stabilité avec un réseau SAN

ou IP qui ont entraîné des expulsions RAC, vous devez résoudre le problème sous-jacent et ne pas facturer les valeurs de l'erreur de décompte ou du dépassement de disque. La modification des délais pour résoudre les erreurs de configuration masque un problème et non pas résout un problème. La modification de ces paramètres pour configurer correctement un environnement RAC basé sur les aspects de conception de l'infrastructure sous-jacente est différente et est conforme aux instructions de support Oracle. Avec le démarrage SAN, il est courant d'ajuster misscount jusqu'à 200 pour correspondre au disktimeout. Voir "ce lien" pour plus d'informations.

Panne du site

Le résultat d'une défaillance du site ou du système de stockage est presque identique au résultat de la perte du lien de réplication. Le site survivant doit subir une pause d'E/S d'environ 15 secondes sur les écritures. Une fois cette période de 15 secondes écoulée, l'E/S reprend sur ce site comme d'habitude.

Si seul le système de stockage a été affecté, le nœud Oracle RAC sur le site en panne perdra les services de stockage et entrera le même compte à rebours de 200 secondes avant la suppression et le redémarrage suivant.

```
2024-09-11 13:44:38.613 [ONMD(3629)]CRS-1615: No I/O has completed after
50% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 99750 milliseconds.
2024-09-11 13:44:51.202 [ORAAGENT(5437)]CRS-5011: Check of resource "NTAP"
failed: details at "(:CLSN00007:)" in
"/gridbase/diag/crs/jfs13/crs/trace/crsd_oraagent_oracle.trc"
2024-09-11 13:44:51.798 [ORAAGENT(75914)]CRS-8500: Oracle Clusterware
ORAAGENT process is starting with operating system process ID 75914
2024-09-11 13:45:28.626 [ONMD(3629)]CRS-1614: No I/O has completed after
75% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 49730 milliseconds.
2024-09-11 13:45:33.339 [ORAAGENT(76328)]CRS-8500: Oracle Clusterware
ORAAGENT process is starting with operating system process ID 76328
2024-09-11 13:45:58.629 [ONMD(3629)]CRS-1613: No I/O has completed after
90% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 19730 milliseconds.
2024-09-11 13:46:18.630 [ONMD(3629)]CRS-1604: CSSD voting file is offline:
/dev/mapper/grid2; details at (:CSSNM00058:) in
/gridbase/diag/crs/jfs13/crs/trace/onmd.trc.
2024-09-11 13:46:18.631 [ONMD(3629)]CRS-1606: The number of voting files
available, 0, is less than the minimum number of voting files required, 1,
resulting in CSSD termination to ensure data integrity; details at
(:CSSNM00018:) in /gridbase/diag/crs/jfs13/crs/trace/onmd.trc
2024-09-11 13:46:18.638 [ONMD(3629)]CRS-1699: The CSS daemon is
terminating due to a fatal error from thread:
clssnmvDiskPingMonitorThread; Details at (:CSSSC00012:) in
/gridbase/diag/crs/jfs13/crs/trace/onmd.trc
2024-09-11 13:46:18.651 [OCSSD(3631)]CRS-1652: Starting clean up of CRSD
resources.
```

L'état du chemin SAN sur le nœud RAC qui a perdu des services de stockage se présente comme suit :

```
oradata7 (3600a0980383041334a3f55676c697347) dm-20 NETAPP,LUN C-Mode
size=128G features='3 queue_if_no_path pg_init_retries 50' hwhandler='1
alua' wp=rw
|-+- policy='service-time 0' prio=0 status=enabled
| `- 34:0:0:18 sdam 66:96  failed faulty running
`-+- policy='service-time 0' prio=0 status=enabled
  `- 33:0:0:18 sdaj 66:48  failed faulty running
```

L'hôte linux a détecté la perte des chemins beaucoup plus rapidement que 200 secondes, mais du point de vue de la base de données, les connexions client à l'hôte sur le site défaillant seront toujours bloquées pendant 200 secondes sous les paramètres Oracle RAC par défaut. Les opérations complètes de la base de données ne reprendront qu'une fois la suppression terminée.

Pendant ce temps, le nœud Oracle RAC sur le site opposé enregistre la perte de l'autre nœud RAC. Dans le cas contraire, le système continue de fonctionner normalement.

```
2024-09-11 13:46:34.152 [ONMD(3547)]CRS-1612: Network communication with
node jfs13 (2) has been missing for 50% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 14.020 seconds
2024-09-11 13:46:41.154 [ONMD(3547)]CRS-1611: Network communication with
node jfs13 (2) has been missing for 75% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 7.010 seconds
2024-09-11 13:46:46.155 [ONMD(3547)]CRS-1610: Network communication with
node jfs13 (2) has been missing for 90% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 2.010 seconds
2024-09-11 13:46:46.470 [OHASD(1705)]CRS-8011: reboot advisory message
from host: jfs13, component: cssmonit, with time stamp: L-2024-09-11-
13:46:46.404
2024-09-11 13:46:46.471 [OHASD(1705)]CRS-8013: reboot advisory message
text: At this point node has lost voting file majority access and
oracssdmonitor is rebooting the node due to unknown reason as it did not
receive local hearbeats for 28180 ms amount of time
2024-09-11 13:46:48.173 [ONMD(3547)]CRS-1632: Node jfs13 is being removed
from the cluster in cluster incarnation 621516934
```

Défaillance du médiateur

Le service médiateur ne contrôle pas directement les opérations de stockage. Il fonctionne comme un chemin de contrôle alternatif entre les clusters. Il existe principalement pour automatiser le basculement sans les risques associés à un scénario « split-brain ».

En conditions normales de fonctionnement, chaque cluster réplique les modifications apportées à son partenaire et chaque cluster peut donc vérifier que le cluster partenaire est en ligne et qu'il transmet les données. Si le lien de réplication échoue, la réplication s'arrête.

Un médiateur est nécessaire pour des opérations automatisées sécurisées, car il serait autrement impossible pour les clusters de stockage de déterminer si la perte de la communication bidirectionnelle était due à une panne du réseau ou à une défaillance réelle du stockage.

Le médiateur fournit un chemin alternatif pour chaque cluster afin de vérifier l'état de santé de son partenaire. Les scénarios sont les suivants :

- Si un cluster peut contacter directement son partenaire, les services de réplication sont opérationnels. Aucune action requise.
- Si un site privilégié ne peut pas contacter son partenaire directement ou via le médiateur, il suppose que le partenaire est réellement indisponible ou a été isolé et a mis ses chemins LUN hors ligne. Le site préféré va ensuite publier l'état RPO=0 et continuer à traiter les E/S en lecture et en écriture.
- Si un site non préféré ne peut pas contacter directement son partenaire, mais peut le contacter via le médiateur, il mettra ses chemins hors ligne et attend le retour de la connexion de réplication.

- Si un site non privilégié ne peut pas contacter son partenaire directement ou via un médiateur opérationnel, il suppose que le partenaire est réellement indisponible ou a été isolé et a mis ses chemins LUN hors ligne. Le site non privilégié va ensuite publier l'état RPO=0 et continuer le traitement des E/S en lecture et en écriture. Il assumera le rôle de la source de réplication et deviendra le nouveau site préféré.

Si le médiateur n'est pas disponible :

- En cas de défaillance des services de réplication, quelle qu'en soit la raison, le site préféré libère l'état RPO=0 et reprend le traitement des E/S en lecture et en écriture. Le site non préféré mettra ses chemins hors ligne.
- La défaillance du site préféré entraînera une panne, car le site non préféré ne pourra pas vérifier que le site opposé est réellement hors ligne et, par conséquent, il ne serait pas sûr que le site non préféré puisse reprendre ses services.

Restauration du service

SnapMirror propose une fonctionnalité d'autorétablissement. La synchronisation active SnapMirror détecte automatiquement la présence d'une relation de réplication défectueuse et la ramène à un état RPO=0. Une fois la réplication synchrone rétablie, les chemins reviennent en ligne.

Dans de nombreux cas, les applications en cluster détectent automatiquement le retour des chemins défaillants, et ces applications sont également reconnectées. Dans d'autres cas, une analyse SAN au niveau de l'hôte peut être nécessaire ou les applications doivent être reconnectées manuellement.

Cela dépend de l'application et de la façon dont elle est configurée et, en général, ces tâches peuvent être facilement automatisées. La synchronisation active SnapMirror elle-même est auto-fixing et ne nécessite aucune intervention de l'utilisateur pour reprendre les opérations de stockage avec un objectif de point de récupération de 0 une fois l'alimentation et la connectivité restaurées.

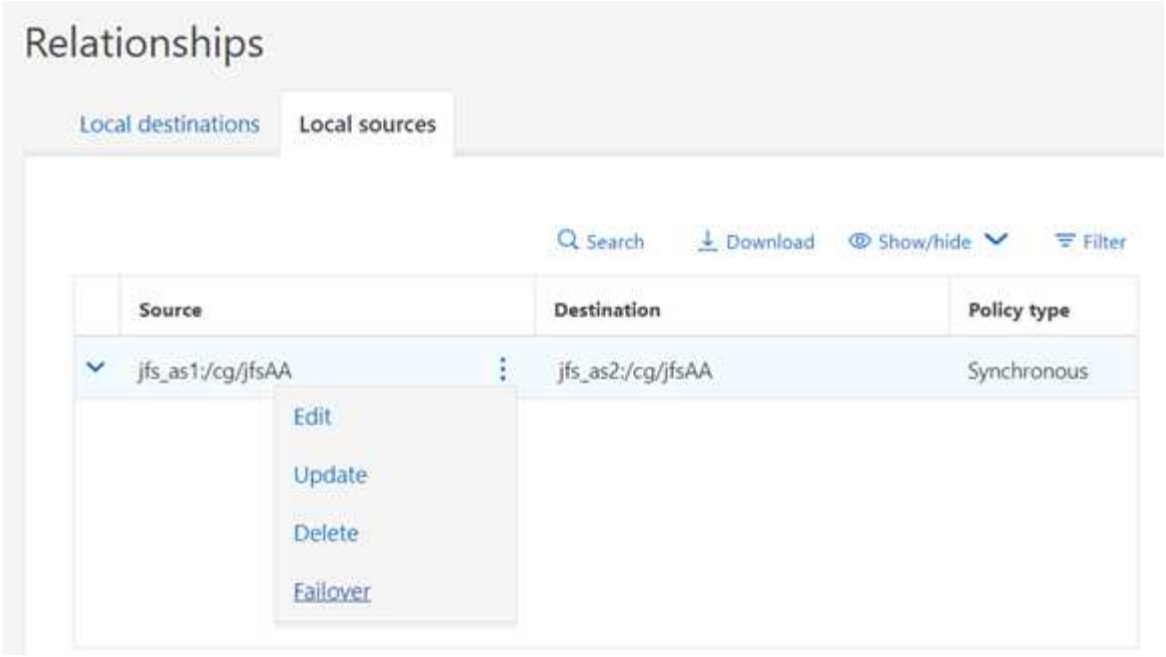
Basculement manuel

Le terme « basculement » ne fait pas référence au sens de la réplication avec la synchronisation active SnapMirror, car il s'agit d'une technologie de réplication bidirectionnelle. En revanche, le terme « basculement » désigne le système de stockage qui sera le site privilégié en cas de défaillance.

Par exemple, vous pouvez effectuer un basculement pour modifier le site préféré avant d'arrêter un site pour des raisons de maintenance ou avant d'effectuer un test de reprise après incident.

La modification du site préféré nécessite une opération simple. L'E/S s'interrompt pendant une ou deux secondes car l'autorité sur le comportement de réplication change entre les clusters, mais l'E/S n'est pas affectée.

Exemple d'interface graphique :



Exemple de modification via l'interface de ligne de commande :

```
Cluster2::> snapmirror failover start -destination-path jfs_as2:/cg/jfsAA
[Job 9575] Job is queued: SnapMirror failover for destination
"jfs_as2:/cg/jfsAA                      ".

Cluster2::> snapmirror failover show

Source    Destination                                           Error
Path      Path        Type     Status    start-time end-time   Reason
--------- ----------  -------- --------- ---------- ---------- ----------
jfs_as1:/cg/jfsAA
          jfs_as2:/cg/jfsAA
                      planned  completed 9/11/2024  9/11/2024
                                         09:29:22   09:29:32

The new destination path can be verified as follows:


Cluster1::> snapmirror show -destination-path jfs_as1:/cg/jfsAA

                           Source Path: jfs_as2:/cg/jfsAA
                        Destination Path: jfs_as1:/cg/jfsAA
                      Relationship Type: XDP
                Relationship Group Type: consistencygroup
                 SnapMirror Policy Type: automated-failover-duplex
                      SnapMirror Policy: AutomatedFailOverDuplex
                            Tries Limit: -
                           Mirror State: Snapmirrored
                    Relationship Status: InSync
```