Base de données Oracle

Enterprise applications

NetApp
February 10, 2026

Sommaire

Base de données Oracle

Bases de données Oracle sur ONTAP

ONTAP est conçu pour les bases de données Oracle. Pendant des décennies, ONTAP a été optimisé pour les demandes uniques d'E/S de bases de données relationnelles. Plusieurs fonctionnalités ONTAP ont été créées spécifiquement pour répondre aux besoins des bases de données Oracle, et même à la demande d'Oracle Inc. Elle-même.

Cette documentation remplace les rapports techniques publiés précédemment *TR-3633 : bases de données Oracle sur ONTAP ; TR-4591 : protection des données Oracle : sauvegarde, restauration, réplication ; TR-4592 : Oracle sur MetroCluster ; et TR-4534 : migration de bases de données Oracle vers des systèmes de stockage NetApp*

Outre les nombreuses possibilités offertes par ONTAP pour valoriser votre environnement de base de données, les besoins des utilisateurs sont très variés, notamment en termes de taille de la base de données, de performances et de protection des données. Les déploiements de systèmes de stockage NetApp prennent des formes diverses, qu'il s'agisse d'un environnement virtualisé incluant environ 6,000 bases de données fonctionnant sous VMware ESX ou d'un data warehouse à instance unique dont la taille de 996 To ne cesse de croître. Par conséquent, il existe peu de bonnes pratiques claires pour la configuration d'une base de données Oracle sur un système de stockage NetApp.

Les exigences relatives à l'exploitation d'une base de données Oracle sur un stockage NetApp sont traitées de deux manières. Tout d'abord, lorsqu'il existe une bonne pratique claire, elle sera appelée spécifiquement. D'une manière générale, de nombreuses considérations de conception à prendre en compte par les architectes de solutions de stockage Oracle en fonction de leurs besoins spécifiques seront expliquées.

Configuration ONTAP sur les systèmes AFF/ FAS

RAID

RAID désigne l'utilisation de la redondance pour protéger les données contre la perte d'un disque.

Des questions se posent parfois au sujet des niveaux RAID dans la configuration du stockage NetApp utilisé pour les bases de données Oracle et d'autres applications d'entreprise. De nombreuses meilleures pratiques Oracle en matière de configuration de baie de stockage contiennent des avertissements concernant l'utilisation de la mise en miroir RAID et/ou l'évitement de certains types de RAID. Bien qu'elles soulèvent des points valides, ces sources ne s'appliquent pas au RAID 4 et aux technologies NetApp RAID DP et RAID-TEC utilisées dans ONTAP.

RAID 4, RAID 5, RAID 6, RAID DP et RAID-TEC utilisent tous la parité pour s'assurer qu'une panne de disque n'entraîne pas de perte de données. Ces options RAID offrent une meilleure utilisation du stockage que la mise en miroir, mais la plupart des implémentations RAID présentent des inconvénients pour les opérations d'écriture. La réalisation d'une opération d'écriture sur d'autres implémentations RAID peut nécessiter plusieurs lectures de disque pour régénérer les données de parité, un processus communément appelé la pénalité RAID.

Cependant, ONTAP n'entraîne pas cette pénalité RAID. Cela est dû à l'intégration de NetApp WAFL (Write Anywhere File Layout) à la couche RAID. Les opérations d'écriture sont fusionnées dans la mémoire RAM et

préparées sous la forme d'une couche RAID complète, y compris la génération de la parité. ONTAP n'a pas besoin d'effectuer de lecture pour effectuer une écriture, ce qui signifie que ONTAP et WAFL évitent la pénalité RAID. Les performances des opérations stratégiques pour la latence, telles que la journalisation de reprise, sont assurées sans aucun obstacle. Les écritures aléatoires des fichiers de données n'entraînent aucune pénalité RAID résultant de la régénération de la parité.

En ce qui concerne la fiabilité statistique, même RAID DP offre une meilleure protection que la mise en miroir RAID. Le problème principal est la demande sur disques lors de la reconstruction RAID. Avec une configuration RAID en miroir, le risque de perte de données en cas de défaillance d'un disque pendant la reconstruction vers son partenaire dans la configuration RAID est bien plus grand que le risque de défaillance simultanée de trois disques dans une configuration RAID DP.

Gestion de la capacité

La gestion d'une base de données ou d'une autre application d'entreprise avec un stockage d'entreprise prévisible, gérable et haute performance requiert de l'espace libre sur les disques pour la gestion des données et des métadonnées. La quantité d'espace libre requise dépend du type de disque utilisé et des processus métier.

L'espace libre est défini comme tout espace qui n'est pas utilisé pour les données réelles et inclut l'espace non alloué sur l'agrégat lui-même et l'espace inutilisé au sein des volumes constitutifs. Le provisionnement fin doit également être envisagé. Par exemple, un volume peut contenir une LUN de 1 To, dont seulement 50 % sont utilisés par des données réelles. Dans un environnement à provisionnement fin, cet espace semble être consommé de 500 Go. Toutefois, dans un environnement entièrement provisionné, la capacité totale de 1 To semble être utilisée. Les 500 Go d'espace non alloué sont masqués. Cet espace n'est pas utilisé par les données réelles et doit donc être inclus dans le calcul de l'espace libre total.

Les recommandations de NetApp pour les systèmes de stockage utilisés pour les applications d'entreprise sont les suivantes :

Des agrégats SSD, y compris les systèmes AFF

NetApp recommande un minimum de 10% d'espace libre. Cela inclut tout l'espace inutilisé, y compris l'espace libre au sein de l'agrégat ou d'un volume, ainsi que tout espace libre alloué en raison de l'utilisation du provisionnement complet, mais qui n'est pas utilisé par les données réelles. L'espace logique n'est pas important, la question est de savoir quelle quantité d'espace physique réellement disponible pour le stockage des données.

La recommandation de 10 % d'espace libre est très prudente. Les agrégats SSD peuvent prendre en charge des charges de travail à des niveaux d'utilisation encore plus élevés, sans affecter les performances. Cependant, à mesure que l'utilisation de l'agrégat augmente, le risque de manquer d'espace augmente également si l'utilisation n'est pas surveillée de près. De plus, même si vous utilisez un système à 99 % de capacité, les performances risquent d'être moins élevées, mais vous devrez probablement interrompre la gestion pour l'empêcher de se remplir complètement lors de la commande de matériel supplémentaire. L'acquisition et l'installation de disques supplémentaires peuvent prendre un certain temps.

Les agrégats HDD, y compris les agrégats Flash Pool

NetApp recommande un minimum de 15 % d'espace libre lorsque des disques rotatifs sont utilisés. Cela inclut tout l'espace inutilisé, y compris l'espace libre au sein de l'agrégat ou d'un volume, ainsi que tout espace libre alloué en raison de l'utilisation du provisionnement complet, mais qui n'est pas utilisé par les données réelles. Les performances seront affectées à mesure que l'espace libre approche 10 %.

Ordinateurs virtuels de stockage

La gestion du stockage des bases de données Oracle est centralisée sur un SVM (Storage Virtual machine)

Un SVM, connu sous le nom de vserver sur l'interface de ligne de commandes ONTAP, est une unité fonctionnelle de base du stockage. Il est utile de comparer un SVM à un invité sur un serveur VMware ESX.

Lors de l'installation initiale, ESX ne possède pas de fonctionnalités préconfigurées, telles que l'hébergement d'un système d'exploitation invité ou la prise en charge d'une application utilisateur. Il s'agit d'un conteneur vide jusqu'à ce qu'une machine virtuelle (VM) soit définie. ONTAP fonctionne de manière similaire : Lors de la première installation de ONTAP, aucune fonctionnalité de service des données n'est disponible tant qu'un SVM n'est pas créé. Pour configurer les services de données.

À l'instar des autres aspects de l'architecture de stockage, les meilleures options pour la conception des SVM et de l'interface logique (LIF) dépendent largement des exigences d'évolutivité et des besoins de l'entreprise.

SVM

Il n'existe aucune bonne pratique officielle de provisionnement des SVM pour ONTAP. La bonne approche dépend des exigences en matière de gestion et de sécurité.

La plupart des clients utilisent un SVM principal pour la plupart de leurs besoins quotidiens, mais ils en créent un petit pour des besoins particuliers. Par exemple, vous pouvez créer :

- SVM d'une base de données stratégique gérée par une équipe de spécialistes
- SVM pour un groupe de développement auquel un contrôle administratif complet a été attribué afin de pouvoir gérer leur propre stockage indépendamment
- SVM pour les données sensibles de l'entreprise, telles que les données de rapports financiers ou de ressources humaines, pour lesquelles l'équipe administrative doit être limitée

Dans un environnement de colocation, on peut attribuer à chaque locataire une SVM dédiée aux données. La limite du nombre de SVM et de LIF par cluster, paire HA et nœud dépend du protocole utilisé, du modèle de nœud et de la version de ONTAP. Consulter le "NetApp Hardware Universe" pour ces limites.

Gestion des performances grâce à la QoS de ONTAP

Pour gérer efficacement et en toute sécurité plusieurs bases de données Oracle, il est nécessaire de disposer d'une stratégie de qualité de service efficace. C'est pourquoi les systèmes de stockage modernes offrent des performances toujours plus élevées.

Plus précisément, l'adoption croissante des systèmes de stockage 100 % Flash a permis de consolider les charges de travail. Les baies de stockage qui reposent sur des supports rotatifs ne prennent généralement en charge qu'un nombre limité de charges de travail exigeantes en E/S, car leurs capacités IOPS sont limitées par rapport aux anciens disques rotatifs. Une ou deux bases de données fortement actives saturaient les disques sous-jacents bien avant que les contrôleurs de stockage n'atteignent leurs limites. Cela a changé. Il

est possible de saturer les contrôleurs de stockage les plus puissants, car le nombre de disques SSD requis est relativement faible. Cela signifie que vous pouvez exploiter pleinement les capacités des contrôleurs sans craindre un effondrement soudain des performances lors de pics de latence des supports rotatifs.

À titre d'exemple de référence, un simple système AFF A800 HA à deux nœuds est capable de traiter jusqu'à un million d'IOPS aléatoires avant que la latence ne dépasse la milliseconde. On pourrait s'attendre à ce que très peu de charges de travail atteignent de tels niveaux. L'utilisation optimale de cette baie AFF A800 implique l'hébergement de plusieurs workloads. Pour ce faire, la sécurité et la prévisibilité exigent des contrôles de QoS.

Il existe deux types de qualité de service (QoS) dans ONTAP : les IOPS et la bande passante. Les contrôles de QoS peuvent être appliqués aux SVM, volumes, LUN et fichiers.

QoS des IOPS

Un contrôle de la QoS pour les IOPS est évidemment basé sur l'ensemble des IOPS d'une ressource donnée, mais il existe un certain nombre d'aspects de la QoS pour les IOPS qui peuvent ne pas être intuitifs. Au départ, quelques clients ont été surpris par l'augmentation apparente de la latence lorsqu'un seuil d'IOPS est atteint. L'augmentation de la latence est la conséquence naturelle de la limitation des IOPS. Logiquement, il fonctionne de la même manière qu'un système de jetons. Par exemple, si un volume donné contenant des fichiers de données dispose d'une limite de 10 000 IOPS, chaque E/S arrivant doit d'abord recevoir un jeton pour poursuivre le traitement. Tant que plus de 10 000 jetons n'ont pas été consommés en une seconde donnée, aucun retard n'est présent. Si les opérations d'E/S doivent attendre la réception de leur jeton, cet attente apparaît comme une latence supplémentaire. Plus une charge de travail est élevée, plus les E/S sont longues à attendre dans la file d'attente pour le traitement de son tour, ce qui apparaît comme une latence plus élevée.

> Soyez prudent lorsque vous appliquez des contrôles QoS aux données des transactions de base de données/journaux de reprise. Alors que les demandes de performances liées à la journalisation de reprise sont généralement très élevées, bien inférieures à celles des fichiers de données, l'activité du journal de reprise est en rafales. L'E/S se produit en de brèves impulsions et une limite de QoS qui semble appropriée pour les niveaux d'E/S de reprise moyens peut être trop basse pour les exigences réelles. Cela peut entraîner de strictes limitations de performance en cas d'engagement de la QoS avec chaque pic de journal de reprise. En général, la journalisation des opérations de reprise et d'archivage ne doit pas être limitée par la QoS.

QoS de la bande passante

Toutes les tailles d'E/S ne sont pas identiques. Par exemple, une base de données peut effectuer de nombreuses lectures de blocs de petite taille, ce qui entraînerait l'atteinte du seuil d'IOPS, mais il est également possible que les bases de données effectuent une analyse de table complète comprenant un très petit nombre de lectures de blocs volumineux, qui consomment une très grande quantité de bande passante, mais relativement peu d'IOPS.

De même, un environnement VMware peut générer un nombre très élevé d'IOPS aléatoires au démarrage, mais exécuter moins d'E/S, mais plus importantes, lors d'une sauvegarde externe.

Pour gérer efficacement les performances, les IOPS ou la bande passante doivent parfois être limitées, voire les deux.

QoS minimale/garantie

De nombreux clients recherchent une solution incluant une QoS garantie, qui semble plus difficile à atteindre qu'elle ne le paraît et qui risque d'être très gaspillée. Par exemple, pour placer 10 bases de données avec une

garantie de 10 000 IOPS, il est nécessaire de dimensionner un système dans le cas où les 10 bases de données s'exécutent simultanément à 10 000 IOPS, pour un total de 100 000.

La meilleure utilisation pour les contrôles QoS minimaux est de protéger les charges de travail stratégiques. Prenons l'exemple d'un contrôleur ONTAP avec un maximum de 500 000 IOPS et un mélange de charges de travail de production et de développement. Vous devez appliquer des règles de QoS maximales aux workloads de développement pour empêcher toute base de données de monopoliser le contrôleur. Vous appliqueriez ensuite des règles de QoS minimales aux charges de travail de production afin de vous assurer que les IOPS requises sont toujours disponibles, le cas échéant.

La QoS adaptative

La QoS adaptative fait référence à la fonctionnalité ONTAP où la limite de QoS repose sur la capacité de l'objet de stockage. Elle est rarement utilisée avec les bases de données, car il n'existe généralement aucun lien entre la taille d'une base de données et ses exigences de performances. Les grandes bases de données peuvent être quasiment inertes, tandis que les bases de données plus petites peuvent être celles qui nécessitent le plus d'IOPS.

La QoS adaptative peut s'avérer très utile avec les datastores de virtualisation, car les exigences en IOPS de ces jeux de données ont tendance à être corrélées à la taille totale de la base de données. Un datastore plus récent contenant 1 To de fichiers VMDK devrait avoir besoin d'environ la moitié des performances pour un datastore de 2 To. La QoS adaptative vous permet d'augmenter automatiquement les limites de qualité de service lorsque le datastore est rempli de données.

Efficacité

Les fonctionnalités ONTAP d'optimisation de l'espace sont optimisées pour les bases de données Oracle. Dans la plupart des cas, la meilleure approche consiste à conserver les valeurs par défaut avec toutes les fonctionnalités d'efficacité activées.

Les fonctionnalités d'optimisation de l'espace, telles que la compression, la compaction et la déduplication, sont conçues pour augmenter la quantité de données logiques correspondant à un volume de stockage physique donné. Vous réduisez ainsi vos coûts et vos frais de gestion.

À un niveau élevé, la compression est un processus mathématique qui permet de détecter et d'encoder des modèles de données de manière à réduire les besoins en espace. En revanche, la déduplication détecte les blocs de données répétés et supprime les copies parasites. La compaction permet à plusieurs blocs logiques de données de partager le même bloc physique sur le support.

> Reportez-vous aux sections ci-dessous sur le provisionnement fin pour une explication de l'interaction entre l'efficacité du stockage et la réservation fractionnaire.

Compression

Avant la disponibilité des systèmes de stockage 100 % Flash, la compression basée sur les baies était d'une valeur limitée, car la plupart des charges de travail exigeantes en E/S nécessitaient un très grand nombre de piles pour obtenir une performance acceptable. Les systèmes de stockage contenaient invariablement beaucoup plus de capacité que nécessaire, ce qui a pour effet d'augmenter le nombre de disques. La situation a changé avec la montée du stockage Solid-State. Il n'est plus nécessaire de surprovisionner des disques uniquement pour obtenir de bonnes performances. L'espace disque d'un système de stockage peut être adapté aux besoins réels en termes de capacité.

La capacité accrue des disques SSD en termes d'IOPS permet presque toujours de réaliser des économies

par rapport aux disques rotatifs. Toutefois, la compression peut réaliser davantage d'économies en augmentant la capacité effective des supports SSD.

Il existe plusieurs façons de compresser les données. De nombreuses bases de données incluent leurs propres fonctionnalités de compression, mais ce phénomène est rarement observé dans les environnements clients. La raison en est généralement la réduction des performances pour un **changement** de données compressées, plus avec certaines applications, il existe des coûts de licence élevés pour la compression au niveau de la base de données. Enfin, il y a les conséquences globales sur les performances des opérations des bases de données. Il est peu judicieux de payer un coût de licence par processeur élevé pour un processeur qui effectue la compression et la décompression des données plutôt que le véritable travail de base de données. Une meilleure option consiste à décharger la tâche de compression sur le système de stockage.

Compression adaptative

La compression adaptative a été testée en profondeur avec des charges de travail exigeantes sans effet sur les performances, même dans un environnement 100 % Flash où la latence se mesure en microsecondes. Certains clients ont même signalé une augmentation des performances due à l'utilisation de la compression, car les données restent compressées dans le cache, augmentant ainsi la quantité de cache disponible dans un contrôleur.

ONTAP gère les blocs physiques dans des unités de 4 Ko. La compression adaptative utilise une taille de bloc de compression par défaut de 8 Ko, ce qui signifie que les données sont compressées dans des unités de 8 Ko. La taille de bloc de 8 Ko la plus utilisée par les bases de données relationnelles est donc identique. Les algorithmes de compression deviennent plus efficaces avec la compression d'un volume croissant de données. Une taille de bloc de compression de 32 Ko serait plus compacte qu'une unité de bloc de compression de 8 Ko. Cela signifie que la compression adaptative utilisant une taille de bloc de 8 Ko par défaut entraîne des taux d'efficacité légèrement inférieurs, mais qu'une taille de bloc de compression inférieure présente également des avantages considérables. Les charges de travail de la base de données incluent une grande quantité d'activités de remplacement. Le remplacement d'un bloc de données de 32 Ko compressé de 8 Ko nécessite la lecture de l'intégralité des 32 Ko de données logiques, leur décompression, la mise à jour de la région de 8 Ko requise, la recompression, puis l'écriture de la totalité des 32 Ko sur les disques. Cette opération est très coûteuse pour un système de stockage. En effet, certaines baies de stockage concurrentes, basées sur des blocs de compression plus volumineux, affectent également considérablement les performances des charges de travail de la base de données.

La taille de bloc utilisée par la compression adaptative peut être augmentée jusqu'à 32 Ko. Cela peut améliorer l'efficacité du stockage et doit être envisagé pour les fichiers de repos tels que les journaux de transactions et les fichiers de sauvegarde lorsqu'une quantité importante de ces données est stockée sur la baie. Dans certains cas, les bases de données actives qui utilisent une taille de bloc de 16 ou 32 Ko peuvent également tirer parti de l'augmentation de la taille de bloc de la compression adaptative pour qu'elle corresponde. Consultez un représentant NetApp ou partenaire pour savoir si cette solution convient à votre charge de travail.

Les tailles de bloc de compression supérieures à 8 Ko ne doivent pas être utilisées avec la déduplication sur les destinations de sauvegarde en streaming. Les petites modifications apportées aux données sauvegardées affectent la fenêtre de compression de 32 Ko. Si la fenêtre change, les données compressées obtenues diffèrent dans l'ensemble du fichier. La déduplication a lieu après la compression, ce qui signifie que le moteur de déduplication voit chaque sauvegarde compressée différemment. Si la déduplication des sauvegardes en continu est nécessaire, seule une compression adaptative de bloc de 8 Ko doit être utilisée. Il est préférable d'utiliser la compression adaptative, car elle fonctionne à des blocs de taille réduite sans perturber l'efficacité de la déduplication. Pour des raisons similaires, la compression côté hôte interfère également avec l'efficacité de la déduplication.

Alignement de compression

La compression adaptative dans un environnement de base de données nécessite un certain respect de l'alignement des blocs de compression. Cela ne préoccupe que les données soumises à des écrasements aléatoires de blocs très spécifiques. Cette approche est similaire à l'alignement global du système de fichiers, où le début d'un système de fichiers doit être aligné sur une limite de périphérique de 4 Ko et la taille de bloc d'un système de fichiers doit être un multiple de 4 Ko.

Par exemple, une écriture de 8 Ko dans un fichier est compressée uniquement si elle s'aligne sur une limite de 8 Ko dans le système de fichiers lui-même. Ce point signifie qu'il doit figurer sur le premier 8 Ko du fichier, le deuxième 8 Ko du fichier, etc. La manière la plus simple de garantir un alignement correct est d'utiliser le type de LUN correct, toute partition créée doit avoir un décalage par rapport au début du périphérique qui est un multiple de 8K, et utiliser une taille de bloc du système de fichiers qui est un multiple de la taille de bloc de la base de données.

Les données telles que les sauvegardes ou les journaux de transactions sont des opérations écrites de manière séquentielle sur plusieurs blocs, qui sont tous compressés. Par conséquent, il n'est pas nécessaire de considérer l'alignement. Le seul modèle d'E/S préoccupant est l'écrasement aléatoire des fichiers.

Compaction

La compaction est une technologie qui améliore l'efficacité de la compression. Comme indiqué précédemment, la compression adaptative à elle seule permet d'économiser 2:1 au maximum, car elle se limite au stockage d'une E/S de 8 Ko dans un bloc WAFL de 4 Ko. Les méthodes de compression avec des blocs de taille supérieure améliorent l'efficacité. Cependant, elles ne conviennent pas aux données soumises à des remplacements de blocs de petite taille. La décompression d'unités de données de 32 Ko, la mise à jour d'une partie de 8 Ko, la recompression et l'écriture sur les disques entraînent une surcharge.

La compaction des données permet de stocker plusieurs blocs logiques dans des blocs physiques. Par exemple, une base de données avec des données fortement compressibles comme des blocs texte ou partiellement pleins peut être compressée de 8 Ko à 1 Ko. Sans compaction, 1 Ko de données occuperaient toujours un bloc complet de 4 Ko. La compaction des données à la volée permet de stocker 1 Ko de données compressées dans un espace physique de seulement 1 Ko, parallèlement à d'autres données compressées. Il ne s'agit pas d'une technologie de compression. Il s'agit simplement d'un moyen plus efficace d'allouer de l'espace sur les disques et, par conséquent, il ne doit pas créer d'effet détectable sur les performances.

Le degré d'économie obtenu varie. En général, les données déjà compressées ou chiffrées ne peuvent pas être compressées davantage et, par conséquent, la compaction de ces datasets ne peut pas être bénéfique. À contrario, les fichiers de données récemment initialisés ne contiennent qu'un petit peu plus que des métadonnées de bloc et des zéros compressent jusqu'à 80:1.

Efficacité du stockage sensible à la température

L'efficacité de stockage sensible à la température (TSSE) est disponible dans ONTAP 9.8 et versions ultérieures. Il s'appuie sur des cartes thermiques d'accès aux blocs pour identifier les blocs peu utilisés et les compresser à l'aide d'une meilleure efficacité.

Déduplication

La déduplication permet de supprimer les tailles de bloc dupliquées d'un dataset. Par exemple, si le même bloc de 4 Ko existe dans 10 fichiers différents, la déduplication redirige ce bloc de 4 Ko au sein des 10 fichiers vers le même bloc physique de 4 Ko. Résultat : une amélioration de l'efficacité de ces données de 10:1.

Les données, telles que les LUN de démarrage invité VMware, se dédupliquent extrêmement bien, car elles sont constituées de plusieurs copies des mêmes fichiers du système d'exploitation. L'efficacité de 100:1 et plus

ont été observées.

Certaines données ne contiennent pas de données dupliquées. Par exemple, un bloc Oracle contient un en-tête globalement unique à la base de données et une bande-annonce presque unique. Par conséquent, la déduplication d'une base de données Oracle permet rarement de réaliser plus de 1 % d'économies. La déduplication avec les bases de données MS SQL est légèrement meilleure, mais les métadonnées uniques au niveau des blocs restent une limitation.

Dans quelques cas, des économies d'espace allant jusqu'à 15 % ont été observées pour les bases de données de 16 Ko et les blocs volumineux. La bande de 4 Ko initiale de chaque bloc contient l'en-tête unique dans le monde, et le bloc de 4 Ko final contient la remorque presque unique. Les blocs internes sont candidats à la déduplication, bien que dans la pratique cela soit presque entièrement attribué à la déduplication des données mises à zéro.

De nombreuses baies concurrentes prétendent être capables de dédupliquer des bases de données en présumant qu'une base de données est copiée plusieurs fois. Il est également possible d'utiliser la déduplication NetApp, mais ONTAP offre une meilleure option : la technologie FlexClone de NetApp. Le résultat final est le même : plusieurs copies d'une base de données qui partagent la plupart des blocs physiques sous-jacents sont créées. L'utilisation de FlexClone est bien plus efficace que de prendre le temps de copier les fichiers de base de données, puis de les dédupliquer. Il s'agit en effet de la non-duplication plutôt que de la déduplication, car un doublon n'est jamais créé à la première place.

Efficacité et provisionnement fin

Les fonctions d'efficacité sont des formes de provisionnement fin. Par exemple, une LUN de 100 Go occupant un volume de 100 Go peut compresser à 50 Go. Aucune économie réelle n'est encore réalisée, car le volume est toujours de 100 Go. Le volume doit d'abord être réduit afin que l'espace économisé puisse être utilisé ailleurs sur le système. Si des modifications ultérieures de la LUN de 100 Go réduisent la taille des données compressibles, la LUN augmente et le volume pourrait se remplir.

Le provisionnement fin est fortement recommandé car il simplifie la gestion tout en améliorant la capacité exploitable avec les économies associées. La raison en est simple : les environnements de base de données comportent souvent beaucoup d'espace vide, un grand nombre de volumes et de LUN, ainsi que des données compressibles. Le provisionnement fin entraîne la réservation d'espace sur le stockage pour les volumes et les LUN au cas où un jour ils se traduirait par une saturation de 100 % et contiendraient des données non compressibles à 100 %. Il est peu probable que cela se produise. Le provisionnement fin permet de récupérer et d'utiliser cet espace ailleurs. Il permet également de gérer la capacité en fonction du système de stockage lui-même, plutôt que de nombreux volumes et LUN plus petits.

Certains clients préfèrent utiliser le provisionnement lourd, soit pour des charges de travail spécifiques, soit généralement en fonction de pratiques opérationnelles et d'approvisionnement établies.

> Si un volume est configuré en mode « Thick provisioning », veillez à désactiver complètement toutes les fonctions d'efficacité de ce volume, y compris la décompression et la suppression de la déduplication à l'aide de la `sis undo` commande. Le volume ne doit pas apparaître en `volume efficiency show` sortie. Si c'est le cas, le volume est encore partiellement configuré pour les fonctions d'efficacité. Par conséquent, les garanties de remplacement fonctionnent différemment, ce qui augmente le risque que les dépassements de configuration entraînent un manque inattendu d'espace du volume, ce qui entraîne des erreurs d'E/S de la base de données.

Meilleures pratiques en matière d'efficacité

NetApp recommande ce qui suit :

AFF par défaut

Les volumes créés sur ONTAP et exécutés sur un système AFF 100 % Flash sont à allocation dynamique, avec l'activation de toutes les fonctionnalités d'efficacité à la volée. Bien que les bases de données ne bénéficient généralement pas de la déduplication et puissent inclure des données non compressibles, les paramètres par défaut conviennent néanmoins à la plupart des charges de travail. ONTAP est conçu pour traiter efficacement tous les types de données et de modèles d'E/S, qu'ils entraînent ou non des économies. Les valeurs par défaut ne doivent être modifiées que si les raisons sont parfaitement comprises et si un écart est bénéfique.

Recommandations générales

- Si les volumes et/ou les LUN ne sont pas à provisionnement fin, vous devez désactiver tous les paramètres d'efficacité car l'utilisation de ces fonctionnalités ne permet aucune économie et la combinaison du provisionnement lourd et de l'optimisation de l'espace peut provoquer des comportements inattendus, notamment des erreurs de manque d'espace.
- Si les données ne sont pas sujettes à des écrasements, par exemple avec des sauvegardes ou des journaux de transactions de base de données, vous pouvez atteindre une meilleure efficacité en activant TSSE avec une période de refroidissement faible.
- Certains fichiers peuvent contenir une quantité importante de données non compressibles, par exemple lorsque la compression est déjà activée au niveau de l'application, les fichiers sont cryptés. Si l'un de ces scénarios est vrai, envisagez de désactiver la compression pour permettre un fonctionnement plus efficace sur d'autres volumes contenant des données compressibles.
- N'utilisez pas la compression et la déduplication de 32 Ko pour les sauvegardes de bases de données. Voir la section Compression adaptative pour plus d'informations.

Provisionnement fin

Le provisionnement fin pour une base de données Oracle nécessite une planification minutieuse, car il en résulte une configuration d'espace sur un système de stockage qui n'est pas nécessairement physiquement disponible. Cela vaut vraiment le coup, car une fois correctement effectué, il en résulte des économies considérables et des améliorations en termes de gestion.

Le provisionnement fin, de nombreuses formes, fait partie intégrante de nombreuses fonctionnalités offertes par ONTAP à l'environnement applicatif d'entreprise. Le provisionnement fin est également étroitement lié aux technologies d'efficacité pour la même raison : les fonctionnalités d'efficacité permettent de stocker davantage de données logiques que ce qui existe techniquement sur le système de stockage.

La plupart des snapshots impliquent un provisionnement fin. Par exemple, une base de données classique de 10 To sur un système de stockage NetApp compte environ 30 jours de copies Snapshot. Cet arrangement donne lieu à environ 10 To de données visibles dans le système de fichiers actif et 300 To dédiés aux snapshots. La capacité totale de stockage de 310 To réside généralement dans un espace d'environ 12 To à 15 To. La base de données active consomme 10 To et les 300 To de données restantes ne nécessitent que 2 à 5 To d'espace, car seules les modifications apportées aux données d'origine sont stockées.

Le clonage est également un exemple de provisionnement fin. Un client NetApp majeur a créé 40 clones d'une base de données de 80 To à utiliser pour le développement. Si les 40 développeurs qui utilisent ces clones surécrivent chaque bloc dans chaque fichier de données, plus de 3,2 po de stockage seraient nécessaires. En pratique, le chiffre d'affaires est faible et l'espace collectif requis est proche de 40 To, car seules les modifications sont stockées sur les disques.

Gestion de l'espace

Le provisionnement fin d'un environnement applicatif doit être extrêmement prudent, car les taux de modification des données peuvent augmenter de manière inattendue. Par exemple, la consommation d'espace due aux snapshots peut augmenter rapidement si les tables de base de données sont réindexées ou si des correctifs à grande échelle sont appliqués aux invités VMware. Une sauvegarde mal placée peut écrire une grande quantité de données dans un délai très court. Enfin, il peut être difficile de restaurer certaines applications si un système de fichiers manque d'espace de façon inattendue.

Avec une configuration soigneuse de, ces risques peuvent être maîtrisés `volume-autogrow` et `snapshot-autodelete` règles. Comme leurs noms l'indiquent, ces options permettent de créer des règles qui effacent automatiquement l'espace consommé par les snapshots ou augmentent un volume pour prendre en charge des données supplémentaires. De nombreuses options sont disponibles et les besoins varient selon les clients.

Voir la "documentation sur la gestion du stockage logique" pour une discussion complète de ces fonctionnalités.

Réservations fractionnaires

La réserve fractionnaire fait référence au comportement d'une LUN dans un volume en ce qui concerne l'efficacité de l'espace. Lorsque l'option `fractional-reserve` est défini sur 100 %. toutes les données du volume peuvent connaître un taux de rotation de 100 % avec n'importe quel modèle de données, sans épuiser l'espace sur le volume.

Par exemple, prenons l'exemple d'une base de données située sur une seule LUN de 250 Go dans un volume de 1 To. La création d'un snapshot entraînerait immédiatement la réservation d'un espace supplémentaire de 250 Go dans le volume, garantissant ainsi que l'espace disponible sur le volume ne serait pas insuffisant pour quelque raison que ce soit. L'utilisation de réserves fractionnaires est généralement inutile car il est très peu probable que chaque octet du volume de base de données ait besoin d'être écrasé. Il n'y a aucune raison de réserver de l'espace pour un événement qui ne se produit jamais. Cependant, si un client ne peut pas surveiller la consommation d'espace dans un système de stockage et doit être certain que l'espace ne sera jamais épuisé, des réservations fractionnaires de 100 % seront nécessaires pour utiliser les snapshots.

Compression et déduplication

La compression et la déduplication sont deux formes de provisionnement fin. Par exemple, une empreinte des données de 50 To peut être compressée jusqu'à 30 To, ce qui permet d'économiser 20 To. Pour que la compression offre tous les avantages, il faut utiliser quelques 20 To pour d'autres données ou acheter le système de stockage avec moins de 50 To. Il en résulte une quantité de données stockées supérieure à ce qui n'est techniquement disponible sur le système de stockage. Du point de vue des données, il y a 50 To de données, même si celles-ci ne occupent que 30 To sur les disques.

Il est toujours possible que la compressibilité d'un dataset change, ce qui entraîne une consommation accrue de l'espace réel. Cette augmentation de la consommation signifie que la compression doit être gérée comme avec les autres formes de provisionnement fin en termes de surveillance et d'utilisation `volume-autogrow` et `snapshot-autodelete`.

La compression et la déduplication sont présentées plus en détail dans la section xref:./oracle/efficiency.html

Compression et réservations fractionnaires

La compression est une forme d'allocation dynamique. Les réservations fractionnaires affectent l'utilisation de la compression, avec une remarque importante ; l'espace est réservé avant la création du snapshot. Normalement, la réserve fractionnaire n'est importante que si un instantané existe. S'il n'y a pas de snapshot,

la réserve fractionnaire n'est pas importante. Ce n'est pas le cas avec la compression. Si une LUN est créée sur un volume avec compression, ONTAP conserve l'espace nécessaire pour prendre en charge un snapshot. Ce comportement peut être déroutant pendant la configuration, mais il est normal.

Prenons l'exemple d'un volume de 10 Go avec une LUN de 5 Go compressée à 2,5 Go sans copie Snapshot. Prenez en compte ces deux scénarios :

- La réserve fractionnaire = 100 entraîne une utilisation de 7,5 Go
- La réserve fractionnaire = 0 entraîne une utilisation de 2,5 Go

Le premier scénario comprend 2,5 Go de consommation d'espace pour les données actuelles et 5 Go d'espace pour représenter 100 % de chiffre d'affaires de la source en prévision de l'utilisation des snapshots. Le deuxième scénario ne réserve pas d'espace supplémentaire.

Bien que cette situation puisse sembler confuse, il est peu probable qu'elle soit rencontrée dans la pratique. La compression implique un provisionnement fin et le provisionnement fin dans un environnement LUN nécessite des réservations fractionnaires. Il est toujours possible d'écraser des données compressées par un élément non compressible, ce qui signifie qu'un volume doit être à provisionnement fin pour la compression, pour réaliser des économies.

NetApp recommande les configurations de réserve suivantes :

- Réglez `fractional-reserve` à 0 lorsque la surveillance de la capacité de base est en place avec `volume-autogrow` et `snapshot-autodelete`.
- Réglez `fractional-reserve` à 100 s'il n'y a pas de capacité de surveillance ou s'il est impossible d'évacuer l'espace en quelque circonstance que ce soit.

Allocation d'espace libre et d'espace LVM

L'efficacité du provisionnement fin des LUN actifs dans un environnement de système de fichiers peut diminuer au fil du temps à mesure que des données sont supprimées. À moins que les données supprimées ne soient écrasées par des zéros (voir aussi ASMRU) ou que l'espace ne soit libéré avec la récupération d'espace TRIM/UNMAP, les données « effacées » occupent de plus en plus d'espace blanc non alloué dans le système de fichiers. De plus, le provisionnement fin des LUN actifs est d'une utilité limitée dans de nombreux environnements de bases de données, car les fichiers de données sont initialisés à leur taille maximale au moment de leur création.

Une planification minutieuse de la configuration de LVM peut améliorer l'efficacité et réduire les besoins en provisionnement du stockage et en redimensionnement des LUN. Lorsqu'un LVM tel que Veritas VxVM ou Oracle ASM est utilisé, les LUN sous-jacentes sont divisés en extensions qui ne sont utilisées que lorsque cela est nécessaire. Par exemple, si un dataset commence à 2 To mais peut atteindre 10 To au fil du temps, ce dataset peut être placé sur 10 To de LUN à provisionnement fin organisées dans un groupe de disques LVM. Elle occupant seulement 2 To d'espace au moment de la création et réclarait uniquement de l'espace supplémentaire, dans la mesure où les extensions sont allouées pour prendre en charge la croissance du volume des données. Ce processus est sûr tant que l'espace est surveillé.

Basculement/basculement ONTAP

Il est nécessaire de bien comprendre les fonctions de basculement et de basculement du stockage pour s'assurer que les opérations de la base de données Oracle ne sont pas interrompues par ces opérations. En outre, les arguments utilisés par les opérations de basculement et de basculement peuvent affecter l'intégrité des données en cas

d’utilisation incorrecte.

- Dans des conditions normales, les écritures entrantes sur un contrôleur donné sont mises en miroir de manière synchrone sur son partenaire. Dans un environnement NetApp MetroCluster, les écritures sont également mises en miroir sur un contrôleur distant. Tant qu’une écriture n’est pas stockée sur un support non volatile dans tous les emplacements, elle n’est pas validée par l’application hôte.
- Le support qui stocke les données d’écriture est appelé mémoire non volatile ou NVMEM. Elle est également parfois appelée mémoire NVRAM, et peut être considérée comme un cache d’écriture, même si elle fonctionne comme un journal. En fonctionnement normal, les données de NVMEM ne sont pas lues ; elles sont uniquement utilisées pour protéger les données en cas de défaillance logicielle ou matérielle. Lors de l’écriture des données sur les disques, les données sont transférées de la mémoire RAM du système, et non de NVMEM.
- Lors d’une opération de basculement, un nœud d’une paire haute disponibilité reprend les opérations de son partenaire. Un basculement est quasiment identique, mais s’applique aux configurations MetroCluster dans lesquelles un nœud distant prend le relais par rapport à un nœud local.

Lors des opérations de maintenance de routine, un basculement du stockage ou un basculement doivent être transparents, sauf en cas de brève pause potentielle dans les opérations en cas de changement des chemins réseau. La mise en réseau peut toutefois être complexe et il est facile d’y faire des erreurs. NetApp recommande donc de tester minutieusement les opérations de basculement et de basculement avant de mettre en production un système de stockage. C’est la seule façon de s’assurer que tous les chemins réseau sont correctement configurés. Dans un environnement SAN, vérifiez soigneusement le résultat de la commande `sanlun lun show -p` pour vous assurer que tous les chemins principaux et secondaires attendus sont disponibles.

Il convient de faire attention lors d’un basculement forcé ou d’un basculement forcé. Forcer une modification de la configuration du stockage avec ces options signifie que l’état du contrôleur propriétaire des disques est ignoré et que le nœud alternatif prend le contrôle des disques. Une force de basculement incorrecte peut entraîner une perte ou une corruption des données. En effet, un basculement forcé ou un basculement forcé peut rejeter le contenu de la NVMEM. Une fois le basculement ou le basculement effectué, la perte de ces données signifie que les données stockées sur les disques peuvent revenir à un état plus ancien du point de vue de la base de données.

Un basculement forcé avec une paire haute disponibilité normale devrait rarement être nécessaire. Dans la plupart des scénarios de défaillance, un nœud s’arrête et informe le partenaire qu’un basculement automatique a lieu. Il existe certains cas à la périphérie, par exemple une panne de déploiement où l’interconnexion entre les nœuds est perdue puis un contrôleur est perdu, dans lequel un basculement forcé est nécessaire. Dans ce cas, la mise en miroir entre les nœuds est perdue avant la panne du contrôleur, ce qui signifie que le contrôleur survivant n’aurait plus de copie des écritures en cours. Le basculement doit ensuite être forcé, ce qui signifie que des données peuvent être perdues.

La même logique s’applique à un basculement MetroCluster. Dans des conditions normales, le basculement est presque transparent. Toutefois, un incident peut entraîner une perte de connectivité entre le site survivant et le site de reprise sur incident. Du point de vue du site survivant, le problème ne pourrait être rien de plus qu’une interruption de la connectivité entre les sites, et le site d’origine pourrait encore traiter les données. Si un nœud ne peut pas vérifier l’état du contrôleur principal, seul un basculement forcé est possible.

NetApp recommande de prendre les précautions suivantes :

- Veillez à ne pas forcer accidentellement un basculement ou un basculement. En règle générale, il n'est pas nécessaire de forcer et le fait de forcer la modification peut entraîner la perte de données.
- Si un basculement ou un basculement forcé s'avère nécessaire, assurez-vous que les applications sont arrêtées, que tous les systèmes de fichiers sont démontés et que les groupes de volumes LVM (Logical Volume Manager) sont proposés en mode Variyoffed. Les groupes de disques ASM doivent être démontés.
- En cas de basculement forcé du MetroCluster, vous pouvez isoler le nœud défaillant de toutes les ressources de stockage restantes. Pour plus d'informations, consultez le Guide de gestion et de reprise sur incident de MetroCluster correspondant à la version appropriée de ONTAP.

MetroCluster et plusieurs agrégats

MetroCluster est une technologie de réplication synchrone qui passe en mode asynchrone en cas d'interruption de la connectivité. Cette demande est la plus courante de la part des clients, car une réplication synchrone garantie signifie que l'interruption de la connectivité du site entraîne un blocage complet des E/S de la base de données, ce qui la met hors service.

Avec MetroCluster, les agrégats sont rapidement resynchronisés une fois la connectivité restaurée. Contrairement à d'autres technologies de stockage, MetroCluster ne devrait jamais nécessiter de mise en miroir complète après une panne de site. Seules les modifications delta doivent être expédiées.

Dans les jeux de données qui couvrent les agrégats, le risque est faible de nécessiter des étapes supplémentaires de restauration des données en cas de sinistre en cas de déploiement. En particulier, si (a) la connectivité entre les sites est interrompue, (b) la connectivité est restaurée, (c) les agrégats atteignent un état dans lequel certains sont synchronisés et d'autres ne le sont pas, puis (d) le site primaire est perdu, le site survivant dans lequel les agrégats ne sont pas synchronisés. Dans ce cas, une partie du dataset est synchronisée et il est impossible d'ouvrir des applications, des bases de données ou des datastores sans restauration. Si un dataset compte plusieurs agrégats, NetApp recommande vivement d'utiliser des sauvegardes basées sur des snapshots avec l'un des nombreux outils disponibles pour vérifier la restauration rapide dans ce scénario inhabituel.

Configuration ONTAP sur les systèmes ASA r2

RAID

Le RAID désigne l'utilisation de la redondance basée sur la parité pour protéger les données contre les pannes de disque. ASA r2 utilise les mêmes technologies RAID ONTAP que les systèmes AFF et FAS , assurant une protection robuste contre les pannes de plusieurs disques.

ONTAP effectue automatiquement la configuration RAID pour les systèmes ASA r2. Il s'agit d'un élément essentiel de l'expérience simplifiée de gestion du stockage introduite avec l'interface ASA r2.

Les principaux détails concernant la configuration RAID automatique sur ASA r2 sont les suivants :

- Zones de disponibilité de stockage (SAZ) : Au lieu de gérer manuellement les agrégats et les groupes RAID traditionnels, ASA r2 utilise des zones de disponibilité de stockage (SAZ). Il s'agit de pools de

disques partagés et protégés par RAID pour une paire HA, où les deux nœuds ont un accès complet au même stockage.

- Placement automatique : lorsqu'une unité de stockage (espace de noms LUN ou NVMe) est créée, ONTAP crée automatiquement un volume au sein de la SAZ et le place pour un équilibre optimal entre performances et capacité.
- Aucune gestion manuelle des agrégats : les commandes traditionnelles de gestion des agrégats et des groupes RAID ne sont pas prises en charge sur ASA r2. Cela élimine la nécessité pour les administrateurs de planifier manuellement la taille des groupes RAID, les disques de parité ou l'affectation des nœuds.
- Provisionnement simplifié : le provisionnement est géré via System Manager ou des commandes CLI simplifiées qui se concentrent sur les unités de stockage plutôt que sur la configuration RAID physique sous-jacente.
- Rééquilibrage de la charge de travail : à partir des versions 2025 (ONTAP 9.17.1), ONTAP rééquilibre automatiquement les charges de travail entre les nœuds de la paire HA afin de garantir que les performances et l'utilisation de l'espace restent équilibrées sans intervention manuelle.

ASA r2 utilise automatiquement les technologies RAID par défaut d'ONTAP : RAID DP pour la plupart des configurations et RAID-TEC pour les très grands pools SSD. Cela élimine le besoin de sélection manuelle du RAID. Ces niveaux RAID basés sur la parité offrent une meilleure efficacité et une meilleure fiabilité de stockage que la mise en miroir, que les anciennes bonnes pratiques d'Oracle recommandent souvent mais qui n'est pas pertinente pour ASA r2. ONTAP évite la pénalité d'écriture RAID traditionnelle grâce à l'intégration WAFL , garantissant des performances optimales pour les charges de travail Oracle telles que la journalisation des journaux de restauration et les écritures aléatoires de fichiers de données. Associé à la gestion automatisée du RAID et aux zones de disponibilité du stockage, ASA r2 offre une haute disponibilité et une protection de niveau entreprise pour les bases de données Oracle.

Gestion de la capacité

La gestion d'une base de données ou d'une autre application d'entreprise avec un stockage d'entreprise prévisible, gérable et haute performance requiert de l'espace libre sur les disques pour la gestion des données et des métadonnées. La quantité d'espace libre requise dépend du type de disque utilisé et des processus métier.

ASA r2 utilise des zones de disponibilité de stockage (SAZ) au lieu d'agrégats, mais le principe reste le même : l'espace libre comprend toute capacité physique non consommée par les données réelles, les instantanés ou la surcharge du système. Il faut également tenir compte du provisionnement fin : les allocations logiques ne reflètent pas l'utilisation physique réelle.

Les recommandations de NetApp concernant les systèmes de stockage ASA r2 utilisés pour les applications d'entreprise sont les suivantes :

Pools SSD dans les systèmes ASA r2

* NetApp recommande* de maintenir un minimum de 10 % d'espace physique libre dans les environnements ASA r2. Cette directive s'applique aux pools composés uniquement de SSD utilisés par les systèmes ASA r2 et inclut tout l'espace inutilisé dans la SAZ et les unités de stockage. L'espace logique n'a pas d'importance ; l'essentiel est de se concentrer sur l'espace physique libre réellement disponible pour le stockage des données.

Bien que ASA r2 puisse supporter une utilisation élevée sans dégradation des performances, un fonctionnement proche de sa pleine capacité augmente le risque d'épuisement de l'espace et les frais administratifs lors de l'extension du stockage. Un taux d'utilisation supérieur à 90 % peut ne pas avoir d'impact

sur les performances, mais peut compliquer la gestion et retarder la mise à disposition de disques supplémentaires.

Les systèmes ASA r2 prennent en charge des unités de stockage jusqu'à 128 To et des tailles SAZ jusqu'à 2 Po par paire HA, avec ONTAP qui équilibre automatiquement la capacité entre les nœuds. Il est essentiel de surveiller l'utilisation au niveau du cluster, de la SAZ et de l'unité de stockage afin de garantir un espace libre suffisant pour les instantanés, les charges de travail à provisionnement fin et la croissance future. Si la capacité approche des seuils critiques (~ 90 % d'utilisation), des SSD supplémentaires doivent être ajoutés par groupes (minimum six disques) pour maintenir les performances et la résilience.

Ordinateurs virtuels de stockage

La gestion du stockage de la base de données Oracle sur les systèmes ASA r2 est également centralisée sur une machine virtuelle de stockage (SVM), connue sous le nom de vserver dans l'interface de ligne de commande ONTAP .

Une SVM est l'unité fondamentale de provisionnement et de sécurité du stockage dans ONTAP , similaire à une VM invitée sur un serveur VMware ESX. Lors de sa première installation sur ASA r2, ONTAP ne dispose d'aucune capacité de service de données jusqu'à la création d'une SVM. Le SVM définit la personnalité et les services de données pour l'environnement SAN.

Les systèmes ASA r2 utilisent une interface ONTAP exclusivement SAN, simplifiée pour prendre en charge les protocoles de blocs (FC, iSCSI, NVMe/FC, NVMe/TCP) et qui supprime les fonctionnalités liées au NAS. Cela simplifie la gestion et garantit que toutes les configurations SVM sont optimisées pour les charges de travail SAN. Contrairement aux systèmes AFF/ FAS , ASA r2 n'expose pas d'options pour les services NAS tels que les répertoires personnels ou les partages NFS.

Lors de la création d'un cluster, ASA r2 provisionne automatiquement une SVM de données par défaut nommée svm1 avec les protocoles SAN activés. Cette SVM est prête pour les opérations de stockage par blocs sans nécessiter de configuration manuelle des services de protocole. Par défaut, les LIF de données IP de cette SVM prennent en charge les protocoles iSCSI et NVMe/TCP et utilisent la politique de service default-data-blocks, ce qui simplifie la configuration initiale des charges de travail SAN. Les administrateurs peuvent ultérieurement créer des SVM supplémentaires ou personnaliser les configurations LIF en fonction des exigences de performance, de sécurité ou de multi-locataires.

> Les interfaces logiques (LIF) pour les protocoles SAN doivent être conçues en fonction des exigences de performance et de disponibilité. ASA r2 prend en charge les LIF iSCSI, FC et NVMe, mais notez que le basculement automatique des LIF iSCSI n'est pas activé par défaut car ASA r2 utilise un réseau partagé pour les hôtes NVMe et SCSI. Pour activer le basculement automatique, créez "LIF iSCSI uniquement".

SVM

Comme pour les autres plateformes ONTAP , il n'existe pas de recommandation officielle concernant le nombre de SVM à créer ; la décision dépend des exigences de gestion et de sécurité.

La plupart des clients utilisent un seul SVM principal pour leurs opérations quotidiennes et créent des SVM supplémentaires pour des besoins spécifiques, tels que :

- Une SVM dédiée pour une base de données d'entreprise critique, gérée par une équipe de spécialistes.
- Un SVM pour un groupe de développement avec contrôle administratif délégué
- Un SVM pour les données sensibles nécessitant un accès administratif restreint

Dans les environnements multi-locataires, chaque locataire peut se voir attribuer une SVM dédiée. La limite du nombre de SVM et de LIF par cluster, paire HA et nœud dépend du protocole utilisé, du modèle de nœud et de la version d' ONTAP. Consultez le "NetApp Hardware Universe" pour ces limites.

ASA r2 prend en charge jusqu'à 256 SVM par cluster et par paire HA à partir d' ONTAP 9.18.1 (auparavant 32 dans les versions précédentes).

Gestion des performances avec ONTAP QoS sur les systèmes ASA r2

La gestion sûre et efficace de plusieurs bases de données Oracle sur ASA r2 nécessite une stratégie QoS efficace. Ceci est particulièrement important car les systèmes ASA r2 sont des plateformes SAN tout flash conçues pour des performances extrêmement élevées et la consolidation des charges de travail.

Un nombre relativement restreint de SSD peut saturer même les contrôleurs les plus puissants ; les contrôles QoS sont donc essentiels pour garantir des performances prévisibles sur de multiples charges de travail. À titre de référence, les systèmes ASA r2 tels que l' ASA A1K ou l'A90 peuvent fournir des centaines de milliers à plus d'un million d'IOPS avec une latence inférieure à la milliseconde. Très peu de charges de travail uniques consommeraient ce niveau de performance ; une utilisation optimale implique donc généralement l'hébergement de plusieurs bases de données ou applications. Pour ce faire en toute sécurité, des politiques QoS sont nécessaires afin d'éviter la contention des ressources.

La QoS ONTAP sur ASA r2 fonctionne de la même manière que sur les systèmes AFF/ FAS , avec deux types de contrôles principaux : IOPS et bande passante. Les contrôles QoS peuvent être appliqués aux SVM et aux LUN.

QoS des IOPS

La QoS basée sur les IOPS limite le nombre total d'IOPS pour une ressource donnée. Dans ASA r2, les politiques QoS peuvent être appliquées au niveau SVM et à des objets de stockage individuels tels que les LUN. Lorsqu'une charge de travail atteint sa limite d'IOPS, des requêtes d'E/S supplémentaires sont mises en file d'attente pour les jetons, ce qui introduit une latence. Il s'agit d'un comportement normal qui empêche toute charge de travail unique de monopoliser les ressources du système.

Soyez prudent lorsque vous appliquez des contrôles QoS aux données de journalisation des transactions/restauration de la base de données. Ces charges de travail sont irrégulières, et une limite de QoS qui semble raisonnable pour une activité moyenne peut être trop basse lors des pics de charge, entraînant de graves problèmes de performance. En général, la journalisation des modifications et des archives ne devrait pas être limitée par la QoS.

QoS de la bande passante

La QoS basée sur la bande passante limite le débit en Mbps. Ceci est utile lorsque les charges de travail effectuent des lectures ou des écritures de blocs importants, comme des analyses complètes de tables ou des opérations de sauvegarde, qui consomment une bande passante importante mais relativement peu d'IOPS. La combinaison des limites d'IOPS et de bande passante permet un contrôle plus précis.

QoS minimale/garantie

Les politiques QoS minimales réservent les performances aux charges de travail critiques. Par exemple, dans un environnement mixte avec des bases de données de production et de développement, appliquez une QoS maximale aux charges de travail de développement et une QoS minimale aux charges de travail de production

afin de garantir des performances prévisibles.

La QoS adaptative

La QoS adaptative ajuste les limites en fonction de la taille de l'objet de stockage. Bien que rarement utilisée pour les bases de données (car la taille n'est pas corrélée aux besoins de performance), elle peut s'avérer utile pour les charges de travail de virtualisation où les exigences de performance évoluent avec la capacité.

Efficacité

Les fonctionnalités d'optimisation de l'espace ONTAP sont entièrement prises en charge et optimisées pour les systèmes ASA r2. Dans la quasi-totalité des cas, la meilleure solution consiste à conserver les paramètres par défaut tout en activant toutes les options d'optimisation des performances.

Les systèmes ASA r2 sont des plateformes SAN entièrement flash, c'est pourquoi les technologies d'efficacité telles que la compression, la compaction et la déduplication sont essentielles pour maximiser la capacité utilisable et réduire les coûts.

Compression

La compression réduit l'espace requis en encodant des motifs dans les données. Avec les systèmes ASA r2 basés sur SSD, la compression permet de réaliser des économies importantes car la mémoire flash élimine le besoin de surdimensionnement pour obtenir des performances optimales. La compression adaptative ONTAP est activée par défaut et a été testée de manière approfondie avec des charges de travail d'entreprise, y compris des bases de données Oracle, sans impact mesurable sur les performances, même dans des environnements où la latence est mesurée en microsecondes. Dans certains cas, les performances s'améliorent car les données compressées occupent moins d'espace cache.

L'efficacité de stockage sensible à la température (TSSE) n'est pas appliquée sur les systèmes ASA r2. Sur les systèmes ASA r2, la compression ne repose pas sur les données chaudes (fréquemment consultées) ou les données froides (rarement consultées). La compression démarre sans attendre que les données soient froides.

Compression adaptative

La compression adaptative utilise par défaut une taille de bloc de 8 Ko, correspondant à la taille de bloc couramment utilisée par les bases de données relationnelles. Des tailles de blocs plus importantes (16 Ko ou 32 Ko) peuvent améliorer l'efficacité des données séquentielles telles que les journaux de transactions ou les sauvegardes, mais doivent être utilisées avec prudence pour les bases de données actives afin d'éviter les surcharges lors des réécritures.

La taille des blocs peut être augmentée jusqu'à 32 Ko pour les fichiers inactifs tels que les journaux ou les sauvegardes. Consultez les instructions de NetApp avant de modifier les paramètres par défaut.

N'utilisez pas la compression 32 Ko avec déduplication pour les sauvegardes en continu. Utilisez une compression de 8 Ko pour maintenir l'efficacité de la déduplication.

Alignement de compression

L'alignement de la compression est important pour les réécritures aléatoires. Assurez-vous que le type de LUN, le décalage de partition (multiple de 8 Ko) et la taille de bloc du système de fichiers sont correctement alignés sur la taille de bloc de la base de données. Les données séquentielles telles que les sauvegardes ou les journaux ne nécessitent pas de considérations d'alignement.

Compaction

La compaction complète la compression en permettant à plusieurs blocs compressés de partager le même bloc physique. Par exemple, si un bloc de 8 Ko est compressé à 1 Ko, la compaction garantit que l'espace restant n'est pas gaspillé. Cette fonctionnalité est intégrée et n'entraîne aucune perte de performance.

Déduplication

La déduplication supprime les blocs dupliqués dans les ensembles de données. Bien que les bases de données Oracle n'offrent généralement que des économies minimales en matière de déduplication en raison des en-têtes et des pieds de page uniques des blocs, la déduplication ONTAP peut tout de même récupérer de l'espace à partir de blocs mis à zéro et de modèles répétés.

Efficacité et provisionnement fin

Les systèmes ASA r2 utilisent le provisionnement fin par défaut. Les fonctionnalités d'efficacité complètent le provisionnement fin pour maximiser la capacité utilisable.

> Les unités de stockage sont toujours provisionnées de manière fine sur les systèmes de stockage ASA r2. Le provisionnement épais n'est pas pris en charge.

Technologie QuickAssist (QAT)

Dans les plateformes NetApp ASA r2, la technologie Intel QuickAssist (QAT) offre une efficacité accélérée par le matériel qui diffère considérablement de l'efficacité du stockage sensible à la température (TSSE) basée sur le logiciel sans QAT.

QAT avec accélération matérielle :

- Décharge les tâches de compression et de chiffrement des cœurs du processeur.
- Permet une efficacité immédiate et intégrée pour les données chaudes (fréquemment consultées) et les données froides (rarement consultées).
- Réduit considérablement la charge du processeur.
- Offre un débit plus élevé et une latence plus faible.
- Améliore l'évolutivité des opérations sensibles aux performances telles que le chiffrement TLS et VPN.

TSSE sans QAT :

- Repose sur des processus pilotés par le processeur pour un fonctionnement efficace.
- L'efficacité n'est appliquée qu'aux données froides après un délai.
- Consomme davantage de ressources du processeur.
- Limite les performances globales par rapport aux systèmes accélérés par QAT.

Les systèmes ASA r2 modernes offrent donc une efficacité accrue grâce à l'accélération matérielle et une

meilleure utilisation du système que les anciennes plateformes TSSE uniquement.

Meilleures pratiques d'efficacité pour ASA r2

NetApp recommande ce qui suit :

Valeurs par défaut de ASA r2

Les unités de stockage créées sur ONTAP exécuté sur des systèmes ASA r2 sont provisionnées de manière fine avec toutes les fonctionnalités d'efficacité en ligne activées par défaut, y compris la compression, le compactage et la déduplication. Bien que les bases de données Oracle ne bénéficient généralement pas de manière significative de la déduplication et puissent inclure des données non compressibles, ces paramètres par défaut conviennent à la quasi-totalité des charges de travail. ONTAP est conçu pour traiter efficacement tous les types de données et de modèles d'E/S, qu'ils génèrent ou non des économies. Les valeurs par défaut ne doivent être modifiées que si les raisons sont parfaitement comprises et s'il existe un avantage clair à s'en écarter.

Recommandations générales

- Désactiver la compression pour les données chiffrées ou compressées par l'application : si les fichiers sont déjà compressés au niveau de l'application ou chiffrés, désactivez la compression pour optimiser les performances et permettre un fonctionnement plus efficace sur d'autres unités de stockage.
- Évitez de combiner de grands blocs de compression avec la déduplication : n'utilisez pas à la fois la compression de 32 Ko et la déduplication pour les sauvegardes de bases de données. Pour les sauvegardes en continu, utilisez une compression de 8 Ko afin de maintenir l'efficacité de la déduplication.
- Suivi des gains d'efficacité : utilisez les outils ONTAP (System Manager, Active IQ) pour suivre les économies d'espace réelles et ajuster les politiques si nécessaire.

Provisionnement fin

Le provisionnement fin d'une base de données Oracle sur ASA r2 nécessite une planification minutieuse car il implique la configuration d'un espace logique supérieur à celui physiquement disponible. Lorsqu'elle est correctement mise en œuvre, l'allocation dynamique de ressources permet de réaliser d'importantes économies et d'améliorer la facilité de gestion.

Le provisionnement fin fait partie intégrante d' ASA r2 et est étroitement lié aux technologies d'efficacité ONTAP , car les deux permettent de stocker plus de données logiques que la capacité physique du système. Les systèmes ASA r2 sont exclusivement SAN, et le provisionnement fin s'applique aux unités de stockage et aux LUN au sein des zones de disponibilité de stockage (SAZ).

Les unités de stockage ASA r2 sont provisionnées dynamiquement par défaut.

Presque toutes les utilisations des snapshots impliquent un provisionnement fin. Par exemple, une base de données classique de 10 Tio avec 30 jours d'instantanés peut apparaître comme 310 Tio de données logiques, mais seulement 12 à 15 Tio d'espace physique sont consommés car les instantanés ne stockent que les blocs modifiés.

De même, le clonage est une autre forme d'approvisionnement léger. Un environnement de développement avec 40 clones d'une base de données de 80 Tio nécessiterait 3,2 PiB s'il était entièrement écrit, mais en pratique il consomme beaucoup moins car seules les modifications sont stockées.

Gestion de l'espace

Il convient d'être prudent avec le provisionnement fin dans un environnement applicatif car les taux de modification des données peuvent augmenter de manière inattendue. Par exemple, la consommation d'espace due aux instantanés peut augmenter rapidement si les tables de base de données sont réindexées ou si des correctifs à grande échelle sont appliqués aux machines virtuelles VMware. Une sauvegarde mal placée peut écrire une grande quantité de données en très peu de temps. Enfin, il peut être difficile de récupérer certaines applications si un LUN vient à manquer d'espace libre de manière inattendue.

Dans ASA r2, ces risques sont atténués grâce à **l'allocation dynamique**, **la surveillance proactive** et **les politiques de redimensionnement des LUN**, plutôt qu'à des fonctionnalités ONTAP comme l'extension automatique des volumes ou la suppression automatique des instantanés. Les administrateurs doivent :

- Activer le provisionnement fin sur les LUN (`space-reserve disabled`) - il s'agit du paramètre par défaut dans ASA r2
- Surveillez la capacité à l'aide des alertes du gestionnaire système ou de l'automatisation basée sur une API
- Utilisez le redimensionnement LUN planifié ou scripté pour accompagner la croissance
- Configurer la réserve de snapshots et la suppression automatique des snapshots via System Manager (GUI)

Une planification minutieuse des seuils d'espace et des scripts d'automatisation est essentielle car ASA r2 ne prend pas en charge la croissance automatique des volumes ni la suppression des instantanés via l'interface de ligne de commande.

ASA r2 n'utilise pas de paramètres de réserve fractionnaire car il s'agit d'une architecture SAN uniquement qui abstrait les options de volume basées sur WAFL. L'efficacité de l'espace et la protection contre l'écrasement sont gérées au niveau du LUN. Par exemple, si vous disposez d'un LUN de 250 Gio provisionné à partir d'une unité de stockage, les snapshots consomment de l'espace en fonction des modifications réelles des blocs plutôt que de réserver une quantité d'espace égale à l'avance. Cela élimine le besoin de réservations statiques importantes, qui étaient courantes dans les environnements ONTAP traditionnels utilisant la réserve fractionnaire.

Si une protection contre l'écrasement garanti est requise et que la surveillance n'est pas possible, les administrateurs doivent prévoir une capacité suffisante dans l'unité de stockage et configurer la réserve de snapshots en conséquence. Cependant, la conception de ASA r2 rend la réserve fractionnaire inutile pour la plupart des charges de travail.

Compression et déduplication

La compression et la déduplication dans ASA r2 sont des technologies d'optimisation de l'espace, et non des mécanismes de provisionnement fin traditionnels. Ces fonctionnalités réduisent l'encombrement physique du stockage en éliminant les données redondantes et en compressant les blocs, ce qui permet de stocker davantage de données logiques que ne le permettrait la capacité brute.

Par exemple, un ensemble de données de 50 Tio peut être compressé à 30 Tio, économisant ainsi 20 Tio d'espace physique. Du point de vue de l'application, il reste 50 Tio de données, même si elles n'occupent que 30 Tio sur le disque.

La compressibilité d'un ensemble de données peut évoluer au fil du temps, ce qui peut augmenter l'espace physique consommé. Par conséquent, la compression et la déduplication doivent être gérées de manière proactive grâce à une surveillance et une planification des capacités.

Allocation d'espace libre et d'espace LVM

Le provisionnement fin dans les environnements ASA r2 peut perdre en efficacité au fil du temps si les blocs supprimés ne sont pas récupérés. À moins que l'espace ne soit libéré à l'aide de TRIM/UNMAP ou écrasé avec des zéros (via ASMRU - Automatic Space Management and Reclamation Utility), les données supprimées continuent de consommer de la capacité physique. Dans de nombreux environnements de bases de données Oracle, le provisionnement fin offre un avantage limité car les fichiers de données sont généralement pré-alloués à leur taille maximale lors de leur création.

Une planification minutieuse de la configuration LVM peut améliorer l'efficacité et minimiser le besoin de provisionnement de stockage et de redimensionnement des LUN. Lorsqu'un LVM tel que Veritas VxVM ou Oracle ASM est utilisé, les LUN sous-jacents sont divisés en extensions qui ne sont utilisées qu'en cas de besoin. Par exemple, si un ensemble de données commence à 2 Tio mais peut atteindre 10 Tio au fil du temps, cet ensemble de données pourrait être placé sur 10 Tio de LUN à provisionnement fin organisés dans un groupe de disques LVM. Il n'occuperait que 2 Tio d'espace au moment de sa création et ne réclamerait d'espace supplémentaire qu'à mesure que des extensions seraient allouées pour absorber la croissance des données. Ce procédé est sûr tant que l'espace est surveillé.

basculement ONTAP

Une bonne compréhension des fonctions de reprise de stockage est nécessaire pour garantir que les opérations de la base de données Oracle ne soient pas interrompues pendant ces opérations. De plus, les arguments utilisés lors des opérations de rachat peuvent affecter l'intégrité des données s'ils sont utilisés incorrectement.

Dans des conditions normales, les écritures entrantes destinées à un contrôleur donné sont répliquées de manière synchrone sur son partenaire HA. Dans un environnement ASA r2 avec SnapMirror Active Sync (SM-as), les écritures sont également répliquées sur un contrôleur distant du site secondaire. Tant qu'une écriture n'est pas enregistrée sur un support non volatil à tous les emplacements, elle n'est pas confirmée à l'application hôte.

Le support de stockage des données écrites est appelé mémoire non volatile (NVMEM). On l'appelle parfois mémoire vive non volatile (NVRAM) et on peut la considérer comme un journal d'écriture plutôt que comme un cache. En fonctionnement normal, les données stockées sur la NVMEM ne sont pas lues ; elle sert uniquement à protéger les données en cas de panne logicielle ou matérielle. Lorsque des données sont écrites sur des disques, elles sont transférées depuis la RAM système, et non depuis la NVMEM.

Lors d'une opération de prise de contrôle, un nœud d'une paire HA prend le relais des opérations de son partenaire. Dans ASA r2, le basculement n'est pas applicable car MetroCluster n'est pas pris en charge ; à la place, SnapMirror Active Sync assure la redondance au niveau du site. Les opérations de prise de contrôle du stockage lors de la maintenance de routine doivent être transparentes, hormis une brève interruption des opérations lors du changement des chemins réseau. La mise en réseau peut être complexe et les erreurs sont faciles à commettre ; c'est pourquoi NetApp recommande fortement de tester minutieusement les opérations de reprise avant de mettre un système de stockage en production. C'est le seul moyen de garantir que tous les chemins réseau sont correctement configurés. Dans un environnement SAN, vérifiez l'état du chemin à l'aide de la commande `sanlun lun show -p` ou les outils de gestion de chemins multiples natifs du système d'exploitation pour garantir la disponibilité de tous les chemins attendus. Les systèmes ASA r2 fournissent tous les chemins optimisés actifs pour les LUN, et les clients utilisant des espaces de noms NVMe doivent

s'appuyer sur des outils natifs du système d'exploitation, car les chemins NVMe ne sont pas couverts par sanlun.

Il convient d'être prudent lors du déclenchement d'une prise de contrôle forcée. Forcer une modification de la configuration de stockage signifie que l'état du contrôleur propriétaire des disques est ignoré et que le nœud alternatif prend de force le contrôle des disques. Un forçage incorrect d'une prise de contrôle peut entraîner une perte ou une corruption de données, car une prise de contrôle forcée peut supprimer le contenu de la NVMEM. Une fois la prise de contrôle terminée, la perte de ces données signifie que les données stockées sur les disques pourraient revenir à un état légèrement antérieur du point de vue de la base de données.

Une prise de contrôle forcée avec une paire HA normale devrait rarement être nécessaire. Dans la quasi-totalité des scénarios de panne, un nœud s'arrête et en informe le partenaire afin qu'un basculement automatique ait lieu. Il existe certains cas particuliers, comme une panne en cascade où l'interconnexion entre les nœuds est perdue puis un contrôleur tombe en panne, nécessitant une prise de contrôle forcée. Dans une telle situation, la réplication entre les nœuds est perdue avant la défaillance du contrôleur, ce qui signifie que le contrôleur survivant ne dispose plus d'une copie des écritures en cours. Il faut alors forcer le rachat, ce qui signifie qu'il y a potentiellement un risque de perte de données.

NetApp recommande de prendre les précautions suivantes :

- Veillez à ne pas provoquer accidentellement une prise de contrôle. Normalement, il ne devrait pas être nécessaire de forcer la modification, et forcer le changement peut entraîner une perte de données.
- Si une prise de contrôle forcée est nécessaire, assurez-vous que les applications sont arrêtées, que tous les systèmes de fichiers sont démontés et que les groupes de volumes LVM (Logical Volume Manager) sont désactivés. Les groupes de disques ASM doivent être démontés.
- En cas de défaillance au niveau du site lors de l'utilisation de SM-as, le basculement automatique non planifié assisté par ONTAP Mediator sera lancé sur le cluster survivant, ce qui entraînera une brève pause d'E/S, puis les transitions de base de données reprendront à partir du cluster survivant. Pour plus d'informations, consultez le "Synchronisation active SnapMirror sur les systèmes ASA r2" pour les étapes de configuration détaillées.

Configuration de la base de données avec les systèmes AFF/ FAS

Tailles de bloc

ONTAP utilise en interne une taille de bloc variable, ce qui signifie que les bases de données Oracle peuvent être configurées avec n'importe quelle taille de bloc. Cependant, la taille des blocs du système de fichiers peut affecter les performances et, dans certains cas, une taille de bloc de reprise supérieure peut améliorer les performances.

Tailles des blocs de fichiers de données

Certains systèmes d'exploitation offrent un choix de tailles de blocs de système de fichiers. Pour les systèmes de fichiers prenant en charge les fichiers de données Oracle, la taille de bloc doit être de 8 Ko lorsque la compression est utilisée. Lorsque la compression n'est pas requise, vous pouvez utiliser une taille de bloc de 8 Ko ou 4 Ko.

Si un fichier de données est placé sur un système de fichiers avec un bloc de 512 octets, des fichiers mal

alignés sont possibles. Il est possible que le LUN et le système de fichiers soient correctement alignés en fonction des recommandations de NetApp, mais les E/S de fichier sont mal alignées. Un tel mauvais alignement entraînerait de graves problèmes de performances.

Les systèmes de fichiers prenant en charge les journaux de reprise doivent utiliser une taille de bloc qui représente un multiple de la taille de bloc de reprise. Cela nécessite généralement que le système de fichiers redo log et le fichier redo log lui-même utilisent une taille de bloc de 512 octets.

Rétablir les tailles des blocs

Avec des taux de reprise très élevés, il est possible que des tailles de bloc de 4 Ko soient plus performantes, car les taux de reprise élevés permettent d'exécuter les E/S en moins d'opérations et de manière plus efficace. Si les taux de reprise sont supérieurs à 50 Mbit/s, envisagez de tester une taille de bloc de 4 Ko.

Quelques problèmes clients ont été identifiés avec les bases de données à l'aide de journaux de reprise avec une taille de bloc de 512 octets sur un système de fichiers d'une taille de bloc de 4 Ko et de nombreuses transactions très petites. La surcharge liée à l'application de plusieurs modifications de 512 octets à un seul bloc du système de fichiers de 4 Ko a entraîné des problèmes de performances qui ont été résolus en changeant le système de fichiers pour qu'il utilise une taille de bloc de 512 octets.

NetApp vous recommande de ne pas modifier la taille du bloc de reprise, sauf si un service client ou un service professionnel vous en informe ou si le changement est basé sur la documentation officielle du produit.

db_file_multibloc_read_count

Le `db_file_multiblock_read_count` Paramètre contrôle le nombre maximal de blocs de base de données Oracle lus par Oracle au cours d'une opération, pendant les E/S séquentielles

Toutefois, ce paramètre n'affecte pas le nombre de blocs lus par Oracle au cours des opérations de lecture, ni le nombre d'E/S aléatoires Seule la taille de bloc des E/S séquentielles est affectée.

Oracle recommande à l'utilisateur de ne pas définir ce paramètre. Cela permet au logiciel de base de données de définir automatiquement la valeur optimale. Cela signifie généralement que ce paramètre est défini sur une valeur qui produit une taille d'E/S de 1 Mo. Par exemple, une lecture de 1 Mo de blocs de 8 Ko nécessite la lecture de 128 blocs. La valeur par défaut de ce paramètre est donc 128.

La plupart des problèmes de performance de base de données observés par NetApp sur les sites des clients provenaient de paramètres incorrects. Des raisons valides ont été données pour modifier cette valeur avec les versions 8 et 9 d'Oracle. Par conséquent, le paramètre peut être présent sans le savoir dans `init.ora` Fichiers car la base de données a été mise à niveau vers Oracle 10 et versions ultérieures. La configuration héritée de 8 ou 16, par rapport à la valeur par défaut 128, nuit de manière significative aux performances d'E/S séquentielles.

NetApp recommande de régler le `db_file_multiblock_read_count` le paramètre ne doit pas être présent dans le `init.ora` fichier. NetApp n'a jamais observé d'amélioration des performances suite à la modification de ce paramètre, mais le débit d'E/S séquentielles subit une importante dégradation dans de nombreux cas.

filesytemio_options

Le paramètre d'initialisation Oracle `filesystemio_options` Contrôle l'utilisation des E/S asynchrones et directes

Contrairement à une idée reçue, ces deux types d'E/S ne s'excluent pas mutuellement. NetApp a observé que ce paramètre est souvent mal configuré dans les environnements des clients. Cette configuration incorrecte est la cause directe de nombreux problèmes de performances.

Les E/S asynchrones offrent la possibilité de paralléliser les opérations Oracle d'E/S. Avant la disponibilité des E/S asynchrones sur différents systèmes d'exploitation, les utilisateurs ont configuré de nombreux processus dbwriter et modifié la configuration du processus serveur. Avec les E/S asynchrones, le système d'exploitation lui-même exécute les E/S en parallèle pour le compte du logiciel de base de données. Ce processus ne présente aucun risque pour les données et les opérations critiques, telles que la journalisation de reprise Oracle, sont toujours exécutées de manière synchrone.

Les E/S directes contournent le cache du tampon du système d'exploitation. Sur un système UNIX, les E/S transitent normalement par le cache du tampon du système d'exploitation. Ceci est utile pour les applications qui ne maintiennent pas de cache interne, mais Oracle dispose de son propre cache de tampon dans la SGA. Dans la plupart des cas, il est préférable d'activer les E/S directes et d'allouer la RAM du serveur à la mémoire SGA plutôt que d'utiliser le cache du tampon du système d'exploitation. La SGA exploite la mémoire plus efficacement. En outre, lors de leur transit via le tampon du se, les E/S sont soumises à un traitement supplémentaire, ce qui augmente les latences. Cette augmentation est particulièrement visible lors des E/S intenses en écriture, pour lesquelles la faible latence est primordiale.

Les options pour `filesystemio_options` sont :

- **Async.** Oracle soumet des demandes d'E/S au système d'exploitation pour traitement. Ce qui lui permet d'effectuer d'autres tâches plutôt que d'attendre la fin des E/S et d'augmenter ainsi la parallélisation des E/S.
- **Directio.** Oracle effectue des E/S directement par rapport aux fichiers physiques plutôt que de router les E/S via le cache du système d'exploitation hôte.
- **None.** Oracle utilise des E/S synchrones et mises en tampon Dans cette configuration, le choix entre les processus serveur partagés et dédiés et le nombre de dbwriter est plus important.
- **Setall.** Oracle utilise des E/S asynchrones et directes Dans presque tous les cas, l'utilisation de `setall` est optimale.

> Le `filesystemio_options` Ce paramètre n'a aucun effet dans les environnements dNFS et ASM. Dans ces environnements, les E/S asynchrones et directes sont automatiquement utilisées

Certains clients ont déjà rencontré des problèmes d'E/S asynchrones, notamment avec les versions précédentes de Red Hat Enterprise Linux 4 (RHEL4). Certains conseils obsolètes sur Internet suggèrent toujours d'éviter les E/S asynchrones en raison d'informations obsolètes. Les E/S asynchrones sont stables sur tous les systèmes d'exploitation actuels. Il n'y a aucune raison de le désactiver, en l'absence d'un bug connu avec le système d'exploitation.

Si une base de données utilise des E/S mises en tampon, un switch vers des E/S directes peut également justifier une modification de la taille de la mémoire SGA. La désactivation des E/S mises en tampon élimine le gain de performance fourni par le cache du se hôte pour la base de données. L'ajout de RAM à la SGA résout ce problème. Et devrait améliorer les performances nettes d'E/S.

Bien qu'il soit presque toujours préférable d'utiliser la RAM pour la SGA d'Oracle plutôt que pour le cache du tampon du système d'exploitation, il peut s'avérer impossible de déterminer ce qui est le plus avantageux. Par exemple, il est parfois préférable d'utiliser des E/S mises en tampon avec une mémoire SGA de très petite taille sur un serveur de base de données comportant de nombreuses instances Oracle actives par intermittence. Cette configuration permet à toutes les instances de base de données en cours d'exécution d'utiliser de manière flexible la RAM restante sur le système d'exploitation. Cette situation est très inhabituelle, mais elle a été observée sur certains sites clients.

NetApp recommande le réglage `filesystemio_options` à `setall`, Mais notez que dans certains cas, la perte du cache du tampon hôte peut nécessiter une augmentation de la SGA d'Oracle.

Délais d'expiration du RAC

Oracle RAC est un produit clusterware qui comporte plusieurs types de processus de pulsation internes qui contrôlent l'intégrité du cluster.

Les informations dans le "misscount" La section contient des informations essentielles pour les environnements RAC Oracle utilisant un stockage en réseau. Dans la plupart des cas, les paramètres RAC Oracle par défaut devront être modifiés pour garantir que le cluster RAC résiste aux modifications de chemin réseau et aux opérations de basculement/basculement du stockage.

disktimeout

Le paramètre RAC principal lié au stockage est `disktimeout`. Ce paramètre contrôle le seuil au sein duquel les E/S du fichier de vote doivent être terminées. Si le `disktimeout` Le paramètre est dépassé, puis le nœud RAC est supprimé du cluster. La valeur par défaut de ce paramètre est 200. Cette valeur doit être suffisante pour les procédures standard de Takeover et and Giveback du stockage.

NetApp recommande fortement de tester soigneusement les configurations RAC avant de les mettre en production, car de nombreux facteurs affectent un basculement ou un rétablissement. Outre le temps nécessaire au basculement du stockage, la propagation des modifications du protocole LACP (Link Aggregation Control Protocol) nécessite également du temps supplémentaire. En outre, le logiciel de chemins d'accès multiples SAN doit détecter un délai d'expiration d'E/S et réessayer sur un autre chemin. Si une base de données est extrêmement active, une grande quantité d'E/S doit être mise en file d'attente et relancée avant le traitement des E/S du disque de vote.

En l'absence d'un basculement ou d'un retour de stockage réel, l'effet peut être simulé à l'aide de tests de câble Pull sur le serveur de base de données.

NetApp recommande ce qui suit :

- En quittant le `disktimeout` paramètre à la valeur par défaut de 200.
- Testez toujours soigneusement une configuration RAC.

misscount

Le `misscount` Le paramètre affecte normalement uniquement la pulsation réseau entre les nœuds RAC. La valeur par défaut est 30 secondes. Si les binaires de la grille se trouvent sur une matrice de stockage ou si le disque d'amorçage du système d'exploitation n'est pas local, ce paramètre peut devenir important. Cela inclut

les hôtes avec des lecteurs de démarrage situés sur un SAN FC, les systèmes d'exploitation démarrés par NFS et les lecteurs de démarrage situés sur les datastores de virtualisation, tels qu'un fichier VMDK.

Si l'accès à un disque de démarrage est interrompu par un basculement ou un rétablissement du stockage, il est possible que l'emplacement binaire de la grille ou l'ensemble du système d'exploitation soit temporairement bloqué. Le temps nécessaire à ONTAP pour terminer l'opération de stockage et au système d'exploitation pour changer les chemins et reprendre les E/S peut être supérieur à `misscount` seuil. Par conséquent, un nœud est immédiatement supprimé une fois la connectivité à la LUN de démarrage ou aux binaires de la grille restaurée. Dans la plupart des cas, l'exclusion et le redémarrage qui s'ensuit se produisent sans message de journalisation indiquant la raison du redémarrage. Toutes les configurations ne sont pas affectées. Testez donc tout hôte de démarrage SAN, de démarrage NFS ou basé sur un datastore dans un environnement RAC afin que RAC reste stable si la communication avec le lecteur de démarrage est interrompue.

Dans le cas de lecteurs de démarrage non locaux ou d'un système de fichiers non local hébergeant `grid` binaires, le `misscount` devra être modifié pour correspondre `disktimeout`. Si ce paramètre est modifié, effectuez des tests supplémentaires pour identifier également les effets sur le comportement du RAC, tels que le temps de basculement du nœud.

NetApp recommande ce qui suit :

- Quittez le `misscount` paramètre à la valeur par défaut de 30, sauf si l'une des conditions suivantes s'applique :
  - `grid` Les fichiers binaires sont situés sur un disque connecté au réseau, y compris les disques basés sur NFS, iSCSI, FC et les datastores.
  - Le système d'exploitation est démarré sur un SAN.
- Dans de tels cas, évaluez l'effet des interruptions de réseau qui affectent l'accès au système d'exploitation ou `GRID_HOME` systèmes de fichiers. Dans certains cas, de telles interruptions provoquent le blocage des démons RAC Oracle, ce qui peut conduire à un `misscount`délai d'expiration et suppression basés sur. Le délai par défaut est de 27 secondes, soit la valeur de `misscount` moins `reboottime`. Dans de tels cas, augmenter `misscount` à 200 pour correspondre `disktimeout`.

Configuration de la base de données avec les systèmes ASA r2

Tailles de bloc

ONTAP utilise en interne une taille de bloc variable, ce qui signifie que les bases de données Oracle peuvent être configurées avec n'importe quelle taille de bloc souhaitée. Cependant, la taille des blocs du système de fichiers peut affecter les performances, et dans certains cas, une taille de bloc de journalisation plus importante peut améliorer les performances.

ASA r2 n'introduit aucun changement dans les recommandations de taille de bloc Oracle par rapport aux systèmes AFF/ FAS . Le comportement ONTAP reste cohérent sur toutes les plateformes.

Tailles des blocs de fichiers de données

Certains systèmes d'exploitation offrent un choix de tailles de blocs de système de fichiers. Pour les systèmes de fichiers prenant en charge les fichiers de données Oracle, la taille de bloc doit être de 8 Ko lorsque la compression est utilisée. Lorsque la compression n'est pas requise, vous pouvez utiliser une taille de bloc de 8 Ko ou 4 Ko.

Si un fichier de données est placé sur un système de fichiers avec un bloc de 512 octets, des fichiers mal alignés sont possibles. Il est possible que le LUN et le système de fichiers soient correctement alignés en fonction des recommandations de NetApp, mais les E/S de fichier sont mal alignées. Un tel mauvais alignement entraînerait de graves problèmes de performances.

Rétablir les tailles des blocs

Les systèmes de fichiers prenant en charge les journaux de reprise doivent utiliser une taille de bloc qui représente un multiple de la taille de bloc de reprise. Cela nécessite généralement que le système de fichiers redo log et le fichier redo log lui-même utilisent une taille de bloc de 512 octets.

Avec des taux de reprise très élevés, il est possible que des tailles de bloc de 4 Ko soient plus performantes, car les taux de reprise élevés permettent d'exécuter les E/S en moins d'opérations et de manière plus efficace. Si les taux de reprise sont supérieurs à 50 Mbit/s, envisagez de tester une taille de bloc de 4 Ko.

Quelques problèmes clients ont été identifiés avec les bases de données à l'aide de journaux de reprise avec une taille de bloc de 512 octets sur un système de fichiers d'une taille de bloc de 4 Ko et de nombreuses transactions très petites. La surcharge liée à l'application de plusieurs modifications de 512 octets à un seul bloc du système de fichiers de 4 Ko a entraîné des problèmes de performances qui ont été résolus en changeant le système de fichiers pour qu'il utilise une taille de bloc de 512 octets.

NetApp vous recommande de ne pas modifier la taille du bloc de reprise, sauf si un service client ou un service professionnel vous en informe ou si le changement est basé sur la documentation officielle du produit.

db_file_multibloc_read_count

Le `db_file_multiblock_read_count` Paramètre contrôle le nombre maximal de blocs de base de données Oracle lus par Oracle au cours d'une opération, pendant les E/S séquentielles

Les recommandations restent inchangées par rapport aux systèmes AFF/ FAS . Le comportement ONTAP et les meilleures pratiques d'Oracle restent identiques sur les plateformes ASA r2, AFF et FAS .

Toutefois, ce paramètre n'affecte pas le nombre de blocs lus par Oracle au cours des opérations de lecture, ni le nombre d'E/S aléatoires Seule la taille de bloc des E/S séquentielles est affectée.

Oracle recommande à l'utilisateur de ne pas définir ce paramètre. Cela permet au logiciel de base de données de définir automatiquement la valeur optimale. Cela signifie généralement que ce paramètre est défini sur une valeur qui produit une taille d'E/S de 1 Mo. Par exemple, une lecture de 1 Mo de blocs de 8 Ko nécessite la lecture de 128 blocs. La valeur par défaut de ce paramètre est donc 128.

La plupart des problèmes de performance de base de données observés par NetApp sur les sites des clients provenaient de paramètres incorrects. Des raisons valides ont été données pour modifier cette valeur avec les versions 8 et 9 d'Oracle. Par conséquent, le paramètre peut être présent sans le savoir dans `init.ora` Fichiers car la base de données a été mise à niveau vers Oracle 10 et versions ultérieures. La configuration

héritée de 8 ou 16, par rapport à la valeur par défaut 128, nuit de manière significative aux performances d'E/S séquentielles.

NetApp recommande de régler le `db_file_multiblock_read_count` le paramètre ne doit pas être présent dans le `init.ora` fichier. NetApp n'a jamais observé d'amélioration des performances suite à la modification de ce paramètre, mais le débit d'E/S séquentielles subit une importante dégradation dans de nombreux cas.

filesytemio_options

Le paramètre d'initialisation Oracle `filesystemio_options` Contrôle l'utilisation des E/S asynchrones et directes

Le comportement et les recommandations concernant filesystemio_options sur ASA r2 sont identiques à ceux des systèmes AFF/ FAS , car ce paramètre est spécifique à Oracle et ne dépend pas de la plateforme de stockage. ASA r2 utilise ONTAP comme AFF/ FAS, donc les mêmes bonnes pratiques s'appliquent.

Contrairement à une idée reçue, ces deux types d'E/S ne s'excluent pas mutuellement. NetApp a observé que ce paramètre est souvent mal configuré dans les environnements des clients. Cette configuration incorrecte est la cause directe de nombreux problèmes de performances.

Les E/S asynchrones offrent la possibilité de paralléliser les opérations Oracle d'E/S. Avant la disponibilité des E/S asynchrones sur différents systèmes d'exploitation, les utilisateurs ont configuré de nombreux processus dbwriter et modifié la configuration du processus serveur. Avec les E/S asynchrones, le système d'exploitation lui-même exécute les E/S en parallèle pour le compte du logiciel de base de données. Ce processus ne présente aucun risque pour les données et les opérations critiques, telles que la journalisation de reprise Oracle, sont toujours exécutées de manière synchrone.

Les E/S directes contournent le cache du tampon du système d'exploitation. Sur un système UNIX, les E/S transitent normalement par le cache du tampon du système d'exploitation. Ceci est utile pour les applications qui ne maintiennent pas de cache interne, mais Oracle dispose de son propre cache de tampon dans la SGA. Dans la plupart des cas, il est préférable d'activer les E/S directes et d'allouer la RAM du serveur à la mémoire SGA plutôt que d'utiliser le cache du tampon du système d'exploitation. La SGA exploite la mémoire plus efficacement. En outre, lors de leur transit via le tampon du se, les E/S sont soumises à un traitement supplémentaire, ce qui augmente les latences. Cette augmentation est particulièrement visible lors des E/S intenses en écriture, pour lesquelles la faible latence est primordiale.

Les options pour `filesystemio_options` sont :

- **Async.** Oracle soumet des demandes d'E/S au système d'exploitation pour traitement. Ce qui lui permet d'effectuer d'autres tâches plutôt que d'attendre la fin des E/S et d'augmenter ainsi la parallélisation des E/S.
- **Directio.** Oracle effectue des E/S directement par rapport aux fichiers physiques plutôt que de router les E/S via le cache du système d'exploitation hôte.
- **None.** Oracle utilise des E/S synchrones et mises en tampon Dans cette configuration, le choix entre les processus serveur partagés et dédiés et le nombre de dbwriter est plus important.
- **Setall.** Oracle utilise des E/S asynchrones et directes Dans presque tous les cas, l'utilisation de `setall` est optimale.

Dans les environnements ASM, Oracle utilise automatiquement les E/S directes et asynchrones pour les disques gérés par ASM. `filesystemio_options` n'a aucun effet sur les groupes de disques ASM. Pour les déploiements non-ASM (par exemple, les systèmes de fichiers sur des LUN SAN), configurez : `filesystemio_options = setall`. Cela permet des E/S asynchrones et directes pour des performances optimales.

Certains systèmes d'exploitation plus anciens présentaient des problèmes avec les entrées/sorties asynchrones, ce qui a conduit à des conseils obsolètes suggérant de les éviter. Cependant, les E/S asynchrones sont stables et entièrement prises en charge par tous les systèmes d'exploitation actuels. Il n'y a aucune raison de le désactiver à moins qu'un bug spécifique du système d'exploitation ne soit identifié.

Si une base de données utilise des E/S mises en tampon, un switch vers des E/S directes peut également justifier une modification de la taille de la mémoire SGA. La désactivation des E/S mises en tampon élimine le gain de performance fourni par le cache du se hôte pour la base de données. L'ajout de RAM à la SGA résout ce problème. Et devrait améliorer les performances nettes d'E/S.

Bien qu'il soit presque toujours préférable d'utiliser la RAM pour la SGA d'Oracle plutôt que pour le cache du tampon du système d'exploitation, il peut s'avérer impossible de déterminer ce qui est le plus avantageux. Par exemple, il est parfois préférable d'utiliser des E/S mises en tampon avec une mémoire SGA de très petite taille sur un serveur de base de données comportant de nombreuses instances Oracle actives par intermittence. Cette configuration permet à toutes les instances de base de données en cours d'exécution d'utiliser de manière flexible la RAM restante sur le système d'exploitation. Cette situation est très inhabituelle, mais elle a été observée sur certains sites clients.

* NetApp recommande* de paramétrer `filesystemio_options` à `setall`, mais sachez que dans certaines circonstances, la perte du cache tampon de l'hôte peut nécessiter une augmentation de la SGA d'Oracle. Les systèmes ASA r2 sont optimisés pour les charges de travail SAN à faible latence, l'utilisation de setall s'aligne donc parfaitement avec la conception ASA pour les déploiements Oracle hautes performances.

Délais d'expiration du RAC

Oracle RAC est un produit clusterware qui comporte plusieurs types de processus de pulsation internes qui contrôlent l'intégrité du cluster.

Les systèmes ASA r2 utilisent ONTAP tout comme AFF/ FAS, donc les mêmes principes s'appliquent aux paramètres de délai d'expiration d'Oracle RAC. Il n'existe aucune modification spécifique à ASA concernant les recommandations relatives aux délais d'expiration de disque ou au nombre d'erreurs. Cependant, ASA r2 est optimisée pour les charges de travail SAN et le basculement à faible latence, ce qui rend ces bonnes pratiques encore plus importantes.

Les informations contenues dans le "misscount" Cette section contient des informations essentielles pour les environnements Oracle RAC utilisant un stockage en réseau, et dans de nombreux cas, les paramètres par défaut d'Oracle RAC devront être modifiés pour garantir que le cluster RAC survive aux changements de chemin réseau et aux opérations de basculement du stockage.

disktimeout

Le paramètre RAC principal lié au stockage est `disktimeout`. Ce paramètre contrôle le seuil au sein duquel les E/S du fichier de vote doivent être terminées. Si le `disktimeout` Le paramètre est dépassé, puis le nœud RAC est supprimé du cluster. La valeur par défaut de ce paramètre est 200. Cette valeur doit être suffisante

pour les procédures standard de Takeover et and Giveback du stockage.

NetApp recommande fortement de tester soigneusement les configurations RAC avant de les mettre en production, car de nombreux facteurs affectent un basculement ou un rétablissement. Outre le temps nécessaire au basculement du stockage, la propagation des modifications du protocole LACP (Link Aggregation Control Protocol) nécessite également du temps supplémentaire. En outre, le logiciel de chemins d'accès multiples SAN doit détecter un délai d'expiration d'E/S et réessayer sur un autre chemin. Si une base de données est extrêmement active, une grande quantité d'E/S doit être mise en file d'attente et relancée avant le traitement des E/S du disque de vote.

En l'absence d'un basculement ou d'un retour de stockage réel, l'effet peut être simulé à l'aide de tests de câble Pull sur le serveur de base de données.

NetApp recommande ce qui suit :

- En quittant le `disktimeout` paramètre à la valeur par défaut de 200.
- Testez toujours soigneusement une configuration RAC.

misscount

Le `misscount` Le paramètre affecte normalement uniquement la pulsation réseau entre les nœuds RAC. La valeur par défaut est 30 secondes. Si les binaires de la grille se trouvent sur une matrice de stockage ou si le disque d'amorçage du système d'exploitation n'est pas local, ce paramètre peut devenir important. Cela inclut les hôtes avec des lecteurs de démarrage situés sur un SAN FC, les systèmes d'exploitation démarrés par NFS et les lecteurs de démarrage situés sur les datastores de virtualisation, tels qu'un fichier VMDK.

Si l'accès à un disque de démarrage est interrompu par un basculement ou un rétablissement du stockage, il est possible que l'emplacement binaire de la grille ou l'ensemble du système d'exploitation soit temporairement bloqué. Le temps nécessaire à ONTAP pour terminer l'opération de stockage et au système d'exploitation pour changer les chemins et reprendre les E/S peut être supérieur à `misscount` seuil. Par conséquent, un nœud est immédiatement supprimé une fois la connectivité à la LUN de démarrage ou aux binaires de la grille restaurée. Dans la plupart des cas, l'exclusion et le redémarrage qui s'ensuit se produisent sans message de journalisation indiquant la raison du redémarrage. Toutes les configurations ne sont pas affectées. Testez donc tout hôte de démarrage SAN, de démarrage NFS ou basé sur un datastore dans un environnement RAC afin que RAC reste stable si la communication avec le lecteur de démarrage est interrompue.

Dans le cas de lecteurs de démarrage non locaux ou d'un système de fichiers non local hébergeant `grid` binaires, le `misscount` devra être modifié pour correspondre `disktimeout`. Si ce paramètre est modifié, effectuez des tests supplémentaires pour identifier également les effets sur le comportement du RAC, tels que le temps de basculement du nœud.

NetApp recommande ce qui suit :

- Quittez le `misscount` paramètre à la valeur par défaut de 30, sauf si l'une des conditions suivantes s'applique :
  - `grid` Les fichiers binaires sont situés sur un lecteur réseau, notamment des lecteurs iSCSI, FC et des lecteurs basés sur un système de stockage de données.
  - Le système d'exploitation est démarré sur un SAN.
- Dans de tels cas, évaluez l'effet des interruptions de réseau qui affectent l'accès au système d'exploitation ou `GRID_HOME` systèmes de fichiers. Dans certains cas, de telles interruptions provoquent le blocage des démons RAC Oracle, ce qui peut conduire à un `misscount`délai d'expiration et suppression basés sur. Le délai par défaut est de 27 secondes, soit la valeur de `misscount` moins `reboottime`. Dans de tels cas, augmenter `misscount` à 200 pour correspondre `disktimeout`.

- La conception optimisée pour SAN de ASA r2 réduit la latence de basculement, mais les délais d'attente doivent toujours être ajustés pour le démarrage en réseau ou les binaires de grille.
- Pour les configurations RAC étendues ou actives-actives (par exemple, la synchronisation active SnapMirror), le réglage du délai d'expiration reste essentiel pour les architectures à RPO nul.

Configuration de l'hôte avec les systèmes AFF/ FAS

AIX

Rubriques de configuration pour la base de données Oracle sous IBM AIX avec ONTAP.

E/S simultanées

Pour obtenir des performances optimales sur IBM AIX, il est nécessaire d'utiliser des E/S simultanées Sans E/S simultanées, les limites de performances sont probablement dues au fait qu'AIX exécute des E/S atomiques sérialisées, ce qui entraîne une surcharge importante.

À l'origine, NetApp a recommandé d'utiliser le `cio` Option de montage pour forcer l'utilisation d'E/S simultanées sur le système de fichiers, mais ce processus présente des inconvénients et n'est plus nécessaire. Depuis l'introduction d'AIX 5.2 et d'Oracle 10gR1, Oracle sous AIX peut ouvrir des fichiers individuels pour des E/S simultanées, au lieu de forcer des E/S simultanées sur l'ensemble du système de fichiers.

La meilleure méthode pour activer les E/S simultanées est de définir le `init.ora` paramètre `filesystemio_options` à `setall`. Oracle peut ainsi ouvrir des fichiers spécifiques pour une utilisation avec des E/S simultanées

À l'aide de `cio` En tant qu'option de montage, force l'utilisation d'E/S simultanées, ce qui peut avoir des conséquences négatives. Par exemple, forcer des E/S simultanées désactive la lecture anticipée sur les systèmes de fichiers, ce qui peut nuire aux performances des E/S se produisant en dehors du logiciel de base de données Oracle, comme la copie de fichiers et les sauvegardes sur bande. En outre, les produits tels qu'Oracle GoldenGate et SAP BR*Tools ne sont pas compatibles avec l'utilisation du `cio` Option de montage avec certaines versions d'Oracle.

NetApp recommande ce qui suit :

- N'utilisez pas le `cio` option de montage au niveau du système de fichiers. Activez plutôt les E/S simultanées via l'utilisation de `filesystemio_options=setall`.
- Utilisez uniquement le `cio` l'option de montage doit être définie si elle n'est pas possible `filesystemio_options=setall`.

Options de montage NFS AIX

Le tableau suivant répertorie les options de montage NFS AIX pour les bases de données Oracle à instance unique.

Type de fichier	Options de montage
Accueil ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Fichiers de contrôle Fichiers de données Journaux de reprise	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
`ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, intr`

Le tableau suivant répertorie les options de montage NFS AIX pour RAC.

Type de fichier	Options de montage
Accueil ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Fichiers de contrôle Fichiers de données Journaux de reprise	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac`
`CRS/Voting`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac`
Ressource dédiée `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Partagée `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr`

La principale différence entre les options de montage à instance unique et RAC est l'ajout de `noac` aux options de montage. Cet ajout a pour effet de désactiver la mise en cache du système d'exploitation hôte qui permet à toutes les instances du cluster RAC d'avoir une vue cohérente de l'état des données.

En utilisant le `cio` option de montage et `init.ora` paramètre `filesystemio_options=setall` a le même effet que la désactivation de la mise en cache de l'hôte, il est toujours nécessaire de l'utiliser `noac`. `noac` est requis pour le partage `ORACLE_HOME` Déploiements pour faciliter la cohérence des fichiers tels que les fichiers

de mots de passe Oracle et `spfile` fichiers de paramètres. Si chaque instance d'un cluster RAC possède un dédié `ORACLE_HOME`, ce paramètre n'est pas requis.

Options de montage AIX jfs/jfs2

Le tableau suivant répertorie les options de montage AIX jfs/jfs2.

Type de fichier	Options de montage
Accueil ADR	Valeurs par défaut
Fichiers de contrôle Fichiers de données Journaux de reprise	Valeurs par défaut
ORACLE_HOME	Valeurs par défaut

Avant d'utiliser AIX `hdisk` dans tout environnement, y compris les bases de données, vérifiez le paramètre `queue_depth`. Ce paramètre n'est pas la profondeur de la file d'attente HBA ; il se rapporte plutôt à la profondeur de la file d'attente SCSI de l'individu `hdisk device. Depending on how the LUNs are configured, the value for `queue_depth` peut être trop faible pour de bonnes performances. Les tests ont montré que la valeur optimale est de 64.

HP-UX

Rubriques de configuration pour la base de données Oracle sur HP-UX avec ONTAP.

Options de montage NFS HP-UX

Le tableau suivant répertorie les options de montage NFS HP-UX pour une seule instance.

Type de fichier	Options de montage
Accueil ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`
Fichiers de contrôle Fichiers de données Journaux de reprise	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144,forc edirectio, nointr,suid`
`ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`

Le tableau suivant répertorie les options de montage NFS HP-UX pour RAC.

Type de fichier	Options de montage
Accueil ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, noac,suid`

Type de fichier	Options de montage
Fichiers de contrôle Fichiers de données Journaux de reprise	`rw, bg,hard, [vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,forcedirectio,suid`
CRS/vote	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac, forcedirectio,suid`
Ressource dédiée `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`
Partagée `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,suid`

La principale différence entre les options de montage à instance unique et RAC est l'ajout de `noac` et `forcedirectio` aux options de montage. Cet ajout a pour effet de désactiver la mise en cache du système d'exploitation hôte, ce qui permet à toutes les instances du cluster RAC d'avoir une vue cohérente de l'état des données. En utilisant le `init.ora` paramètre `filesystemio_options=setall` a le même effet que la désactivation de la mise en cache de l'hôte, il est toujours nécessaire de l'utiliser `noac` et `forcedirectio`.

La raison `noac` est requis pour le partage `ORACLE_HOME` Les déploiements visent à faciliter la cohérence des fichiers tels que les fichiers de mots de passe Oracle et les fichiers spfiles. Si chaque instance d'un cluster RAC possède un dédié `ORACLE_HOME`, ce paramètre n'est pas requis.

Options de montage HP-UX VxFS

Utilisez les options de montage suivantes pour les systèmes de fichiers hébergeant les binaires Oracle :

```
delaylog,nodatainlog
```

Utilisez les options de montage suivantes pour les systèmes de fichiers contenant des fichiers de données, des journaux de reprise, des journaux d'archivage et des fichiers de contrôle dans lesquels la version de HP-UX ne prend pas en charge les E/S simultanées :

```
nodatainlog,mincache=direct,convosync=direct
```

Lorsque des E/S simultanées sont prises en charge (VxFS 5.0.1 et versions ultérieures, ou avec ServiceGuard Storage Management Suite), utilisez ces options de montage pour les systèmes de fichiers contenant des fichiers de données, des journaux de reprise, des journaux d'archivage et des fichiers de contrôle :

```
delaylog,cio
```

Le paramètre `db_file_multiblock_read_count` Est particulièrement critique dans les environnements VxFS. Oracle recommande que ce paramètre ne soit pas défini dans Oracle 10g R1 et versions ultérieures, sauf indication contraire. La taille de bloc Oracle de 8 Ko par défaut est 128. Si la valeur de ce paramètre est forcée à 16 ou moins, retirer le `convosync=direct` Option de montage car elle peut endommager les performances des E/S séquentielles. Cette étape nuit à d'autres aspects de la performance et ne doit être prise que si la valeur de `db_file_multiblock_read_count` doit être modifié par rapport à la valeur par défaut.

Linux

Rubriques de configuration spécifiques au système d'exploitation Linux.

Tables d'emplacements TCP Linux NFSv3

Les tables d'emplacements TCP sont l'équivalent NFSv3 de la profondeur de file d'attente de l'adaptateur de bus hôte (HBA). Ces tableaux contrôlent le nombre d'opérations NFS qui peuvent être en attente à la fois. La valeur par défaut est généralement 16, un chiffre bien trop faible pour assurer des performances optimales. Le problème inverse se produit sur les noyaux Linux plus récents : la limite de la table des emplacements TCP augmente automatiquement par envoi de demandes, jusqu'à atteindre le niveau de saturation du serveur NFS.

Pour des performances optimales et pour éviter les problèmes de performances, ajustez les paramètres du noyau qui contrôlent les tables d'emplacements TCP.

Exécutez le `sysctl -a | grep tcp.*.slot_table` et observez les paramètres suivants :

```
# sysctl -a | grep tcp.*.slot_table
sunrpc.tcp_max_slot_table_entries = 128
sunrpc.tcp_slot_table_entries = 128
```

Tous les systèmes Linux doivent inclure `sunrpc.tcp_slot_table_entries`, mais seulement certains incluent `sunrpc.tcp_max_slot_table_entries`. Ils doivent tous deux être réglés sur 128.

Si vous ne définissez pas ces paramètres, vous risquez d'avoir des effets importants sur les performances. Dans certains cas, les performances sont limitées car le système d'exploitation linux n'émet pas suffisamment d'E/S. Dans d'autres cas, les latences d'E/S augmentent à mesure que le système d'exploitation linux tente d'émettre plus d'E/S que ce qui peut être traité.

Options de montage NFS Linux

Le tableau suivant répertorie les options de montage NFS Linux pour une seule instance.

Type de fichier	Options de montage
Accueil ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Fichiers de contrôle Fichiers de données Journaux de reprise	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr`

Type de fichier	Options de montage
ORACLE_HOME	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr`

Le tableau suivant répertorie les options de montage NFS Linux pour RAC.

Type de fichier	Options de montage
Accueil ADR	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, actimeo=0`
Fichiers de contrôle Fichiers de données Journaux de reprise	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,actimeo=0`
CRS/vote	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,actimeo=0`
Ressource dédiée `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Partagée `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,actimeo=0`

La principale différence entre les options de montage à instance unique et RAC est l'ajout de `actimeo=0` aux options de montage. Cet ajout a pour effet de désactiver la mise en cache du système d'exploitation hôte, ce qui permet à toutes les instances du cluster RAC d'avoir une vue cohérente de l'état des données. En utilisant le `init.ora` paramètre `filesystemio_options=setall` a le même effet que la désactivation de la mise en cache de l'hôte, il est toujours nécessaire de l'utiliser `actimeo=0`.

La raison `actimeo=0` est requis pour le partage `ORACLE_HOME` Les déploiements visent à faciliter la cohérence des fichiers tels que les fichiers de mots de passe Oracle et les fichiers spfiles. Si chaque instance d'un cluster RAC possède un dédié `ORACLE_HOME`, ce paramètre n'est pas requis.

En règle générale, les fichiers ne provenant pas de bases de données doivent être montés avec les mêmes options que celles utilisées pour les fichiers de données à instance unique. Toutefois, certaines applications peuvent avoir des exigences différentes. Évitez les options de montage `noac` et `actimeo=0` si possible parce que ces options désactivent la lecture et la mise en mémoire tampon au niveau du système de fichiers. Cela peut entraîner de graves problèmes de performances pour les processus tels que l'extraction, la translation et le chargement.

ACCESS et GETATTR

Certains clients ont remarqué qu'un niveau extrêmement élevé d'autres IOPS, comme L'ACCÈS et GETATTR, peut dominer leurs charges de travail. Dans des cas extrêmes, les opérations telles que les lectures et les écritures peuvent représenter jusqu'à 10 % du total. Il s'agit d'un comportement normal avec toute base de données qui inclut l'utilisation de `actimeo=0` et/ou `noac` Sous Linux car ces options font que le système d'exploitation Linux recharge en permanence les métadonnées de fichiers à partir du système de stockage. Les opérations telles que ACCESS et GETATTR sont des opérations à faible impact qui sont traitées à partir du cache ONTAP dans un environnement de base de données. Elles ne doivent pas être considérées comme

des IOPS authentiques, comme les lectures et les écritures, qui génèrent une véritable demande pour les systèmes de stockage. Cependant, ces autres IOPS créent une certaine charge, en particulier dans les environnements RAC. Pour résoudre ce problème, activez dNFS, qui contourne le cache du tampon du système d'exploitation et évite ces opérations de métadonnées inutiles.

NFS direct Linux

Une option de montage supplémentaire, appelée `nosharecache`, Est requis lorsque (a) dNFS est activé et (b) qu'un volume source est monté plusieurs fois sur un seul serveur (c) avec un montage NFS imbriqué. Cette configuration est principalement utilisée dans les environnements prenant en charge les applications SAP. Par exemple, un seul volume sur un système NetApp peut avoir un répertoire situé sur `/vol/oracle/base` et une seconde à `/vol/oracle/home`. Si `/vol/oracle/base` est monté à `/oracle` et `/vol/oracle/home` est monté à `/oracle/home`, Le résultat est des montages NFS imbriqués qui proviennent de la même source.

Le système d'exploitation peut détecter le fait que `/oracle` et `/oracle/home` résident sur le même volume, qui est le même système de fichiers source. Le système d'exploitation utilise ensuite le même descripteur de périphérique pour accéder aux données. Cela améliore l'utilisation de la mise en cache du système d'exploitation et de certaines autres opérations, mais interfère avec dNFS. Si dNFS doit accéder à un fichier, par exemple `spfile` , sur , `/oracle/home` il peut essayer par erreur d'utiliser le mauvais chemin d'accès aux données. Le résultat est une opération d'E/S défaillante. Dans ces configurations, ajoutez `nosharecache` l'option de montage à tout système de fichiers NFS qui partage un volume source avec un autre système de fichiers NFS de cet hôte. Cela force le système d'exploitation Linux à allouer un descripteur de périphérique indépendant pour ce système de fichiers.

Linux Direct NFS et Oracle RAC

DNFS présente des avantages spéciaux en matière de performances pour Oracle RAC sur le système d'exploitation Linux. En effet, Linux ne dispose pas d'une méthode permettant de forcer les E/S directes, qui est requise avec RAC pour assurer la cohérence entre les nœuds. Pour contourner ce problème, Linux nécessite l'utilisation du `actimeo=0` Mount option, qui entraîne l'expiration immédiate des données de fichier à partir du cache du système d'exploitation. Cette option force à son tour le client Linux NFS à relire en permanence les données d'attributs, ce qui endommage la latence et augmente la charge sur le contrôleur de stockage.

L'activation de dNFS contourne le client NFS hôte et évite ces dommages. Plusieurs clients ont signalé une amélioration significative des performances sur les clusters RAC et une baisse significative de la charge ONTAP (en particulier par rapport aux autres IOPS) lors de l'activation de dNFS.

Linux Direct NFS et fichier orangfstab

Si vous utilisez dNFS sur Linux avec l'option de chemins d'accès multiples, vous devez utiliser plusieurs sous-réseaux. Sur d'autres systèmes d'exploitation, vous pouvez établir plusieurs canaux dNFS à l'aide du `LOCAL` et `DONTROUTE` Options de configuration de plusieurs canaux dNFS sur un même sous-réseau. Cependant, cela ne fonctionne pas correctement sur Linux et des problèmes de performances inattendus peuvent survenir. Sous Linux, chaque carte réseau utilisée pour le trafic dNFS doit se trouver sur un sous-réseau différent.

Planificateur d'E/S.

Le noyau Linux permet un contrôle de bas niveau sur la façon dont les E/S sont planifiées pour bloquer les périphériques. Les valeurs par défaut sur les différentes distributions de Linux varient considérablement. Les tests montrent que la date limite offre habituellement les meilleurs résultats, mais il arrive que le NOOP ait été légèrement meilleur. La différence de performance est minime, mais testez les deux options s'il est nécessaire d'extraire les performances maximales d'une configuration de base de données. Dans de nombreuses configurations, le paramètre CFQ est le paramètre par défaut. Il a démontré des problèmes de performances

significatifs avec les charges de travail de la base de données.

Pour plus d'informations sur la configuration du planificateur d'E/S, reportez-vous à la documentation du fournisseur Linux correspondant.

Chemins d'accès multiples

Certains clients ont rencontré des pannes durant une interruption du réseau, car le démon multivoie ne s'exécutait pas sur leur système. Sur les versions récentes de Linux, le processus d'installation du système d'exploitation et le démon de chemins d'accès multiples peuvent exposer ces systèmes d'exploitation à ce problème. Les packages sont installés correctement, mais ils ne sont pas configurés pour un démarrage automatique après un redémarrage.

Par exemple, la valeur par défaut du démon multiacheminement sur RHEL5.5 peut apparaître comme suit :

```
[root@host1 iscsi]# chkconfig --list | grep multipath
multipathd      0:off   1:off   2:off   3:off   4:off   5:off   6:off
```

Ceci peut être corrigé à l'aide des commandes suivantes :

```
[root@host1 iscsi]# chkconfig multipathd on
[root@host1 iscsi]# chkconfig --list | grep multipath
multipathd      0:off   1:off   2:on    3:on    4:on    5:on    6:off
```

Mise en miroir ASM

La mise en miroir ASM peut nécessiter des modifications des paramètres de chemins d'accès multiples Linux pour permettre à ASM de reconnaître un problème et de basculer vers un autre groupe de pannes. La plupart des configurations ASM sur ONTAP reposent sur une redondance externe. La protection des données est assurée par la baie externe et ASM ne met pas en miroir les données. Certains sites utilisent ASM avec redondance normale pour fournir une mise en miroir bidirectionnelle, généralement entre différents sites.

Les paramètres Linux indiqués dans le "Documentation des utilitaires hôtes NetApp" Incluez les paramètres de chemins d'accès multiples qui entraînent une mise en file d'attente illimitée des E/S. Cela signifie qu'une E/S sur un périphérique LUN sans chemin d'accès actif attend tant que les E/S sont terminées. Cette opération est généralement souhaitable, car les hôtes Linux attendent tant que nécessaire la fin des modifications du chemin SAN, le redémarrage des commutateurs FC ou le basculement d'un système de stockage.

Ce comportement de mise en file d'attente illimité cause un problème de mise en miroir ASM car ASM doit recevoir une erreur d'E/S pour qu'il puisse réessayer d'E/S sur une autre LUN.

Définissez les paramètres suivants dans Linux `multipath.conf` Fichier pour les LUN ASM utilisés avec la mise en miroir ASM :

```
polling_interval 5
no_path_retry 24
```

Ces paramètres créent une temporisation de 120 secondes pour les périphériques ASM. Le délai d'attente est calculé comme étant le `polling_interval` * `no_path_retry` en secondes. Il peut être nécessaire

d'ajuster la valeur exacte dans certaines circonstances, mais un délai de 120 secondes doit être suffisant pour la plupart des utilisations. En particulier, 120 secondes doivent permettre un basculement ou un retour du contrôleur sans générer d'erreur d'E/S susceptible de mettre le groupe défaillant hors ligne.

Un plus bas `no_path_retry` La valeur peut réduire le temps nécessaire à ASM pour passer à un autre groupe de pannes, mais augmente également le risque de basculement indésirable lors des activités de maintenance, telles qu'une prise de contrôle. Le risque peut être atténué par une surveillance attentive de l'état de mise en miroir ASM. Si un basculement indésirable se produit, les miroirs peuvent être rapidement resynchronisés si la resynchronisation est effectuée relativement rapidement. Pour plus d'informations, consultez la documentation Oracle sur ASM Fast Mirror Resync pour la version du logiciel Oracle utilisé.

Options de montage Linux xfs, ext3 et ext4

NetApp recommande d'utiliser les options de montage par défaut.

ASMLib/AFD (pilote de filtre ASM)

Rubriques de configuration spécifiques au système d'exploitation Linux utilisant AFD et ASMlib

Tailles de bloc ASMlib

ASMlib est une bibliothèque de gestion ASM facultative et des utilitaires associés. Sa valeur principale est la capacité de tamponner un LUN ou un fichier NFS en tant que ressource ASM avec une étiquette lisible par l'utilisateur.

Les versions récentes d'ASMlib détectent un paramètre LUN appelé blocs logiques par exposant de bloc physique (LBPPBE). Cette valeur n'a été signalée que récemment par la cible SCSI ONTAP. Elle renvoie désormais une valeur qui indique qu'une taille de bloc de 4 Ko est recommandée. Il ne s'agit pas d'une définition de la taille de bloc, mais il est un indice pour toute application utilisant LBPPBE que les E/S d'une certaine taille peuvent être gérées plus efficacement. Cependant, ASMlib interprète LBPPBE comme une taille de bloc et estampille constamment l'en-tête ASM lors de la création du périphérique ASM.

Ce processus peut causer des problèmes avec les mises à niveau et les migrations de différentes manières, tous en fonction de l'incapacité à mélanger des périphériques ASMlib avec des tailles de bloc différentes dans le même groupe de disques ASM.

Par exemple, des tableaux plus anciens ont généralement signalé une valeur LBPPBE de 0 ou n'ont pas signalé cette valeur du tout. ASMlib l'interprète comme une taille de bloc de 512 octets. Pour les baies plus récentes, la taille de bloc est de 4 Ko. Il n'est pas possible de mélanger des périphériques de 512 octets et de 4 Ko dans le même groupe de disques ASM. Cela empêche un utilisateur d'augmenter la taille du groupe de disques ASM en utilisant des LUN de deux baies ou en utilisant ASM comme outil de migration. Dans d'autres cas, RMAN pourrait ne pas permettre la copie de fichiers entre un groupe de disques ASM avec une taille de bloc de 512 octets et un groupe de disques ASM avec une taille de bloc de 4 Ko.

La solution préférée est de corriger ASMlib. L'ID de bug Oracle est 13999609 et le correctif est présent dans oracleasm-support-2.1.8-1 et versions ultérieures. Ce correctif permet à un utilisateur de définir le paramètre `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` à `true` dans le `/etc/sysconfig/oracleasm` fichier de configuration. Cela empêche ASMlib d'utiliser le paramètre LBPPBE, ce qui signifie que les LUN de la nouvelle baie sont maintenant reconnues comme des périphériques de bloc de 512 octets.

L'option ne modifie pas la taille de bloc sur les LUN précédemment estampées par ASMlib. Par exemple, si un groupe de disques ASM avec des blocs de 512 octets doit être migré vers un nouveau système de stockage qui signale un bloc de 4 Ko, l'option `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` Doit être défini avant que les nouvelles LUN soient estampées avec ASMlib. Si les périphériques ont déjà été estampillés par oracleasm, ils doivent être reformatés avant d'être repoussées avec une nouvelle taille de bloc. Commencez par déconfigurer le périphérique avec `oracleasm deletedisk`, Puis effacez le premier 1 Go du périphérique avec `dd if=/dev/zero of=/dev/mapper/device bs=1048576 count=1024`. Enfin, si le périphérique a déjà été partitionné, utilisez le `kpartx` Commande permettant de supprimer les partitions obsolètes ou de simplement redémarrer le système d'exploitation.

Si ASMlib ne peut pas être corrigé, ASMlib peut être supprimé de la configuration. Ce changement est perturbateur et nécessite le démarquage des disques ASM et s'assurer que le `asm_diskstring` le paramètre est défini correctement. Toutefois, cette modification ne nécessite pas la migration des données.

Tailles de bloc d'entraînement de filtre ASM (AFD)

AFD est une bibliothèque de gestion ASM facultative qui remplace ASMlib. Du point de vue du stockage, il est très similaire à ASMlib, mais il inclut des fonctionnalités supplémentaires telles que la capacité de bloquer les E/S non-Oracle afin de réduire les risques d'erreurs d'utilisateur ou d'application susceptibles de corrompre les données.

Tailles des blocs de périphériques

Comme ASMlib, AFD lit également le paramètre LUN blocs logiques par exposant de bloc physique (LBPPBE) et utilise par défaut la taille de bloc physique, et non la taille de bloc logique.

Cela peut créer un problème si l'AFD est ajouté à une configuration existante où les périphériques ASM sont déjà formatés comme des périphériques de bloc de 512 octets. Le pilote AFD reconnaîtrait le LUN comme un périphérique 4K et l'incompatibilité entre l'étiquette ASM et le périphérique physique empêcherait l'accès. De même, les migrations seraient affectées, car il n'est pas possible de combiner des périphériques de 512 octets et de 4 Ko dans le même groupe de disques ASM. Cela empêche un utilisateur d'augmenter la taille du groupe de disques ASM en utilisant des LUN de deux baies ou en utilisant ASM comme outil de migration. Dans d'autres cas, RMAN pourrait ne pas permettre la copie de fichiers entre un groupe de disques ASM avec une taille de bloc de 512 octets et un groupe de disques ASM avec une taille de bloc de 4 Ko.

La solution est simple - AFD inclut un paramètre pour contrôler si elle utilise les tailles de bloc logiques ou physiques. Il s'agit d'un paramètre global affectant tous les périphériques du système. Pour forcer AFD à utiliser la taille de bloc logique, définissez `options oracleafd oracleafd_use_logical_block_size=1` dans le `/etc/modprobe.d/oracleafd.conf` fichier.

Tailles de transfert multivoie

Les modifications récentes du noyau linux appliquent des restrictions de taille d'E/S envoyées aux périphériques à chemins d'accès multiples, et AFD ne respecte pas ces restrictions. Les E/S sont ensuite rejetées, ce qui entraîne la mise hors ligne du chemin d'accès à la LUN. Il en résulte une incapacité à installer Oracle Grid, à configurer ASM ou à créer une base de données.

La solution consiste à spécifier manuellement la longueur de transfert maximale dans le fichier multipath.conf pour les LUN ONTAP :

```
devices {
            device {
                vendor "NETAPP"
                product "LUN.*"
                max_sectors_kb 4096
        }
    }
```

Même si aucun problème n'existe actuellement, ce paramètre doit être défini si l'AFD est utilisé pour garantir qu'une future mise à niveau de linux ne provoque pas de problèmes inattendus.

Microsoft Windows

Rubriques de configuration pour la base de données Oracle sous Microsoft Windows avec ONTAP.

NFS

Oracle prend en charge l'utilisation de Microsoft Windows avec le client NFS direct. Cette fonctionnalité offre les avantages de NFS en termes de gestion, notamment la possibilité d'afficher les fichiers dans les différents environnements, de redimensionner les volumes de façon dynamique et d'exploiter un protocole IP moins onéreux. Pour plus d'informations sur l'installation et la configuration d'une base de données sous Microsoft Windows à l'aide de dNFS, reportez-vous à la documentation officielle d'Oracle. Il n'existe pas de meilleures pratiques spéciales.

SAN

Pour une efficacité de compression optimale, assurez-vous que le système de fichiers NTFS utilise une unité d'allocation de 8 Ko ou plus. L'utilisation d'une unité d'allocation 4K, qui est généralement la valeur par défaut, a un impact négatif sur l'efficacité de la compression.

Solaris

Rubriques de configuration spécifiques au système d'exploitation Solaris.

Options de montage Solaris NFS

Le tableau suivant répertorie les options de montage Solaris NFS pour une seule instance.

Type de fichier	Options de montage
Accueil ADR	`rw,bg,hard,[vers=3,vers=4.1], roto=tcp, timeo=600,rsize=262144,wsize=262144`
Fichiers de contrôle Fichiers de données Journaux de reprise	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,llock,suid`

Type de fichier	Options de montage
ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,suid

L'utilisation de `llock` il a été prouvé qu'il améliorait considérablement les performances dans les environnements des clients en supprimant la latence associée à l'acquisition et au déblocage du système de stockage. Utilisez cette option avec soin dans les environnements dans lesquels de nombreux serveurs sont configurés pour monter les mêmes systèmes de fichiers et où Oracle est configuré pour monter ces bases de données. Bien qu'il s'agisse d'une configuration très inhabituelle, elle est utilisée par un petit nombre de clients. Si une instance est démarrée une seconde fois par erreur, une corruption des données peut se produire, car Oracle ne peut pas détecter les fichiers de verrouillage sur le serveur étranger. Les verrous NFS n'offrent pas de protection ; comme dans la version NFS 3, ils sont réservés à des conseils.

Parce que le `llock` et `forcedirectio` les paramètres s'excluent mutuellement, il est important que `filesystemio_options=setall` est présent dans le `init.ora` classez-les de sorte que `directio` est utilisé. Sans ce paramètre, la mise en cache du tampon du système d'exploitation hôte est utilisée et les performances peuvent être affectées.

Le tableau suivant répertorie les options de montage de Solaris NFS RAC.

Type de fichier	Options de montage
Accueil ADR	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,noac
Fichiers de contrôle Fichiers de données Journaux de reprise	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,forcedirectio
CRS/vote	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,forcedirectio
Ressource dédiée ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,suid
Partagée ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,suid

La principale différence entre les options de montage à instance unique et RAC est l'ajout de `noac` et `forcedirectio` aux options de montage. Cet ajout a pour effet de désactiver la mise en cache du système d'exploitation hôte, ce qui permet à toutes les instances du cluster RAC d'avoir une vue cohérente de l'état des données. En utilisant le `init.ora` paramètre `filesystemio_options=setall` a le même effet que la désactivation de la mise en cache de l'hôte, il est toujours nécessaire de l'utiliser `noac` et `forcedirectio`.

La raison `actimeo=0` est requis pour le partage `ORACLE_HOME` Les déploiements visent à faciliter la cohérence des fichiers tels que les fichiers de mots de passe Oracle et les fichiers spfiles. Si chaque instance d'un cluster RAC possède un dédié `ORACLE_HOME`, ce paramètre n'est pas requis.

Options de montage Solaris UFS

NetApp recommande fortement d'utiliser l'option de montage de journalisation afin de préserver l'intégrité des données en cas de panne de l'hôte Solaris ou d'interruption de la connectivité FC. L'option de montage de la journalisation préserve également l'utilisation des sauvegardes Snapshot.

ZFS Solaris

Solaris ZFS doit être installé et configuré avec soin pour offrir des performances optimales.

mvector

Solaris 11 a inclus un changement dans la façon dont il traite les opérations d'E/S importantes, ce qui peut entraîner de graves problèmes de performances sur les baies de stockage SAN. Le problème est documenté NetApp suivi bug report 630173, "Solaris 11 ZFS Performance regression".

Il ne s'agit pas d'un bogue de ONTAP. Il s'agit d'un défaut Solaris suivi sous les défauts Solaris 7199305 et 7082975.

Vous pouvez consulter le support Oracle pour savoir si votre version de Solaris 11 est affectée, ou vous pouvez tester la solution de contournement en la changeant `zfs_mvector_max_size` à une valeur plus petite.

Pour ce faire, exécutez la commande suivante en tant que root :

```
[root@host1 ~]# echo "zfs_mvector_max_size/W 0t131072" |mdb -kw
```

En cas de problème inattendu résultant de cette modification, vous pouvez facilement l'inverser en exécutant la commande suivante en tant que root :

```
[root@host1 ~]# echo "zfs_mvector_max_size/W 0t1048576" |mdb -kw
```

Noyau

Pour des performances ZFS fiables, un noyau Solaris est nécessaire pour résoudre les problèmes d'alignement des LUN. Le correctif a été introduit avec le correctif 147440-19 dans Solaris 10 et avec SRU 10.5 pour Solaris 11. Utilisez uniquement Solaris 10 et versions ultérieures avec ZFS.

Configuration du LUN

Pour configurer une LUN, effectuez les opérations suivantes :

1. Créer une LUN de type `solaris`.
2. Installez le kit d'utilitaire hôte (HUK) approprié spécifié par le "Matrice d'interopérabilité NetApp (IMT)".
3. Suivez les instructions du HUK exactement comme décrit. Les étapes de base sont décrites ci-dessous, mais reportez-vous au "documentation la plus récente" pour connaître la procédure adéquate.
    a. Exécutez le `host_config` utilitaire de mise à jour du `sd.conf/sdd.conf` fichier. Les disques SCSI seront ainsi en mesure de détecter correctement les LUN ONTAP.
    b. Suivez les instructions fournies par le `host_config` Utilitaire permettant d'activer les entrées/sorties

multivoies (MPIO).

c. Redémarrez. Cette étape est nécessaire pour que les modifications soient reconnues dans l'ensemble du système.

4. Partitionnez les LUN et vérifiez qu'ils sont correctement alignés. Voir « Annexe B : Vérification de l'alignement WAFL » pour obtenir des instructions sur la façon de tester et de confirmer directement l'alignement.

zpools

Un zpool ne doit être créé qu'après les étapes de la "Configuration du LUN" sont effectuées. Si la procédure n'est pas effectuée correctement, les performances risquent d'être sérieusement dégradées en raison de l'alignement des E/S. Pour des performances optimales sur ONTAP, les E/S doivent être alignées sur une limite de 4 Ko sur un disque. Les systèmes de fichiers créés sur un zpool utilisent une taille de bloc effective qui est contrôlée par un paramètre appelé `ashift`, qui peut être affiché en exécutant la commande `zdb -C`.

La valeur de `ashift` la valeur par défaut est 9, ce qui signifie 2^9, ou 512 octets. Pour des performances optimales, le `ashift` La valeur doit être 12 (2^12=4K). Cette valeur est définie au moment de la création du zpool et ne peut pas être modifiée, ce qui signifie que les données dans zpools avec `ashift` une migration autre que 12 doit être effectuée en copiant les données vers un nouveau zpool.

Après avoir créé un zpool, vérifiez la valeur de `ashift` avant de continuer. Si la valeur n'est pas 12, les LUN n'ont pas été détectées correctement. Détruisez le zpool, vérifiez que toutes les étapes indiquées dans la documentation des utilitaires hôtes correspondante ont été effectuées correctement et recréez le zpool.

Zpools et LDOMS Solaris

Les LDOMS Solaris créent une exigence supplémentaire pour s'assurer que l'alignement des E/S est correct. Bien qu'un LUN soit correctement découvert en tant que périphérique 4K, un périphérique virtuel vdsk sur un LDOM n'hérite pas de la configuration du domaine d'E/S. Le vdsk basé sur cette LUN revient par défaut à un bloc de 512 octets.

Un fichier de configuration supplémentaire est requis. Tout d'abord, les LDOM individuels doivent être corrigés pour le bogue Oracle 15824910 afin d'activer les options de configuration supplémentaires. Ce correctif a été porté dans toutes les versions actuellement utilisées de Solaris. Une fois le logiciel LDOM corrigé, il est prêt à configurer les nouveaux LUN correctement alignés comme suit :

1. Identifiez la ou les LUN à utiliser dans le nouveau zpool. Dans cet exemple, il s'agit du périphérique c2d1.

```
[root@LDOM1 ~]# echo | format
Searching for disks...done
AVAILABLE DISK SELECTIONS:
  0. c2d0 <Unknown-Unknown-0001-100.00GB>
     /virtual-devices@100/channel-devices@200/disk@0
  1. c2d1 <SUN-ZFS Storage 7330-1.0 cyl 1623 alt 2 hd 254 sec 254>
     /virtual-devices@100/channel-devices@200/disk@1
```

2. Récupérez l'instance vdc des systèmes à utiliser pour un pool ZFS :

```
[root@LDOM1 ~]#  cat /etc/path_to_inst
#
# Caution! This file contains critical kernel state
#
"/fcoe" 0 "fcoe"
"/iscsi" 0 "iscsi"
"/pseudo" 0 "pseudo"
"/scsi_vhci" 0 "scsi_vhci"
"/options" 0 "options"
"/virtual-devices@100" 0 "vnex"
"/virtual-devices@100/channel-devices@200" 0 "cnex"
"/virtual-devices@100/channel-devices@200/disk@0" 0 "vdc"
"/virtual-devices@100/channel-devices@200/pciv-communication@0" 0 "vpci"
"/virtual-devices@100/channel-devices@200/network@0" 0 "vnet"
"/virtual-devices@100/channel-devices@200/network@1" 1 "vnet"
"/virtual-devices@100/channel-devices@200/network@2" 2 "vnet"
"/virtual-devices@100/channel-devices@200/network@3" 3 "vnet"
"/virtual-devices@100/channel-devices@200/disk@1" 1 "vdc" << We want
this one
```

3. Modifier `/platform/sun4v/kernel/drv/vdc.conf`:

```
block-size-list="1:4096";
```

Cela signifie que l'instance de périphérique 1 se voit attribuer une taille de bloc de 4096.

Par exemple, supposons que les instances vdsk 1 à 6 doivent être configurées pour une taille de bloc de 4 Ko et `/etc/path_to_inst` se lit comme suit :

```
"/virtual-devices@100/channel-devices@200/disk@1" 1 "vdc"
"/virtual-devices@100/channel-devices@200/disk@2" 2 "vdc"
"/virtual-devices@100/channel-devices@200/disk@3" 3 "vdc"
"/virtual-devices@100/channel-devices@200/disk@4" 4 "vdc"
"/virtual-devices@100/channel-devices@200/disk@5" 5 "vdc"
"/virtual-devices@100/channel-devices@200/disk@6" 6 "vdc"
```

4. La finale `vdc.conf` le fichier doit contenir les éléments suivants :

```
block-size-list="1:8192","2:8192","3:8192","4:8192","5:8192","6:8192";
```

Avertissement
Le LDOM doit être redémarré après la configuration de vdc.conf et la création du vdsk. Cette étape ne peut pas être évitée. La modification de la taille de bloc n'est effective qu'après un redémarrage. Procéder à la configuration du pool de zpool et s'assurer que le module de transmission automatique est correctement réglé sur 12 comme décrit précédemment.

Journal des intentions ZFS (ZIL)

En général, il n'y a aucune raison de localiser le ZFS Intent Log (ZIL) sur un autre périphérique. Le journal peut partager de l'espace avec le pool principal. L'utilisation principale d'une ZIL distincte est l'utilisation de disques physiques qui n'offrent pas les fonctionnalités de mise en cache des écritures dans les baies de stockage modernes.

biais logique

Réglez le `logbias` Paramètre sur les systèmes de fichiers ZFS hébergeant les données Oracle.

```
zfs set logbias=throughput <filesystem>
```

Ce paramètre réduit les niveaux d'écriture globaux. Sous les valeurs par défaut, les données écrites sont d'abord validées dans le ZIL, puis dans le pool de stockage principal. Cette approche est adaptée à une configuration utilisant une configuration de disque simple, qui inclut un périphérique ZIL SSD et un support rotatif pour le pool de stockage principal. En effet, elle permet une validation dans une seule transaction d'E/S sur le support à latence la plus faible disponible.

Lorsque vous utilisez une baie de stockage moderne qui inclut sa propre capacité de mise en cache, cette approche n'est généralement pas nécessaire. Dans de rares cas, il peut être souhaitable d'effectuer une écriture avec une seule transaction dans le journal, par exemple une charge de travail composée d'écritures aléatoires hautement concentrées et sensibles à la latence. L'amplification d'écriture peut avoir des conséquences, car les données consignées sont finalement écrites dans le pool de stockage principal, ce qui double l'activité d'écriture.

E/S directes

De nombreuses applications, y compris les produits Oracle, peuvent contourner le cache du tampon hôte en activant des E/S directes Cette stratégie ne fonctionne pas comme prévu avec les systèmes de fichiers ZFS. Bien que le cache du tampon hôte soit contourné, ZFS lui-même continue à mettre en cache les données. Cette action peut entraîner des résultats trompeurs lors de l'utilisation d'outils tels que fio ou Sio pour effectuer des tests de performances. En effet, il est difficile de prévoir si les E/S atteignent le système de stockage ou si elles sont mises en cache localement au sein du système d'exploitation. Cette action rend également très difficile l'utilisation de tels tests synthétiques pour comparer les performances ZFS aux autres systèmes de fichiers. D'un point de vue pratique, les performances du système de fichiers varient considérablement, voire nulle, pour les charges de travail réelles des utilisateurs.

Plusieurs zpools

Les sauvegardes, les restaurations, les clones et l'archivage des données ZFS basés sur des snapshots doivent être effectués au niveau du zpool et requièrent généralement plusieurs zpools. Un zpool est similaire à un groupe de disques LVM et doit être configuré à l'aide des mêmes règles. Par exemple, il est probablement préférable de définir au mieux une base de données avec les fichiers de données résidant sur `zpool1` ainsi que les journaux d'archivage, les fichiers de contrôle et les journaux de reprise qui résident sur `zpool2`. Cette

approche permet une sauvegarde à chaud standard dans laquelle la base de données est placée en mode de sauvegarde à chaud, suivie d'un snapshot de `zpool1`. La base de données est alors supprimée du mode de sauvegarde à chaud, l'archivage des journaux est forcé et un instantané de `zpool2` est créé. Une opération de restauration nécessite de démonter les systèmes de fichiers zfs et de mettre hors ligne le zpool dans son intégralité, après une opération de restauration SnapRestore. Le zpool peut alors être remis en ligne et la base de données récupérée.

filesytemio_options

Le paramètre Oracle `filesystemio_options` Fonctionne différemment avec ZFS. Si `setall` ou `directio` Est utilisé, les opérations d'écriture sont synchrones et contournent le cache du tampon du système d'exploitation, mais les lectures sont mises en tampon par ZFS. Cette action engendre des difficultés dans l'analyse des performances, car les E/S sont parfois interceptées et traitées par le cache ZFS, ce qui rend la latence du stockage et les E/S totales inférieures à ce qu'elles semblent être.

Configuration de l'hôte avec les systèmes ASA r2

AIX

Sujets de configuration pour la base de données Oracle sur IBM AIX avec ASA r2 ONTAP.

AIX est pris en charge avec NetApp ASA r2 pour l'hébergement de bases de données Oracle, à condition que :

- Vous configurez correctement Oracle pour les E/S simultanées.
- Vous utilisez les protocoles SAN pris en charge (FC/iSCSI/NVMe).
- Vous exécutez ONTAP 9.16.x ou une version ultérieure sur ASA r2.

E/S simultanées

Pour obtenir des performances optimales sur IBM AIX avec ASA r2, il est nécessaire d'utiliser des E/S simultanées. Sans E/S simultanées, les limitations de performances sont probables car AIX effectue des E/S sérialisées et atomiques, ce qui engendre une surcharge importante.

À l'origine, NetApp recommandait d'utiliser `cio` L'option de montage permettait de forcer les E/S simultanées sur le système de fichiers, mais ce processus présentait des inconvénients et n'est plus nécessaire. Depuis l'introduction d'AIX 5.2 et d'Oracle 10gR1, Oracle sur AIX peut ouvrir des fichiers individuels pour des E/S simultanées, au lieu de forcer des E/S simultanées sur l'ensemble du système de fichiers.

La meilleure méthode pour activer les E/S simultanées est de définir le `init.ora` paramètre `filesystemio_options` à `setall`. Oracle peut ainsi ouvrir des fichiers spécifiques pour une utilisation avec des E/S simultanées

L'utilisation de cio comme option de montage force l'utilisation d'E/S simultanées, ce qui peut avoir des conséquences négatives. Par exemple, forcer les E/S simultanées désactive la lecture anticipée sur les systèmes de fichiers, ce qui peut nuire aux performances des E/S effectuées en dehors du logiciel de base de données Oracle, telles que la copie de fichiers et l'exécution de sauvegardes sur bande. De plus, des produits tels qu'Oracle GoldenGate et SAP BR*Tools ne sont pas compatibles avec l'utilisation de l'option de montage cio avec certaines versions d'Oracle.

NetApp recommande ce qui suit :

- N'utilisez pas le `cio` option de montage au niveau du système de fichiers. Activez plutôt les E/S simultanées via l'utilisation de `filesystemio_options=setall`.
- Utilisez uniquement le `cio` option de montage s'il n'est pas possible de définir `filesystemio_options=setall`.

Étant donné que ASA r2 ne prend pas en charge le NAS, tous les déploiements Oracle sur AIX doivent utiliser des protocoles de blocs.

Options de montage AIX jfs/jfs2

Le tableau suivant répertorie les options de montage AIX jfs/jfs2.

Type de fichier	Options de montage
Accueil ADR	Valeurs par défaut
fichiers de contrôle	Valeurs par défaut
Fichiers de données	Valeurs par défaut
Journaux de rétablissement	Valeurs par défaut
ORACLE_HOME	Valeurs par défaut

Avant d'utiliser AIX `hdisk` Dans tous les environnements, y compris les bases de données, vérifiez les paramètres des appareils. `queue_depth`. Ce paramètre ne correspond pas à la profondeur de la file d'attente HBA ; il se rapporte plutôt à la profondeur de la file d'attente SCSI de l'appareil individuel. `hdisk device`. Selon la configuration des LUN ASA r2, la valeur de `queue_depth` Cette valeur pourrait être trop faible pour de bonnes performances. Les tests ont montré que la valeur optimale était de 64.

HP-UX

Sujets de configuration pour la base de données Oracle sur HP-UX avec ASA r2 ONTAP.

HP-UX est pris en charge avec NetApp ASA r2 pour l'hébergement de bases de données Oracle, à condition que :

- La version ONTAP est 9.16.x ou ultérieure.
- Utilisez les protocoles SAN (FC/iSCSI/NVMe). Le NAS n'est pas pris en charge sur ASA r2.
- Appliquez les meilleures pratiques de montage et de réglage des E/S spécifiques à HP-UX.

Options de montage HP-UX VxFS

Utilisez les options de montage suivantes pour les systèmes de fichiers hébergeant les binaires Oracle :

```
delaylog,nodatainlog
```

Utilisez les options de montage suivantes pour les systèmes de fichiers contenant des fichiers de données,

des journaux de reprise, des journaux d'archivage et des fichiers de contrôle dans lesquels la version de HP-UX ne prend pas en charge les E/S simultanées :

```
nodatainlog,mincache=direct,convosync=direct
```

Lorsque des E/S simultanées sont prises en charge (VxFS 5.0.1 et versions ultérieures, ou avec ServiceGuard Storage Management Suite), utilisez ces options de montage pour les systèmes de fichiers contenant des fichiers de données, des journaux de reprise, des journaux d'archivage et des fichiers de contrôle :

```
delaylog,cio
```

Le paramètre `db_file_multiblock_read_count` Est particulièrement critique dans les environnements VxFS. Oracle recommande que ce paramètre ne soit pas défini dans Oracle 10g R1 et versions ultérieures, sauf indication contraire. La taille de bloc Oracle de 8 Ko par défaut est 128. Si la valeur de ce paramètre est forcée à 16 ou moins, retirer le `convosync=direct` Option de montage car elle peut endommager les performances des E/S séquentielles. Cette étape nuit à d'autres aspects de la performance et ne doit être prise que si la valeur de `db_file_multiblock_read_count` doit être modifié par rapport à la valeur par défaut.

Linux

Sujets de configuration spécifiques au système d'exploitation Linux avec ASA r2 ONTAP.

Linux (Oracle Linux, RHEL, SUSE) est pris en charge avec ASA r2 pour les bases de données Oracle. Utilisez les protocoles SAN, configurez correctement le multipathing et appliquez les meilleures pratiques Oracle pour l'optimisation ASM et E/S.

Planificateur d'E/S.

Le noyau Linux permet un contrôle de bas niveau sur la façon dont les E/S sont planifiées pour bloquer les périphériques. Les valeurs par défaut sur les différentes distributions de Linux varient considérablement. Les tests montrent que la date limite offre habituellement les meilleurs résultats, mais il arrive que le NOOP ait été légèrement meilleur. La différence de performance est minime, mais testez les deux options s'il est nécessaire d'extraire les performances maximales d'une configuration de base de données. Dans de nombreuses configurations, le paramètre CFQ est le paramètre par défaut. Il a démontré des problèmes de performances significatifs avec les charges de travail de la base de données.

Pour plus d'informations sur la configuration du planificateur d'E/S, reportez-vous à la documentation du fournisseur Linux correspondant.

Chemins d'accès multiples

Certains clients ont rencontré des pannes durant une interruption du réseau, car le démon multivoie ne s'exécutait pas sur leur système. Sur les versions récentes de Linux, le processus d'installation du système d'exploitation et le démon de chemins d'accès multiples peuvent exposer ces systèmes d'exploitation à ce problème. Les packages sont installés correctement, mais ils ne sont pas configurés pour un démarrage automatique après un redémarrage.

Par exemple, la configuration par défaut du démon multipath sur RHEL 9.7 pourrait ressembler à ceci :

```
[root@host1 ~]# systemctl list-unit-files --type=service | grep multipathd
multipathd.service                        disabled
```

Ceci peut être corrigé à l'aide des commandes suivantes :

```
[root@host1 ~]# systemctl enable multipathd.service
[root@host1 ~]# systemctl list-unit-files --type=service | grep multipathd
multipathd.service                        enabled
```

Profondeur de la file d'attente

Définissez une profondeur de file d'attente appropriée pour les périphériques SAN afin d'éviter les goulots d'étranglement d'E/S. La profondeur de file d'attente par défaut sous Linux est souvent fixée à 128, ce qui peut entraîner des problèmes de performance avec les bases de données Oracle. Un niveau de profondeur de file d'attente trop élevé peut entraîner une mise en file d'attente excessive des E/S, ce qui augmente la latence et réduit le débit. Un réglage trop bas peut limiter le nombre de requêtes d'E/S en attente, réduisant ainsi les performances globales. Une profondeur de file d'attente de 64 est souvent un bon point de départ pour les charges de travail de base de données Oracle sur ASA r2, mais elle peut devoir être ajustée en fonction des caractéristiques spécifiques de la charge de travail et des tests de performance.

Mise en miroir ASM

La mise en miroir ASM peut nécessiter des modifications des paramètres de chemins d'accès multiples Linux pour permettre à ASM de reconnaître un problème et de basculer vers un autre groupe de pannes. La plupart des configurations ASM sur ONTAP reposent sur une redondance externe. La protection des données est assurée par la baie externe et ASM ne met pas en miroir les données. Certains sites utilisent ASM avec redondance normale pour fournir une mise en miroir bidirectionnelle, généralement entre différents sites.

Pour les systèmes ASA r2 prenant en charge le multipathing actif-actif, ces paramètres multipath doivent être ajustés. Comme tous les chemins sont actifs et à charge équilibrée, une mise en file d'attente indéfinie n'est pas nécessaire. Les paramètres multipath devraient plutôt privilégier les performances et la restauration rapide en cas de panne. Ce comportement est important pour la mise en miroir ASM car ASM doit recevoir une défaillance d'E/S pour pouvoir réessayer l'E/S sur un LUN alternatif. Si les E/S sont mises en file d'attente indéfiniment, ASM ne peut pas déclencher de basculement.

Définissez les paramètres suivants dans Linux `multipath.conf` Fichier pour les LUN ASM utilisés avec la mise en miroir ASM :

```
polling_interval 5
no_path_retry 24
failback immediate
path_grouping_policy multibus
path_selector "service-time 0"
```

Ces paramètres créent une temporisation de 120 secondes pour les périphériques ASM. Le délai d'attente est calculé comme étant le `polling_interval` * `no_path_retry` en secondes. Il peut être nécessaire

d'ajuster la valeur exacte dans certaines circonstances, mais un délai de 120 secondes doit être suffisant pour la plupart des utilisations. En particulier, 120 secondes doivent permettre un basculement ou un retour du contrôleur sans générer d'erreur d'E/S susceptible de mettre le groupe défaillant hors ligne.

Un plus bas `no_path_retry` La valeur peut réduire le temps nécessaire à ASM pour passer à un autre groupe de pannes, mais augmente également le risque de basculement indésirable lors des activités de maintenance, telles qu'une prise de contrôle. Le risque peut être atténué par une surveillance attentive de l'état de mise en miroir ASM. Si un basculement indésirable se produit, les miroirs peuvent être rapidement resynchronisés si la resynchronisation est effectuée relativement rapidement. Pour plus d'informations, consultez la documentation Oracle sur ASM Fast Mirror Resync pour la version du logiciel Oracle utilisé.

Options de montage Linux xfs, ext3 et ext4

* NetApp recommande* d'utiliser les options de montage par défaut. Veillez à un alignement correct lors de la création de systèmes de fichiers sur les LUN.

ASMLib/AFD (pilote de filtre ASM)

Sujets de configuration spécifiques au système d'exploitation Linux utilisant AFD et ASMlib avec ASA r2 ONTAP.

Tailles de bloc ASMlib

ASMlib est une bibliothèque de gestion ASM optionnelle et des utilitaires associés. Sa principale valeur réside dans la possibilité d'étiqueter un LUN comme une ressource ASM avec une étiquette lisible par l'homme.

Les versions récentes d'ASMlib détectent un paramètre LUN appelé blocs logiques par exposant de bloc physique (LBPPBE). Cette valeur n'a été signalée que récemment par la cible SCSI ONTAP. Elle renvoie désormais une valeur qui indique qu'une taille de bloc de 4 Ko est recommandée. Il ne s'agit pas d'une définition de la taille de bloc, mais il est un indice pour toute application utilisant LBPPBE que les E/S d'une certaine taille peuvent être gérées plus efficacement. Cependant, ASMlib interprète LBPPBE comme une taille de bloc et estampille constamment l'en-tête ASM lors de la création du périphérique ASM.

Ce processus peut causer des problèmes avec les mises à niveau et les migrations de différentes manières, tous en fonction de l'incapacité à mélanger des périphériques ASMlib avec des tailles de bloc différentes dans le même groupe de disques ASM.

Par exemple, des tableaux plus anciens ont généralement signalé une valeur LBPPBE de 0 ou n'ont pas signalé cette valeur du tout. ASMlib l'interprète comme une taille de bloc de 512 octets. Pour les baies plus récentes, la taille de bloc est de 4 Ko. Il n'est pas possible de mélanger des périphériques de 512 octets et de 4 Ko dans le même groupe de disques ASM. Cela empêche un utilisateur d'augmenter la taille du groupe de disques ASM en utilisant des LUN de deux baies ou en utilisant ASM comme outil de migration. Dans d'autres cas, RMAN pourrait ne pas permettre la copie de fichiers entre un groupe de disques ASM avec une taille de bloc de 512 octets et un groupe de disques ASM avec une taille de bloc de 4 Ko.

La solution préférée est de corriger ASMlib. L'ID de bug Oracle est 13999609 et le correctif est présent dans oracleasm-support-2.1.8-1 et versions ultérieures. Ce correctif permet à un utilisateur de définir le paramètre `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` à `true` dans le `/etc/sysconfig/oracleasm` fichier de configuration. Cela empêche ASMlib d'utiliser le paramètre LBPPBE, ce qui signifie que les LUN de la nouvelle baie sont maintenant reconnues comme des périphériques de bloc de 512 octets.

L'option ne modifie pas la taille de bloc sur les LUN précédemment estampées par ASMlib. Par exemple, si un groupe de disques ASM avec des blocs de 512 octets doit être migré vers un nouveau système de stockage qui signale un bloc de 4 Ko, l'option `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` Doit être défini avant que les nouvelles LUN soient estampées avec ASMlib. Si les périphériques ont déjà été estampillés par oracleasm, ils doivent être reformatés avant d'être repoussées avec une nouvelle taille de bloc. Commencez par déconfigurer le périphérique avec `oracleasm deletedisk`, Puis effacez le premier 1 Go du périphérique avec `dd if=/dev/zero of=/dev/mapper/device bs=1048576 count=1024`. Enfin, si le périphérique a déjà été partitionné, utilisez le `kpartx` Commande permettant de supprimer les partitions obsolètes ou de simplement redémarrer le système d'exploitation.

Si ASMlib ne peut pas être corrigé, ASMlib peut être supprimé de la configuration. Ce changement est perturbateur et nécessite le démarquage des disques ASM et s'assurer que le `asm_diskstring` le paramètre est défini correctement. Toutefois, cette modification ne nécessite pas la migration des données.

Tailles de bloc d'entraînement de filtre ASM (AFD)

AFD est une bibliothèque de gestion ASM facultative qui remplace ASMlib. Du point de vue du stockage, il est très similaire à ASMlib, mais il inclut des fonctionnalités supplémentaires telles que la capacité de bloquer les E/S non-Oracle afin de réduire les risques d'erreurs d'utilisateur ou d'application susceptibles de corrompre les données.

Tailles des blocs de périphériques

Comme ASMlib, AFD lit également le paramètre LUN blocs logiques par exposant de bloc physique (LBPPBE) et utilise par défaut la taille de bloc physique, et non la taille de bloc logique.

Cela peut créer un problème si l'AFD est ajouté à une configuration existante où les périphériques ASM sont déjà formatés comme des périphériques de bloc de 512 octets. Le pilote AFD reconnaîtrait le LUN comme un périphérique 4K et l'incompatibilité entre l'étiquette ASM et le périphérique physique empêcherait l'accès. De même, les migrations seraient affectées, car il n'est pas possible de combiner des périphériques de 512 octets et de 4 Ko dans le même groupe de disques ASM. Cela empêche un utilisateur d'augmenter la taille du groupe de disques ASM en utilisant des LUN de deux baies ou en utilisant ASM comme outil de migration. Dans d'autres cas, RMAN pourrait ne pas permettre la copie de fichiers entre un groupe de disques ASM avec une taille de bloc de 512 octets et un groupe de disques ASM avec une taille de bloc de 4 Ko.

La solution est simple - AFD inclut un paramètre pour contrôler si elle utilise les tailles de bloc logiques ou physiques. Il s'agit d'un paramètre global affectant tous les périphériques du système. Pour forcer AFD à utiliser la taille de bloc logique, définissez `options oracleafd oracleafd_use_logical_block_size=1` dans le `/etc/modprobe.d/oracleafd.conf` fichier.

Tailles de transfert multivoie

Les modifications récentes du noyau linux appliquent des restrictions de taille d'E/S envoyées aux périphériques à chemins d'accès multiples, et AFD ne respecte pas ces restrictions. Les E/S sont ensuite rejetées, ce qui entraîne la mise hors ligne du chemin d'accès à la LUN. Il en résulte une incapacité à installer Oracle Grid, à configurer ASM ou à créer une base de données.

La solution consiste à spécifier manuellement la longueur de transfert maximale dans le fichier multipath.conf pour les LUN ONTAP :

```
devices {
            device {
                vendor "NETAPP"
                product "LUN.*"
                max_sectors_kb 4096
        }
    }
```

Même si aucun problème n'existe actuellement, ce paramètre doit être défini si l'AFD est utilisé pour garantir qu'une future mise à niveau de linux ne provoque pas de problèmes inattendus.

Microsoft Windows

Sujets de configuration pour la base de données Oracle sur Microsoft Windows avec ASA r2 ONTAP.

SAN

Pour une efficacité de compression optimale, assurez-vous que le système de fichiers NTFS utilise une unité d'allocation de 8 Ko ou plus. L'utilisation d'une unité d'allocation 4K, qui est généralement la valeur par défaut, a un impact négatif sur l'efficacité de la compression.

Solaris

Sujets de configuration spécifiques au système d'exploitation Solaris avec ASA r2 ONTAP.

Options de montage Solaris UFS

NetApp recommande fortement d'utiliser l'option de montage de journalisation afin de préserver l'intégrité des données en cas de panne de l'hôte Solaris ou d'interruption de la connectivité FC. L'option de montage de la journalisation préserve également l'utilisation des sauvegardes Snapshot.

ZFS Solaris

Solaris ZFS doit être installé et configuré avec soin pour offrir des performances optimales.

mvector

Solaris 11 a inclus un changement dans la façon dont il traite les opérations d'E/S importantes, ce qui peut entraîner de graves problèmes de performances sur les baies de stockage SAN. Le problème est documenté NetApp suivi bug report 630173, "Solaris 11 ZFS Performance regression".

Il ne s'agit pas d'un bogue de ONTAP. Il s'agit d'un défaut Solaris suivi sous les défauts Solaris 7199305 et 7082975.

Vous pouvez consulter le support Oracle pour savoir si votre version de Solaris 11 est affectée, ou vous pouvez tester la solution de contournement en la changeant `zfs_mvector_max_size` à une valeur plus petite.

Pour ce faire, exécutez la commande suivante en tant que root :

```
[root@host1 ~]# echo "zfs_mvector_max_size/W 0t131072" |mdb -kw
```

En cas de problème inattendu résultant de cette modification, vous pouvez facilement l'inverser en exécutant la commande suivante en tant que root :

```
[root@host1 ~]# echo "zfs_mvector_max_size/W 0t1048576" |mdb -kw
```

Noyau

Pour des performances ZFS fiables, un noyau Solaris est nécessaire pour résoudre les problèmes d'alignement des LUN. Le correctif a été introduit avec le correctif 147440-19 dans Solaris 10 et avec SRU 10.5 pour Solaris 11. Utilisez uniquement Solaris 10 et versions ultérieures avec ZFS.

Configuration du LUN

Pour configurer une LUN, effectuez les opérations suivantes :

1. Créer une LUN de type `solaris`.
2. Installez le kit d'utilitaire hôte (HUK) approprié spécifié par le "Matrice d'interopérabilité NetApp (IMT)".
3. Suivez les instructions du HUK exactement comme décrit. Les étapes de base sont décrites ci-dessous, mais reportez-vous au "documentation la plus récente" pour connaître la procédure adéquate.
   a. Exécutez le `host_config` utilitaire de mise à jour du `sd.conf/sdd.conf` fichier. Les disques SCSI seront ainsi en mesure de détecter correctement les LUN ONTAP.
   b. Suivez les instructions fournies par le `host_config` Utilitaire permettant d'activer les entrées/sorties multivoies (MPIO).
   c. Redémarrez. Cette étape est nécessaire pour que les modifications soient reconnues dans l'ensemble du système.
4. Partitionnez les LUN et vérifiez qu'ils sont correctement alignés. Voir « Annexe B : Vérification de l'alignement WAFL » pour obtenir des instructions sur la façon de tester et de confirmer directement l'alignement.

zpools

Un zpool ne doit être créé qu'après les étapes de la "Configuration du LUN" sont effectuées. Si la procédure n'est pas effectuée correctement, les performances risquent d'être sérieusement dégradées en raison de l'alignement des E/S. Pour des performances optimales sur ONTAP, les E/S doivent être alignées sur une limite de 4 Ko sur un disque. Les systèmes de fichiers créés sur un zpool utilisent une taille de bloc effective qui est contrôlée par un paramètre appelé `ashift`, qui peut être affiché en exécutant la commande `zdb -C`.

La valeur de `ashift` la valeur par défaut est 9, ce qui signifie 2^9, ou 512 octets. Pour des performances optimales, le `ashift` La valeur doit être 12 (2^12=4K). Cette valeur est définie au moment de la création du zpool et ne peut pas être modifiée, ce qui signifie que les données dans zpools avec `ashift` une migration autre que 12 doit être effectuée en copiant les données vers un nouveau zpool.

Après avoir créé un zpool, vérifiez la valeur de `ashift` avant de continuer. Si la valeur n'est pas 12, les LUN n'ont pas été détectées correctement. Détruisez le zpool, vérifiez que toutes les étapes indiquées dans la

documentation des utilitaires hôtes correspondante ont été effectuées correctement et recréez le zpool.

Zpools et LDOMS Solaris

Les LDOMS Solaris créent une exigence supplémentaire pour s'assurer que l'alignement des E/S est correct. Bien qu'un LUN soit correctement découvert en tant que périphérique 4K, un périphérique virtuel vdsk sur un LDOM n'hérite pas de la configuration du domaine d'E/S. Le vdsk basé sur cette LUN revient par défaut à un bloc de 512 octets.

Un fichier de configuration supplémentaire est requis. Tout d'abord, les LDOM individuels doivent être corrigés pour le bogue Oracle 15824910 afin d'activer les options de configuration supplémentaires. Ce correctif a été porté dans toutes les versions actuellement utilisées de Solaris. Une fois le logiciel LDOM corrigé, il est prêt à configurer les nouveaux LUN correctement alignés comme suit :

1. Identifiez la ou les LUN à utiliser dans le nouveau zpool. Dans cet exemple, il s'agit du périphérique c2d1.

```
[root@LDOM1 ~]# echo | format
Searching for disks...done
AVAILABLE DISK SELECTIONS:
  0. c2d0 <Unknown-Unknown-0001-100.00GB>
     /virtual-devices@100/channel-devices@200/disk@0
  1. c2d1 <SUN-ZFS Storage 7330-1.0 cyl 1623 alt 2 hd 254 sec 254>
     /virtual-devices@100/channel-devices@200/disk@1
```

2. Récupérez l'instance vdc des systèmes à utiliser pour un pool ZFS :

```
[root@LDOM1 ~]#  cat /etc/path_to_inst
#
# Caution! This file contains critical kernel state
#
"/fcoe" 0 "fcoe"
"/iscsi" 0 "iscsi"
"/pseudo" 0 "pseudo"
"/scsi_vhci" 0 "scsi_vhci"
"/options" 0 "options"
"/virtual-devices@100" 0 "vnex"
"/virtual-devices@100/channel-devices@200" 0 "cnex"
"/virtual-devices@100/channel-devices@200/disk@0" 0 "vdc"
"/virtual-devices@100/channel-devices@200/pciv-communication@0" 0 "vpci"
"/virtual-devices@100/channel-devices@200/network@0" 0 "vnet"
"/virtual-devices@100/channel-devices@200/network@1" 1 "vnet"
"/virtual-devices@100/channel-devices@200/network@2" 2 "vnet"
"/virtual-devices@100/channel-devices@200/network@3" 3 "vnet"
"/virtual-devices@100/channel-devices@200/disk@1" 1 "vdc" << We want
this one
```

3. Modifier `/platform/sun4v/kernel/drv/vdc.conf`:

```
block-size-list="1:4096";
```

Cela signifie que l'instance de périphérique 1 se voit attribuer une taille de bloc de 4096.

Par exemple, supposons que les instances vdsk 1 à 6 doivent être configurées pour une taille de bloc de 4 Ko et `/etc/path_to_inst` se lit comme suit :

```
"/virtual-devices@100/channel-devices@200/disk@1" 1 "vdc"
"/virtual-devices@100/channel-devices@200/disk@2" 2 "vdc"
"/virtual-devices@100/channel-devices@200/disk@3" 3 "vdc"
"/virtual-devices@100/channel-devices@200/disk@4" 4 "vdc"
"/virtual-devices@100/channel-devices@200/disk@5" 5 "vdc"
"/virtual-devices@100/channel-devices@200/disk@6" 6 "vdc"
```

4. La finale `vdc.conf` le fichier doit contenir les éléments suivants :

```
block-size-list="1:8192","2:8192","3:8192","4:8192","5:8192","6:8192";
```

> Le LDOM doit être redémarré après la configuration de vdc.conf et la création du vdsk. Cette étape ne peut pas être évitée. La modification de la taille de bloc n'est effective qu'après un redémarrage. Procéder à la configuration du pool de zpool et s'assurer que le module de transmission automatique est correctement réglé sur 12 comme décrit précédemment.

Journal des intentions ZFS (ZIL)

En général, il n'y a aucune raison de localiser le ZFS Intent Log (ZIL) sur un autre périphérique. Le journal peut partager de l'espace avec le pool principal. L'utilisation principale d'une ZIL distincte est l'utilisation de disques physiques qui n'offrent pas les fonctionnalités de mise en cache des écritures dans les baies de stockage modernes.

biais logique

Réglez le `logbias` Paramètre sur les systèmes de fichiers ZFS hébergeant les données Oracle.

```
zfs set logbias=throughput <filesystem>
```

Ce paramètre réduit les niveaux d'écriture globaux. Sous les valeurs par défaut, les données écrites sont d'abord validées dans le ZIL, puis dans le pool de stockage principal. Cette approche est adaptée à une configuration utilisant une configuration de disque simple, qui inclut un périphérique ZIL SSD et un support rotatif pour le pool de stockage principal. En effet, elle permet une validation dans une seule transaction d'E/S sur le support à latence la plus faible disponible.

Lorsque vous utilisez une baie de stockage moderne qui inclut sa propre capacité de mise en cache, cette approche n'est généralement pas nécessaire. Dans de rares cas, il peut être souhaitable d'effectuer une

écriture avec une seule transaction dans le journal, par exemple une charge de travail composée d'écritures aléatoires hautement concentrées et sensibles à la latence. L'amplification d'écriture peut avoir des conséquences, car les données consignées sont finalement écrites dans le pool de stockage principal, ce qui double l'activité d'écriture.

E/S directes

De nombreuses applications, y compris les produits Oracle, peuvent contourner le cache du tampon hôte en activant des E/S directes Cette stratégie ne fonctionne pas comme prévu avec les systèmes de fichiers ZFS. Bien que le cache du tampon hôte soit contourné, ZFS lui-même continue à mettre en cache les données. Cette action peut entraîner des résultats trompeurs lors de l'utilisation d'outils tels que fio ou Sio pour effectuer des tests de performances. En effet, il est difficile de prévoir si les E/S atteignent le système de stockage ou si elles sont mises en cache localement au sein du système d'exploitation. Cette action rend également très difficile l'utilisation de tels tests synthétiques pour comparer les performances ZFS aux autres systèmes de fichiers. D'un point de vue pratique, les performances du système de fichiers varient considérablement, voire nulle, pour les charges de travail réelles des utilisateurs.

Plusieurs zpools

Les sauvegardes, les restaurations, les clones et l'archivage des données ZFS basés sur des snapshots doivent être effectués au niveau du zpool et requièrent généralement plusieurs zpools. Un zpool est similaire à un groupe de disques LVM et doit être configuré à l'aide des mêmes règles. Par exemple, il est probablement préférable de définir au mieux une base de données avec les fichiers de données résidant sur `zpool1` ainsi que les journaux d'archivage, les fichiers de contrôle et les journaux de reprise qui résident sur `zpool2`. Cette approche permet une sauvegarde à chaud standard dans laquelle la base de données est placée en mode de sauvegarde à chaud, suivie d'un snapshot de `zpool1`. La base de données est alors supprimée du mode de sauvegarde à chaud, l'archivage des journaux est forcé et un instantané de `zpool2` est créé. Une opération de restauration nécessite de démonter les systèmes de fichiers zfs et de mettre hors ligne le zpool dans son intégralité, après une opération de restauration SnapRestore. Le zpool peut alors être remis en ligne et la base de données récupérée.

filesytemio_options

Le paramètre Oracle `filesystemio_options` Fonctionne différemment avec ZFS. Si `setall` ou `directio` Est utilisé, les opérations d'écriture sont synchrones et contournent le cache du tampon du système d'exploitation, mais les lectures sont mises en tampon par ZFS. Cette action engendre des difficultés dans l'analyse des performances, car les E/S sont parfois interceptées et traitées par le cache ZFS, ce qui rend la latence du stockage et les E/S totales inférieures à ce qu'elles semblent être.

Configuration réseau sur les systèmes AFF/ FAS

Interfaces logiques

Les bases de données Oracle doivent accéder au stockage. Les interfaces logiques (LIF) correspondent à la tuyauterie réseau qui connecte une machine virtuelle de stockage (SVM) au réseau et, par conséquent, à la base de données. Une conception correcte des LIF est requise pour s'assurer qu'il y a suffisamment de bande passante pour chaque charge de travail de la base de données, et le basculement ne provoque pas de perte des services de stockage.

Cette section présente les principes clés de conception des LIF. Pour obtenir une documentation plus complète, reportez-vous au "Documentation de gestion de réseau ONTAP". Comme pour les autres aspects

de l'architecture de la base de données, les meilleures options pour la conception des machines virtuelles de stockage (SVM, appelé SVM au niveau de l'interface de ligne de commande) et de l'interface logique (LIF) dépendent largement des besoins en termes d'évolutivité et des besoins de l'entreprise.

Tenez compte des principaux sujets suivants lors de l'élaboration d'une stratégie LIF :

- **Performances.** la bande passante du réseau est-elle suffisante ?
- **Résilience.** y a-t-il des points de défaillance uniques dans la conception?
- **Gérabilité.** le réseau peut-il être mis à l'échelle sans interruption ?

Ces rubriques s'appliquent à la solution de bout en bout, de l'hôte aux commutateurs et au système de stockage.

Types de LIF

Il existe plusieurs types de LIF. "Documentation ONTAP sur les types de LIF" Fournir des informations plus complètes à ce sujet, mais d'un point de vue fonctionnel, les LIF peuvent être divisées en plusieurs groupes :

- **LIFs de gestion de clusters et de nœuds.** utilisées pour gérer le cluster de stockage.
- **LIF de gestion SVM.** interfaces permettant l'accès à une SVM via l'API REST ou ONTAPI (aussi connue sous le nom de ZAPI) pour des fonctions telles que la création de snapshots ou le redimensionnement de volumes. Des produits tels que SnapManager pour Oracle (SMO) doivent avoir accès à une LIF de gestion SVM.
- **Interfaces de données LIF.** pour FC, iSCSI, NVMe/FC, NVMe/TCP, NFS, ou SMB/CIFS.

> Une LIF de données utilisée pour le trafic NFS peut également être utilisée à des fins de gestion en modifiant la politique de pare-feu de `data` à `mgmt` Ou une autre règle autorisant HTTP, HTTPS ou SSH. Ce changement peut simplifier la configuration du réseau en évitant la configuration de chaque hôte pour l'accès à la fois à la LIF de données NFS et à une LIF de gestion distincte. Il n'est pas possible de configurer une interface pour l'iSCSI et le trafic de gestion, bien que les deux utilisent un protocole IP. Une LIF de gestion distincte est requise dans les environnements iSCSI.

Conception de SAN LIF

La conception de LIF dans un environnement SAN est relativement simple pour une raison : les chemins d'accès multiples. Toutes les implémentations SAN modernes permettent à un client d'accéder aux données sur plusieurs chemins réseau indépendants et de sélectionner le ou les chemins d'accès les plus adaptés. Par conséquent, les performances du design LIF sont plus simples à gérer, car les clients SAN équilibrent automatiquement la charge en E/S sur les meilleurs chemins disponibles.

Si un chemin devient indisponible, le client sélectionne automatiquement un autre chemin. La simplicité de conception qui en résulte rend les LIF SAN généralement plus faciles à gérer. Cela ne signifie pas pour autant qu'un environnement SAN est toujours plus facile à gérer, car de nombreux autres aspects du stockage SAN sont bien plus complexes que NFS. Cela signifie simplement que la conception de la LIF SAN est plus facile.

Performance

La bande passante est l'élément le plus important à prendre en compte dans les performances de LIF dans un environnement SAN. Par exemple, un cluster ONTAP AFF à deux nœuds doté de deux ports FC 16 Gb par nœud permet d'obtenir jusqu'à 32 Go de bande passante vers/depuis chaque nœud.

Résilience

Les LIF SAN ne basculent pas sur un système de stockage AFF. Si une LIF SAN échoue en raison du basculement du contrôleur, le logiciel de chemins d'accès multiples du client détecte la perte d'un chemin et redirige les E/S vers une autre LIF. Avec les systèmes de stockage ASA, les LIF basculent après un court délai, mais cela n'interrompt pas les E/S, car il existe déjà des chemins actifs sur l'autre contrôleur. Le processus de basculement a lieu afin de restaurer l'accès de l'hôte sur tous les ports définis.

Gestion aisée

La migration des LIF est une tâche beaucoup plus courante dans un environnement NFS, car elle est souvent associée au déplacement des volumes au sein du cluster. Il n'est pas nécessaire de migrer une LIF dans un environnement SAN lorsque les volumes sont déplacés au sein de la paire HA. En effet, une fois le déplacement de volume terminé, ONTAP envoie une notification au SAN concernant un changement de chemins et les clients SAN se réoptimisent automatiquement. La migration de LIF avec SAN est principalement associée à des modifications matérielles physiques majeures. Par exemple, si une mise à niveau des contrôleurs sans interruption est requise, une LIF SAN est migrée vers le nouveau matériel. Si un port FC est défectueux, une LIF peut être migrée vers un port non utilisé.

Recommandations de conception

NetApp fait les recommandations suivantes :

- Ne créez pas plus de chemins que nécessaire. Un nombre excessif de chemins complique la gestion globale et peut entraîner des problèmes de basculement de chemin sur certains hôtes. De plus, certains hôtes ont des limites de chemin inattendues pour les configurations comme le démarrage SAN.
- Très peu de configurations doivent nécessiter plus de quatre chemins vers une LUN. L'intérêt d'avoir plus de deux nœuds de chemins publicitaires vers les LUN est limité, car l'agrégat hébergeant une LUN est inaccessible en cas de défaillance du nœud qui détient la LUN et de son partenaire haute disponibilité. Dans ce cas, la création de chemins sur des nœuds autres que la paire haute disponibilité principale n'est pas utile.
- Même si vous pouvez gérer le nombre de chemins de LUN visibles en sélectionnant les ports inclus dans les zones FC, il est généralement plus facile d'inclure tous les points cibles potentiels dans la zone FC et de contrôler la visibilité des LUN au niveau des ONTAP.
- Dans ONTAP 8.3 et versions ultérieures, la fonction de mappage de LUN sélectif (SLM) est la fonction par défaut. Avec SLM, toute nouvelle LUN est automatiquement annoncée à partir du nœud qui possède l'agrégat sous-jacent et du partenaire HA du nœud. Cet arrangement évite de créer des ensembles de ports ou de configurer le zoning pour limiter l'accessibilité des ports. Chaque LUN est disponible sur le nombre minimal de nœuds requis pour des performances et une résilience optimales.
  *Dans le cas où un LUN doit être migré en dehors des deux contrôleurs, les nœuds supplémentaires peuvent être ajoutés avec le `lun mapping add-reporting-nodes` De sorte que les LUN soient annoncées sur les nouveaux nœuds. Vous créez ainsi des chemins SAN supplémentaires vers les LUN pour la migration des LUN. Toutefois, l'hôte doit effectuer une opération de découverte pour utiliser les nouveaux chemins.
- Ne vous souciez pas trop du trafic indirect. Dans un environnement très exigeant en E/S, il est préférable d'éviter le trafic indirect pour lequel chaque microseconde de latence est critique, mais l'impact visible sur la performance est négligeable pour les charges de travail classiques.

Conception de LIF NFS

Contrairement aux protocoles SAN, NFS dispose d'une capacité limitée de définir plusieurs chemins d'accès aux données. Les extensions NFS parallèles (pNFS) à NFSv4 répondent à cette limitation, mais l'ajout de chemins d'accès supplémentaires devient rarement intéressant dans la mesure où les vitesses ethernet

atteignent 100 Go et au-delà.

Performances et résilience

Bien que la mesure des performances d'une LIF SAN consiste principalement à calculer la bande passante totale à partir de tous les chemins principaux, la détermination des performances d'une LIF NFS nécessite d'étudier de plus près la configuration réseau exacte. Par exemple, deux ports 10 Gbit peuvent être configurés comme ports physiques bruts ou en tant que groupe d'interface LACP (Link Aggregation Control Protocol). S'ils sont configurés en tant que groupe d'interface, plusieurs stratégies d'équilibrage de charge sont disponibles et fonctionnent différemment selon que le trafic est commuté ou routé. Enfin, Oracle Direct NFS (dNFS) propose des configurations d'équilibrage de charge qui n'existent pour le moment dans aucun client OS NFS.

Contrairement aux protocoles SAN, les systèmes de fichiers NFS nécessitent une résilience au niveau de la couche de protocole. Par exemple, une LUN est toujours configurée avec les chemins d'accès multiples activés, ce qui signifie que plusieurs canaux redondants sont disponibles pour le système de stockage, chacun utilisant le protocole FC. Un système de fichiers NFS, en revanche, dépend de la disponibilité d'un seul canal TCP/IP qui ne peut être protégé qu'au niveau de la couche physique. C'est pourquoi des options telles que le basculement de port et l'agrégation de ports LACP existent.

Dans un environnement NFS, les performances et la résilience sont fournies au niveau de la couche du protocole réseau. En conséquence, ces deux sujets sont étroitement liés et doivent être discutés ensemble.

Lier les LIFs aux groupes de ports

Pour lier une LIF à un port group, associez l'adresse IP de la LIF à un groupe de ports physiques. La méthode principale pour agréger les ports physiques est le LACP. La fonctionnalité de tolérance aux pannes de LACP est assez simple : chaque port d'un groupe LACP est surveillé et supprimé du groupe de ports en cas de dysfonctionnement. Cependant, il existe de nombreuses idées fausses sur le fonctionnement de LACP en matière de performances :

- LACP ne requiert pas que la configuration sur le switch corresponde au terminal. Par exemple, ONTAP peut être configuré avec un équilibrage de charge basé sur IP, tandis qu'un commutateur peut utiliser un équilibrage de charge basé sur MAC.
- Chaque noeud final utilisant une connexion LACP peut choisir indépendamment le port de transmission des paquets, mais il ne peut pas choisir le port utilisé pour la réception. Cela signifie que le trafic de ONTAP vers une destination particulière est lié à un port particulier, et que le trafic de retour peut arriver sur une interface différente. Cela ne cause cependant aucun problème.
- LACP ne distribue pas uniformément le trafic en permanence. Dans un grand environnement comptant de nombreux clients NFS, le résultat est même généralement l'utilisation de tous les ports d'une agrégation LACP. Cependant, tout système de fichiers NFS dans l'environnement est limité à la bande passante d'un seul port, et non à l'agrégation complète.
- Bien que les politiques LACP robin-Robin soient disponibles sur ONTAP, ces règles n'abordent pas la connexion entre un switch et un hôte. Par exemple, une configuration avec une jonction LACP à quatre ports sur un hôte et une jonction LACP à quatre ports sur ONTAP ne peut toujours lire un système de fichiers qu'à l'aide d'un seul port. Bien que ONTAP puisse transmettre des données via les quatre ports, aucune technologie de commutation n'est actuellement disponible, qui envoie du commutateur à l'hôte via les quatre ports. Un seul est utilisé.

L'approche la plus courante dans les grands environnements composés de nombreux hôtes de base de données est de créer un agrégat LACP comportant un nombre approprié d'interfaces 10 Gbit (ou plus rapides) en utilisant l'équilibrage de la charge IP. Cette approche permet à ONTAP d'assurer une utilisation uniforme de tous les ports, tant qu'il y a suffisamment de clients. L'équilibrage de la charge est défaillant lorsque la configuration compte moins de clients, car les ressources en ligne LACP ne redistribuent pas la charge de

manière dynamique.

Lorsqu'une connexion est établie, le trafic dans une direction particulière est placé sur un seul port. Par exemple, une base de données effectuant une analyse de table complète sur un système de fichiers NFS connecté via une jonction LACP à quatre ports lit les données via une seule carte d'interface réseau (NIC). Si seulement trois serveurs de base de données se trouvent dans un tel environnement, il est possible que les trois derniers lisent à partir du même port, alors que les trois autres ports sont inactifs.

Lier les LIF à des ports physiques

La liaison d'une LIF à un port physique permet un contrôle plus granulaire de la configuration du réseau, car une adresse IP donnée sur un système ONTAP n'est associée qu'à un seul port réseau à la fois. La résilience s'obtient ensuite via la configuration des groupes de basculement et des règles de basculement.

Stratégies de basculement et groupes de basculement

Le comportement des LIF durant une interruption du réseau est contrôlé par des règles de basculement et des groupes de basculement. Les options de configuration ont été modifiées avec les différentes versions de ONTAP. Consulter le "Documentation de gestion de réseau ONTAP pour les groupes et politiques de basculement" Pour plus d'informations sur la version de ONTAP déployée.

Les versions ONTAP 8.3 et supérieures permettent la gestion du basculement des LIF sur la base des domaines de diffusion. Par conséquent, un administrateur peut définir tous les ports ayant accès à un sous-réseau donné et autoriser ONTAP à sélectionner une LIF de basculement appropriée. Cette approche peut être utilisée par certains clients, mais elle est limitée dans un environnement de réseau de stockage haut débit en raison du manque de prévisibilité. Par exemple, un environnement peut inclure à la fois des ports 1 Gbit pour l'accès aux systèmes de fichiers de routine et des ports 10 Gbit pour les E/S des fichiers de données Si les deux types de ports existent dans le même broadcast domain, le basculement de LIF peut entraîner le déplacement des E/S des fichiers de données d'un port 10 Gb vers un port 1 Gb.

En résumé, tenez compte des pratiques suivantes :

1. Configurez un groupe de basculement comme défini par l'utilisateur.
2. Remplissez le groupe de basculement avec les ports du contrôleur partenaire de basculement de stockage (SFO) de sorte que les LIF suivent les agrégats lors d'un basculement de stockage. Cela évite de créer du trafic indirect.
3. Utilisez les ports de basculement avec des caractéristiques de performance correspondantes à la LIF d'origine. Par exemple, une LIF située sur un seul port physique de 10 Go doit inclure un groupe de basculement doté d'un seul port 10 Go. Une LIF LACP à quatre ports doit basculer vers une autre LIF LACP à quatre ports. Ces ports seraient un sous-ensemble des ports définis dans le domaine de diffusion.
4. Définissez la politique de basculement sur partenaire SFO uniquement. Veillez donc à ce que la LIF suive l'agrégat lors du failover.

Restauration automatique

Réglez le `auto-revert` paramètre selon vos besoins. La plupart des clients préfèrent définir ce paramètre sur `true` Pour que la LIF rerevienne sur son port home. Cependant, dans certains cas, les clients ont défini cette option sur `false `afin qu'un basculement inattendu puisse être recherché avant de renvoyer une LIF à son port de attache.

Rapport LIF/volume

On croit souvent, à tort, qu'il doit y avoir une relation 1:1 entre les volumes et les LIFs NFS. Même si cette

configuration est requise pour déplacer un volume n'importe où dans un cluster sans jamais créer de trafic d'interconnexion supplémentaire, elle n'est pas obligatoire de manière catégorique. Le trafic intercluster doit être envisagé, mais la simple présence du trafic intercluster ne crée pas de problèmes. Nombre des bancs d'essai publiés pour ONTAP portent sur des E/S principalement indirectes

Par exemple, un projet de base de données contenant un nombre relativement limité de bases de données pour lesquelles seuls 40 volumes nécessitent des performances élevées peut justifier un rapport volume 1:1 vers une stratégie LIF, un arrangement qui nécessiterait 40 adresses IP. N'importe quel volume peut ensuite être déplacé n'importe où dans le cluster avec la LIF associée, et le trafic serait toujours direct, minimisant ainsi chaque source de latence, même à des niveaux d'une microseconde.

Par exemple, un grand environnement hébergé peut être plus facilement géré avec une relation 1:1 entre les clients et les LIF. Au fil du temps, un volume peut avoir besoin d'être migré vers un autre nœud, ce qui provoque du trafic indirect. Cependant, l'effet sur les performances doit être indétectable à moins que les ports réseau du commutateur d'interconnexion ne soient saturés. En cas de problème, une nouvelle LIF peut être établie sur des nœuds supplémentaires et l'hôte peut être mis à jour dans la fenêtre de maintenance suivante afin de supprimer le trafic indirect de la configuration.

Configuration TCP/IP et ethernet

De nombreux clients d'Oracle sur ONTAP utilisent ethernet, le protocole réseau de NFS, iSCSI, NVMe/TCP, en particulier le cloud.

Paramètres du système d'exploitation hôte

La plupart des documents des fournisseurs d'applications incluent des paramètres TCP et ethernet spécifiques destinés à garantir le fonctionnement optimal de l'application. Ces mêmes paramètres suffisent généralement pour assurer des performances de stockage IP optimales.

Contrôle de flux Ethernet

Cette technologie permet à un client de demander à un expéditeur d'arrêter temporairement la transmission de données. Cela est généralement fait parce que le récepteur est incapable de traiter les données entrantes assez rapidement. À un moment donné, demander à un expéditeur de cesser la transmission était moins perturbant que d'avoir un récepteur de paquets de rejet parce que les tampons étaient pleins. Ce n'est plus le cas avec les piles TCP utilisées dans les systèmes d'exploitation d'aujourd'hui. En fait, le contrôle de flux cause plus de problèmes qu'il ne résout.

Les problèmes de performances causés par le contrôle de flux Ethernet ont augmenté ces dernières années. En effet, le contrôle de flux Ethernet fonctionne au niveau de la couche physique. Si une configuration réseau permet à un système d'exploitation hôte d'envoyer une demande de contrôle de flux Ethernet à un système de stockage, il en résulte une pause des E/S pour tous les clients connectés. Étant donné qu'un nombre croissant de clients sont servis par un seul contrôleur de stockage, la probabilité qu'un ou plusieurs de ces clients envoient des demandes de contrôle de flux augmente. Le problème a été fréquemment rencontré sur les sites des clients qui possèdent une virtualisation étendue du système d'exploitation.

Une carte réseau sur un système NetApp ne doit pas recevoir de demandes de contrôle de flux. La méthode utilisée pour obtenir ce résultat varie en fonction du fabricant du commutateur réseau. Dans la plupart des cas, le contrôle de flux sur un commutateur Ethernet peut être réglé sur `receive desired` ou `receive on`, ce qui signifie qu'une demande de contrôle de flux n'est pas transmise au contrôleur de stockage. Dans d'autres cas, la connexion réseau sur le contrôleur de stockage risque de ne pas permettre la désactivation du contrôle de flux. Dans ce cas, les clients doivent être configurés pour ne jamais envoyer de demandes de contrôle de flux, soit en changeant la configuration NIC sur le serveur hôte lui-même, soit en changeant les ports de commutateur auxquels le serveur hôte est connecté.

NetApp recommande de s'assurer que les contrôleurs de stockage NetApp ne reçoivent pas de paquets de contrôle de flux Ethernet. Pour ce faire, il est généralement possible de définir les ports de commutateur auxquels le contrôleur est connecté, mais certains matériels de commutateur ont des limites qui peuvent nécessiter des modifications côté client.

Tailles du MTU

L'utilisation de trames Jumbo a été démontrée afin d'améliorer les performances des réseaux 1 Gbit en réduisant la surcharge du processeur et du réseau, mais l'avantage n'est généralement pas significatif.

NetApp recommande d'implémenter des trames Jumbo lorsque cela est possible, à la fois pour réaliser des avantages potentiels en termes de performances et pour pérenniser la solution.

L'utilisation de trames Jumbo dans un réseau de 10 Gb est presque obligatoire. En effet, la plupart des implémentations de 10 Gbits atteignent une limite de paquets par seconde sans trames Jumbo avant d'atteindre le seuil de 10 Gbits. L'utilisation de trames jumbo améliore l'efficacité du traitement TCP/IP car elle permet au système d'exploitation, au serveur, aux cartes réseau et au système de stockage de traiter moins de paquets, mais des paquets plus volumineux. L'amélioration des performances varie d'une carte réseau à l'autre, mais elle est significative.

Dans le cas des implémentations de trames Jumbo, il est courant, mais incorrect, que tous les périphériques connectés doivent prendre en charge les trames Jumbo et que la taille MTU doit correspondre de bout en bout Au lieu de cela, les deux extrémités du réseau négocient la taille de trame la plus élevée mutuellement acceptable lors de l'établissement d'une connexion. Dans un environnement standard, un commutateur réseau est défini sur une taille MTU de 9 9216, le contrôleur NetApp est défini sur 9000 et les clients sont configurés sur une combinaison de 9000 et 1514. Les clients qui prennent en charge un MTU de 9 9000 peuvent utiliser des trames jumbo, et les clients qui ne peuvent prendre en charge que 1514 peuvent négocier une valeur inférieure.

Les problèmes avec cet arrangement sont rares dans un environnement complètement commuté. Cependant, dans un environnement routé, veillez à ce qu'aucun routeur intermédiaire ne soit forcé de fragmenter des trames jumbo.

NetApp recommande de configurer les éléments suivants :

- Les trames Jumbo sont souhaitables, mais non requises avec Ethernet 1 Gb (GbE).
- Les trames Jumbo sont requises pour des performances maximales avec 10GbE et plus rapides.

Paramètres TCP

Trois paramètres sont souvent mal configurés : les horodatages TCP, l'acquittement sélectif (SACK) et la mise à l'échelle de la fenêtre TCP. De nombreux documents obsolètes sur Internet recommandent de désactiver un ou plusieurs de ces paramètres pour améliorer les performances. Cette recommandation a été très utile il y a de nombreuses années, lorsque les capacités du processeur étaient beaucoup plus faibles et qu'il y avait un avantage à réduire la surcharge sur le traitement TCP chaque fois que cela était possible.

Cependant, avec les systèmes d'exploitation modernes, la désactivation de l'une de ces fonctionnalités TCP n'entraîne généralement aucun avantage détectable, tout en pouvant nuire aux performances. Dans les environnements réseau virtualisés, les performances peuvent être endommagées, car ces fonctionnalités sont nécessaires pour gérer efficacement la perte de paquets et les modifications de la qualité du réseau.

NetApp recommande d'activer les horodatages TCP, le SACK et la mise à l'échelle des fenêtres TCP sur l'hôte, et ces trois paramètres doivent être activés par défaut dans tout système d'exploitation actuel.

Configuration FC SAN

La configuration de FC SAN pour les bases de données Oracle consiste principalement à suivre les meilleures pratiques quotidiennes en matière de SAN.

Il s'agit notamment de mesures de planification courantes, telles que l'assurance d'une bande passante suffisante sur le SAN entre l'hôte et le système de stockage, la vérification de la présence de tous les chemins SAN entre tous les périphériques requis, l'utilisation des paramètres de port FC requis par le fournisseur du commutateur FC, la prévention des conflits ISL, et à l'utilisation d'un système de surveillance de la structure SAN approprié.

Segmentation

Une zone FC ne doit jamais contenir plusieurs initiateurs. Un tel arrangement peut sembler fonctionner au départ, mais la diaphonie entre les initiateurs finit par interférer avec la performance et la stabilité.

Les zones à cibles multiples sont généralement considérées comme sûres, bien que dans de rares circonstances le comportement des ports cibles FC de fournisseurs différents ait causé des problèmes. Par exemple, évitez d'inclure les ports cibles d'une baie de stockage NetApp et non NetApp dans la même zone. En outre, le fait de placer un système de stockage NetApp et un dispositif de bande dans la même zone est encore plus susceptible de causer des problèmes.

Réseau à connexion directe

Les administrateurs du stockage préfèrent parfois simplifier leurs infrastructures en supprimant les commutateurs réseau de la configuration. Cela peut être pris en charge dans certains scénarios.

ISCSI et NVMe/TCP

Un hôte utilisant iSCSI ou NVMe/TCP peut être directement connecté à un système de stockage et fonctionner normalement. La raison en est le chemin d'accès. Les connexions directes à deux contrôleurs de stockage distincts donnent lieu à deux chemins de flux de données indépendants. La perte du chemin, du port ou du contrôleur n'empêche pas l'autre chemin d'être utilisé.

NFS

Vous pouvez utiliser un stockage NFS à connexion directe, mais avec une limitation importante : le basculement ne fonctionnera pas sans script important, ce qui incombera au client.

Ce qui complique la reprise après incident avec un stockage NFS à connexion directe, c'est le routage qui se produit sur le système d'exploitation local. Par exemple, supposons qu'un hôte a une adresse IP 192.168.1.1/24 et qu'il est directement connecté à un contrôleur ONTAP avec une adresse IP 192.168.1.50/24. Lors du basculement, cette adresse 192.168.1.50 peut basculer vers l'autre contrôleur et sera disponible pour l'hôte, mais comment l'hôte peut-il détecter sa présence ? L'adresse 192.168.1.1 d'origine existe toujours sur la carte réseau hôte qui ne se connecte plus à un système opérationnel. Le trafic destiné à 192.168.1.50 continuerait d'être envoyé à un port réseau inutilisable.

Le second NIC du système d'exploitation peut être configuré sur 19 2.168.1.2 et serait capable de communiquer avec l'adresse en panne sur 192.168.1.50, mais les tables de routage locales auraient par défaut l'utilisation d'une adresse **et d'une seule adresse** pour communiquer avec le sous-réseau 192.168.1.0/24. Un administrateur système pourrait créer un framework de scripts qui détecterait une connexion réseau défaillante et modifierait les tables de routage locales ou rendrait les interfaces « up and down ». La procédure exacte dépend du système d'exploitation utilisé.

Dans la pratique, les clients NetApp disposent d'un protocole NFS à connexion directe, mais généralement uniquement pour les charges de travail où une pause des E/S est acceptable pendant les basculements. Lorsque des montages durs sont utilisés, aucune erreur d'E/S ne doit se produire lors de ces pauses. L'E/S doit se bloquer jusqu'à ce que les services soient restaurés, soit par un retour arrière, soit par une intervention manuelle pour déplacer les adresses IP entre les cartes réseau de l'hôte.

Connexion directe FC

Il n'est pas possible de connecter directement un hôte à un système de stockage ONTAP à l'aide du protocole FC. La raison en est l'utilisation de NPIV. Le WWN qui identifie un port FC ONTAP sur le réseau FC utilise un type de virtualisation appelé NPIV. Tout périphérique connecté à un système ONTAP doit pouvoir reconnaître un WWN NPIV. Aucun fournisseur actuel de HBA ne propose de HBA pouvant être installé sur un hôte et capable de prendre en charge une cible NPIV.

Configuration réseau sur les systèmes ASA r2

Interfaces logiques

Les bases de données Oracle doivent accéder au stockage. Les interfaces logiques (LIF) correspondent à la tuyauterie réseau qui connecte une machine virtuelle de stockage (SVM) au réseau et, par conséquent, à la base de données. Une conception correcte des LIF est requise pour s'assurer qu'il y a suffisamment de bande passante pour chaque charge de travail de la base de données, et le basculement ne provoque pas de perte des services de stockage.

Cette section présente un aperçu des principaux principes de conception LIF pour les systèmes ASA r2, qui sont optimisés pour les environnements SAN uniquement. Pour une documentation plus complète, consultez le "Documentation de gestion de réseau ONTAP". Comme pour d'autres aspects de l'architecture de base de données, les meilleures options pour la conception de la machine virtuelle de stockage (SVM, connue sous le nom de vserver dans l'interface de ligne de commande) et de l'interface logique (LIF) dépendent fortement des exigences d'évolutivité et des besoins de l'entreprise.

Tenez compte des principaux sujets suivants lors de l'élaboration d'une stratégie LIF :

- **Performance.** La bande passante du réseau est-elle suffisante pour les charges de travail Oracle ?
- **Résilience.** y a-t-il des points de défaillance uniques dans la conception?
- **Gérabilité.** le réseau peut-il être mis à l'échelle sans interruption ?

Ces rubriques s'appliquent à la solution de bout en bout, de l'hôte aux commutateurs et au système de stockage.

Types de LIF

Il existe plusieurs types de LIF. "Documentation ONTAP sur les types de LIF" Fournir des informations plus complètes à ce sujet, mais d'un point de vue fonctionnel, les LIF peuvent être divisées en plusieurs groupes :

- **LIFs de gestion de clusters et de nœuds.** utilisées pour gérer le cluster de stockage.
- **LIF de gestion SVM.** interfaces permettant l'accès à une SVM via l'API REST ou ONTAPI (aussi connue sous le nom de ZAPI) pour des fonctions telles que la création de snapshots ou le redimensionnement de volumes. Des produits tels que SnapManager pour Oracle (SMO) doivent avoir accès à une LIF de gestion SVM.
- **LIF de données.** Interfaces pour protocoles SAN uniquement : FC, iSCSI, NVMe/FC, NVMe/TCP. Les protocoles NAS (NFS, SMB/CIFS) ne sont pas pris en charge sur les systèmes ASA r2.

> Il n'est pas possible de configurer une interface à la fois pour le trafic iSCSI (ou NVMe/TCP) et le trafic de gestion, bien que les deux utilisent un protocole IP. Une interface LIF de gestion distincte est requise dans les environnements iSCSI ou NVMe/TCP. Pour garantir la résilience et les performances, configurez plusieurs LIF de données SAN par protocole et par nœud, et répartissez-les sur différents ports et infrastructures physiques. Contrairement aux systèmes AFF/ FAS , ASA r2 n'autorise pas le trafic NFS ou SMB, il n'y a donc aucune possibilité de réutiliser une LIF de données NAS pour la gestion.

Conception de SAN LIF

La conception de LIF dans un environnement SAN est relativement simple pour une raison : les chemins d'accès multiples. Toutes les implémentations SAN modernes permettent à un client d'accéder aux données sur plusieurs chemins réseau indépendants et de sélectionner le ou les chemins d'accès les plus adaptés. Par conséquent, les performances du design LIF sont plus simples à gérer, car les clients SAN équilibrent automatiquement la charge en E/S sur les meilleurs chemins disponibles.

Si un chemin devient indisponible, le client sélectionne automatiquement un autre chemin. La simplicité de conception qui en résulte rend les LIF SAN généralement plus faciles à gérer. Cela ne signifie pas pour autant qu'un environnement SAN est toujours plus facile à gérer, car de nombreux autres aspects du stockage SAN sont bien plus complexes que NFS. Cela signifie simplement que la conception de la LIF SAN est plus facile.

Performance

Le facteur le plus important à prendre en compte concernant les performances des interfaces LIF dans un environnement SAN est la bande passante. Par exemple, un cluster ASA r2 à deux nœuds avec deux ports FC 32Gb par nœud permet jusqu'à 64Gb de bande passante vers/depuis chaque nœud. De même, pour NVMe/TCP ou iSCSI, assurez-vous d'une connectivité suffisante de 25 GbE ou 100 GbE pour les charges de travail Oracle.

Résilience

Les interfaces logiques SAN ne basculent pas de la même manière que les interfaces logiques NAS. Les systèmes ASA r2 s'appuient sur le multipathing hôte (MPIO/ALUA) pour assurer la résilience. Si une interface LIF SAN devient indisponible en raison d'un basculement du contrôleur, le logiciel de gestion de chemins multiples du client détecte la perte d'un chemin et redirige les E/S vers un chemin alternatif. ASA r2 peut effectuer une relocalisation LIF après un court délai pour rétablir la disponibilité complète du chemin, mais cela n'interrompt pas les E/S car des chemins actifs existent déjà sur le nœud partenaire. Le processus de basculement a lieu afin de rétablir l'accès à l'hôte sur tous les ports définis.

Gestion aisée

Il n'est pas nécessaire de migrer une LIF dans un environnement SAN lorsque les volumes sont déplacés au sein de la paire HA. En effet, une fois le déplacement du volume terminé, ONTAP envoie une notification au SAN concernant un changement de chemin, et les clients SAN se réoptimisent automatiquement. La migration LIF vers SAN est principalement associée à des changements matériels physiques majeurs. Par exemple, si

une mise à niveau non perturbatrice des contrôleurs est nécessaire, une interface SAN LIF est migrée vers le nouveau matériel. Si un port FC s'avère défectueux, une LIF peut être migrée vers un port inutilisé.

Recommandations de conception

NetApp formule les recommandations suivantes pour les environnements SAN ASA r2 :

- Ne créez pas plus de chemins que nécessaire. Un nombre excessif de chemins complique la gestion globale et peut entraîner des problèmes de basculement de chemin sur certains hôtes. De plus, certains hôtes ont des limites de chemin inattendues pour les configurations comme le démarrage SAN.
- Très peu de configurations doivent nécessiter plus de quatre chemins vers une LUN. L'intérêt d'avoir plus de deux nœuds de chemins publicitaires vers les LUN est limité, car l'agrégat hébergeant une LUN est inaccessible en cas de défaillance du nœud qui détient la LUN et de son partenaire haute disponibilité. Dans ce cas, la création de chemins sur des nœuds autres que la paire haute disponibilité principale n'est pas utile.
- Même si vous pouvez gérer le nombre de chemins de LUN visibles en sélectionnant les ports inclus dans les zones FC, il est généralement plus facile d'inclure tous les points cibles potentiels dans la zone FC et de contrôler la visibilité des LUN au niveau des ONTAP.
- Utilisez la fonction de mappage LUN sélectif (SLM), activée par défaut. Avec SLM, tout nouveau LUN est automatiquement annoncé à partir du nœud qui possède l'agrégat sous-jacent et du partenaire HA du nœud. Cette configuration évite d'avoir à créer des ensembles de ports ou à configurer un zonage pour limiter l'accessibilité des ports. Chaque LUN est disponible sur le nombre minimal de nœuds requis pour des performances et une résilience optimales.
- Si une LUN doit être migrée en dehors des deux contrôleurs, les nœuds supplémentaires peuvent être ajoutés avec le `lun mapping add-reporting-nodes` commande permettant d'annoncer les LUN sur les nouveaux nœuds. Cela crée des chemins SAN supplémentaires vers les LUN pour la migration des LUN. Toutefois, l'hôte doit effectuer une opération de découverte pour utiliser les nouveaux chemins.
- Ne vous souciez pas trop du trafic indirect. Dans un environnement très exigeant en E/S, il est préférable d'éviter le trafic indirect pour lequel chaque microseconde de latence est critique, mais l'impact visible sur la performance est négligeable pour les charges de travail classiques.

Configuration TCP/IP et ethernet

De nombreux clients Oracle sur ASA r2 ONTAP utilisent Ethernet, le protocole réseau d'iSCSI et de NVMe/TCP.

Paramètres du système d'exploitation hôte

La plupart des documents des fournisseurs d'applications incluent des paramètres TCP et ethernet spécifiques destinés à garantir le fonctionnement optimal de l'application. Ces mêmes paramètres suffisent généralement pour assurer des performances de stockage IP optimales.

Contrôle de flux Ethernet

Cette technologie permet à un client de demander à un expéditeur d'arrêter temporairement la transmission de données. Cela est généralement fait parce que le récepteur est incapable de traiter les données entrantes assez rapidement. À un moment donné, demander à un expéditeur de cesser la transmission était moins perturbant que d'avoir un récepteur de paquets de rejet parce que les tampons étaient pleins. Ce n'est plus le cas avec les piles TCP utilisées dans les systèmes d'exploitation d'aujourd'hui. En fait, le contrôle de flux cause plus de problèmes qu'il ne résout.

Les problèmes de performances causés par le contrôle de flux Ethernet ont augmenté ces dernières années.

En effet, le contrôle de flux Ethernet fonctionne au niveau de la couche physique. Si une configuration réseau permet à un système d'exploitation hôte d'envoyer une demande de contrôle de flux Ethernet à un système de stockage, il en résulte une pause des E/S pour tous les clients connectés. Étant donné qu'un nombre croissant de clients sont servis par un seul contrôleur de stockage, la probabilité qu'un ou plusieurs de ces clients envoient des demandes de contrôle de flux augmente. Le problème a été fréquemment rencontré sur les sites des clients qui possèdent une virtualisation étendue du système d'exploitation.

Une carte réseau sur un système NetApp ne doit pas recevoir de demandes de contrôle de flux. La méthode utilisée pour obtenir ce résultat varie en fonction du fabricant du commutateur réseau. Dans la plupart des cas, le contrôle de flux sur un commutateur Ethernet peut être réglé sur `receive desired` ou `receive on`, ce qui signifie qu'une demande de contrôle de flux n'est pas transmise au contrôleur de stockage. Dans d'autres cas, la connexion réseau sur le contrôleur de stockage risque de ne pas permettre la désactivation du contrôle de flux. Dans ce cas, les clients doivent être configurés pour ne jamais envoyer de demandes de contrôle de flux, soit en changeant la configuration NIC sur le serveur hôte lui-même, soit en changeant les ports de commutateur auxquels le serveur hôte est connecté.

Pour les systèmes ASA r2, qui sont exclusivement des SAN, les considérations relatives au contrôle de flux Ethernet s'appliquent principalement au trafic iSCSI et NVMe/TCP.

* NetApp recommande* de s'assurer que les contrôleurs de stockage NetApp ASA r2 ne reçoivent pas de paquets de contrôle de flux Ethernet. Cela peut généralement se faire en configurant les ports du commutateur auxquels le contrôleur est connecté, mais certains matériels de commutation présentent des limitations qui peuvent nécessiter des modifications côté client.

Tailles du MTU

L'utilisation de trames Jumbo a été démontrée afin d'améliorer les performances des réseaux 1 Gbit en réduisant la surcharge du processeur et du réseau, mais l'avantage n'est généralement pas significatif.

NetApp recommande d'implémenter des trames Jumbo lorsque cela est possible, à la fois pour réaliser des avantages potentiels en termes de performances et pour pérenniser la solution.

Pour les systèmes ASA r2, qui sont exclusivement SAN, les trames jumbo s'appliquent uniquement aux protocoles SAN basés sur Ethernet (iSCSI et NVMe/TCP).

L'utilisation de trames Jumbo dans un réseau de 10 Gb est presque obligatoire. En effet, la plupart des implémentations de 10 Gbits atteignent une limite de paquets par seconde sans trames Jumbo avant d'atteindre le seuil de 10 Gbits. L'utilisation de trames jumbo améliore l'efficacité du traitement TCP/IP car elle permet au système d'exploitation, au serveur, aux cartes réseau et au système de stockage de traiter moins de paquets, mais des paquets plus volumineux. L'amélioration des performances varie d'une carte réseau à l'autre, mais elle est significative.

Dans le cas des implémentations de trames Jumbo, il est courant, mais incorrect, que tous les périphériques connectés doivent prendre en charge les trames Jumbo et que la taille MTU doit correspondre de bout en bout Au lieu de cela, les deux extrémités du réseau négocient la taille de trame la plus élevée mutuellement acceptable lors de l'établissement d'une connexion. Dans un environnement standard, un commutateur réseau est défini sur une taille MTU de 9 9216, le contrôleur NetApp est défini sur 9000 et les clients sont configurés sur une combinaison de 9000 et 1514. Les clients qui prennent en charge un MTU de 9 9000 peuvent utiliser des trames jumbo, et les clients qui ne peuvent prendre en charge que 1514 peuvent négocier une valeur inférieure.

Les problèmes avec cet arrangement sont rares dans un environnement complètement commuté. Cependant, dans un environnement routé, veillez à ce qu'aucun routeur intermédiaire ne soit forcé de fragmenter des

trames jumbo.

NetApp recommande la configuration suivante pour les environnements SAN ASA r2 :

- Les trames jumbo sont souhaitables mais non obligatoires avec le 1 GbE.
- Les trames Jumbo sont nécessaires pour des performances optimales avec une vitesse de 10 GbE et plus pour le trafic iSCSI et NVMe/TCP.

Paramètres TCP

Trois paramètres sont souvent mal configurés : les horodatages TCP, l'acquittement sélectif (SACK) et la mise à l'échelle de la fenêtre TCP. De nombreux documents obsolètes sur Internet recommandent de désactiver un ou plusieurs de ces paramètres pour améliorer les performances. Cette recommandation a été très utile il y a de nombreuses années, lorsque les capacités du processeur étaient beaucoup plus faibles et qu'il y avait un avantage à réduire la surcharge sur le traitement TCP chaque fois que cela était possible.

Cependant, avec les systèmes d'exploitation modernes, la désactivation de l'une de ces fonctionnalités TCP n'entraîne généralement aucun avantage détectable, tout en pouvant nuire aux performances. Dans les environnements réseau virtualisés, les performances peuvent être endommagées, car ces fonctionnalités sont nécessaires pour gérer efficacement la perte de paquets et les modifications de la qualité du réseau.

NetApp recommande d'activer les horodatages TCP, le SACK et la mise à l'échelle des fenêtres TCP sur l'hôte, et ces trois paramètres doivent être activés par défaut dans tout système d'exploitation actuel.

Configuration FC SAN

La configuration d'un SAN FC pour les bases de données Oracle sur les systèmes ASA r2 consiste principalement à suivre les meilleures pratiques SAN standard.

ASA r2 est optimisée pour les charges de travail SAN uniquement, les principes restent donc les mêmes que pour AFF/ FAS, avec un accent mis sur les performances, la résilience et la simplicité. Cela inclut des mesures de planification classiques telles que s'assurer de l'existence d'une bande passante suffisante sur le SAN entre l'hôte et le système de stockage, vérifier que tous les chemins SAN existent entre tous les périphériques requis, utiliser les paramètres de port FC requis par votre fournisseur de commutateur FC, éviter la contention ISL et utiliser une surveillance appropriée de la structure SAN.

Segmentation

Une zone FC ne doit jamais contenir plusieurs initiateurs. Un tel arrangement peut sembler fonctionner au départ, mais la diaphonie entre les initiateurs finit par interférer avec la performance et la stabilité.

Les zones à cibles multiples sont généralement considérées comme sûres, bien que dans de rares circonstances le comportement des ports cibles FC de fournisseurs différents ait causé des problèmes. Par exemple, évitez d'inclure les ports cibles d'une baie de stockage NetApp et non NetApp dans la même zone. En outre, le fait de placer un système de stockage NetApp et un dispositif de bande dans la même zone est encore plus susceptible de causer des problèmes.

- ASA r2 utilise des zones de disponibilité de stockage au lieu d'agrégats, mais cela ne change pas les principes de zonage FC.
- Le multipathing (MPIO) reste le principal mécanisme de résilience ; cependant, pour les systèmes ASA r2 qui prennent en charge le multipathing actif-actif symétrique, tous les chemins vers un LUN sont actifs et utilisés simultanément pour les E/S.

Réseau à connexion directe

Les administrateurs du stockage préfèrent parfois simplifier leurs infrastructures en supprimant les commutateurs réseau de la configuration. Cela peut être pris en charge dans certains scénarios.

ISCSI et NVMe/TCP

Un hôte utilisant iSCSI ou NVMe/TCP peut être connecté directement à un système de stockage ASA r2 et fonctionner normalement. La raison est liée au cheminement. Les connexions directes à deux contrôleurs de stockage différents créent deux chemins indépendants pour le flux de données. La perte d'un chemin, d'un port ou d'un contrôleur n'empêche pas l'utilisation de l'autre chemin, à condition que le multipathing soit correctement configuré.

Connexion directe FC

Il n'est pas possible de connecter directement un hôte à un système de stockage ASA r2 en utilisant le protocole FC. La raison est la même qu'avec les systèmes AFF/ FAS , l'utilisation de NPIV. Le WWN qui identifie un port ONTAP FC sur le réseau FC utilise un type de virtualisation appelé NPIV. Tout appareil connecté à un système ONTAP doit être capable de reconnaître un WWN NPIV. Il n'existe actuellement aucun fournisseur de HBA proposant un HBA pouvant être installé sur un hôte capable de prendre en charge une cible NPIV.

Configuration du stockage sur les systèmes AFF/FAS

SAN FC

Alignement de LUN

L'alignement des LUN fait référence à l'optimisation des E/S par rapport à la disposition du système de fichiers sous-jacent.

Sur un système ONTAP, le stockage est organisé en unités de 4 Ko. Un bloc de 8 Ko de base de données ou de système de fichiers doit être mappé à exactement deux blocs de 4 Ko. Si une erreur de configuration de LUN déplace l'alignement de 1 Ko dans les deux sens, chaque bloc de 8 Ko existerait sur trois blocs de stockage de 4 Ko différents au lieu de deux. Cette configuration entraîne une augmentation de la latence et des E/S supplémentaires au sein du système de stockage.

L'alignement affecte également les architectures LVM. Si un volume physique au sein d'un groupe de volumes logiques est défini sur l'unité entière (aucune partition n'est créée), le premier bloc de 4 Ko de la LUN s'aligne sur le premier bloc de 4 Ko du système de stockage. Il s'agit d'un alignement correct. Des problèmes surviennent avec les partitions car elles déplacent l'emplacement de départ où le système d'exploitation utilise le LUN. Tant que le décalage est décalé en unités entières de 4 Ko, la LUN est alignée.

Dans les environnements Linux, créez des groupes de volumes logiques sur l'ensemble de l'unité de disque.

Lorsqu'une partition est requise, vérifiez l'alignement en exécutant et en `fdisk -u` vérifiant que le début de chaque partition est un multiple de huit. Cela signifie que la partition démarre à un multiple de huit secteurs de 512 octets, soit 4 Ko.

Voir également la discussion sur l'alignement des blocs de compression dans la section "Efficacité". Toute disposition alignée avec les limites des blocs de compression de 8 Ko est également alignée avec les limites de 4 Ko.

Avertissements de mauvais alignement

La journalisation des opérations de reprise et des transactions de la base de données génère normalement des E/S non alignées qui peuvent entraîner des avertissements erronés concernant les LUN mal alignées sur ONTAP.

La journalisation effectue une écriture séquentielle du fichier journal avec des écritures de taille variable. Une opération d'écriture de journal qui ne s'aligne pas sur les limites de 4 Ko ne provoque généralement pas de problèmes de performances, car l'opération d'écriture de journal suivante termine le bloc. ONTAP est ainsi en mesure de traiter la quasi-totalité des écritures sous forme de blocs complets de 4 Ko, même si les données de blocs de 4 Ko ont été écrites dans deux opérations distinctes.

Vérifiez l'alignement à l'aide d'utilitaires tels que `sio` ou `dd` Qui peuvent générer des E/S à une taille de bloc définie. Les statistiques d'alignement des E/S sur le système de stockage peuvent être affichées à l'aide du `stats` commande. Voir "Vérification de l'alignement WAFL" pour en savoir plus.

L'alignement dans les environnements Solaris est plus compliqué. Reportez-vous à la section "Configuration de l'hôte SAN ONTAP" pour en savoir plus.

Avertissement
Dans les environnements Solaris x86, prenez davantage soin de l'alignement approprié car la plupart des configurations comportent plusieurs couches de partitions. Les tranches de partition Solaris x86 existent généralement au-dessus d'une table de partition d'enregistrement d'amorçage maître standard.

Dimensionnement des LUN et nombre de LUN

Il est essentiel de sélectionner la taille de LUN optimale et le nombre de LUN à utiliser pour optimiser les performances et la gestion des bases de données Oracle.

Une LUN est un objet virtualisé sur ONTAP qui existe sur tous les disques de l'agrégat d'hébergement. Par conséquent, les performances de la LUN ne sont pas affectées par sa taille, car la LUN exploite tout le potentiel de performance de l'agrégat, quelle que soit sa taille.

À titre de commodité, les clients peuvent souhaiter utiliser un LUN de taille spécifique. Par exemple, si une base de données est construite sur un groupe de disques LVM ou Oracle ASM composé de deux LUN de 1 To chacune, ce groupe de disques doit être développé par incréments de 1 To. Il peut être préférable de créer le groupe de disques à partir de huit LUN de 500 Go chacune, de sorte que le groupe de disques puisse être augmenté par incréments plus petits.

Il n'est pas recommandé d'établir une taille de LUN standard universelle, car cela peut compliquer la gestion. Par exemple, une taille de LUN standard de 100 Go peut fonctionner correctement lorsqu'une base de données ou un datastore se situe entre 1 et 2 To, mais qu'une base de données ou un datastore de 20 To nécessite 200 LUN. Cela signifie que les délais de redémarrage du serveur sont plus longs, que les différents utilisateurs doivent gérer davantage d'objets et que des produits tels que SnapCenter doivent effectuer des recherches sur de nombreux objets. L'utilisation d'un nombre inférieur de LUN de plus grande taille permet

d’éviter de tels problèmes.

- Le nombre de LUN est plus important que la taille de LUN.
- La taille de LUN est principalement contrôlée par les exigences liées au nombre de LUN.
- Évitez de créer plus de LUN que nécessaire.

Nombre de LUN

Contrairement à la taille de LUN, le nombre de LUN affecte les performances. La performance des applications dépend souvent de la capacité à réaliser des E/S parallèles via la couche SCSI. Ainsi, deux LUN offrent de meilleures performances qu’une seule LUN. L’utilisation d’un LVM tel que Veritas VxVM, Linux LVM2 ou Oracle ASM est la méthode la plus simple pour augmenter le parallélisme.

Les clients NetApp n’ont généralement pas eu l’avantage d’augmenter le nombre de LUN au-delà de seize. Toutefois, le test d’environnements 100 % SSD avec des E/S aléatoires très lourdes a permis d’améliorer encore jusqu’à 64 LUN.

NetApp recommande ce qui suit :

En général, de quatre à seize LUN suffisent pour prendre en charge les besoins en E/S d’une charge de travail de base de données donnée. Moins de quatre LUN peuvent créer des limites de performances en raison de limites dans les implémentations SCSI hôte.

Placement des LUN

Le placement optimal des LUN de base de données dans les volumes ONTAP dépend principalement de l’utilisation des différentes fonctionnalités ONTAP.

Volumes

L’un des points de confusion les plus courants avec les nouveaux clients ONTAP est l’utilisation des volumes FlexVol, communément appelés simplement « volumes ».

Un volume n’est pas une LUN. Ces termes sont utilisés de façon synonymie avec de nombreux autres produits de fournisseurs, y compris les fournisseurs de cloud. Les volumes ONTAP sont des conteneurs de gestion simples. Ils ne fournissent pas les données en eux-mêmes, ni n’occupent l’espace. Il s’agit de conteneurs pour les fichiers et les LUN. Ils permettent d’améliorer et de simplifier la gestion, notamment à grande échelle.

Volumes et LUN

Les LUN associées sont généralement situées en colocation dans un seul volume. Par exemple, une base de données qui nécessite 10 LUN doit généralement avoir les 10 LUN placées sur le même volume.

- L'utilisation d'un rapport LUN/volumes de 1:1, c'est-à-dire une LUN par volume, n'est **pas** une bonne pratique formelle.
- À la place, les volumes doivent être considérés comme des conteneurs pour les charges de travail ou les datasets. Il peut y avoir une seule LUN par volume ou il peut y en avoir plusieurs. La bonne réponse dépend des exigences de gestion.
- La diffusion des LUN sur un nombre inutile de volumes peut entraîner une surcharge supplémentaire et des problèmes de planification pour des opérations telles que les opérations de snapshot, un nombre excessif d'objets affichés dans l'interface utilisateur et entraîner l'atteinte des limites de volume de la plate-forme avant que la limite de LUN ne soit atteinte.

Volumes, LUN et snapshots

Les règles et planifications Snapshot sont placées sur le volume, et non sur la LUN. Un jeu de données composé de 10 LUN ne nécessite qu'une seule règle de snapshot lorsque ces LUN sont co-localisées dans le même volume.

En outre, la colocation de toutes les LUN associées à un jeu de données donné dans un seul volume permet d'effectuer des opérations de snapshot atomiques. Par exemple, une base de données résidant sur 10 LUN ou un environnement d'application VMware comprenant 10 systèmes d'exploitation différents peut être protégé comme un objet unique et cohérent si les LUN sous-jacentes sont tous placés sur un seul volume. S'ils sont placés sur des volumes différents, les snapshots peuvent être synchronisés à 100 %, même s'ils sont programmés en même temps.

Dans certains cas, il peut être nécessaire de diviser un jeu de LUN associé en deux volumes différents en raison des exigences de restauration. Par exemple, une base de données peut contenir quatre LUN pour les fichiers de données et deux LUN pour les journaux. Dans ce cas, un volume de fichiers de données avec 4 LUN et un volume de journaux avec 2 LUN peuvent être la meilleure option. La raison en est une capacité de restauration indépendante. Par exemple, le volume des fichiers de données peut être restauré de manière sélective à un état antérieur, ce qui signifie que les quatre LUN seraient rétablies à l'état du snapshot, tandis que le volume du journal contenant ses données stratégiques ne serait pas affecté.

Volumes, LUN et SnapMirror

Les règles et opérations SnapMirror sont, tout comme les opérations Snapshot, exécutées sur le volume, et non sur la LUN.

La colocation de LUN associées dans un seul volume vous permet de créer une relation SnapMirror unique et de mettre à jour toutes les données qu'elle contient en une seule mise à jour. Comme pour les instantanés, la mise à jour sera également une opération atomique. La destination SnapMirror dispose d'une réplique instantanée unique des LUN source. Si les LUN ont été réparties sur plusieurs volumes, les répliques peuvent être cohérentes les unes avec les autres.

Volumes, LUN et QoS

S'il est possible d'appliquer la QoS de manière sélective à chaque LUN, il est généralement plus facile de la configurer au niveau du volume. Par exemple, toutes les LUN utilisées par les invités dans un serveur ESX donné peuvent être placées sur un seul volume, puis une règle de qualité de service adaptative de ONTAP peut être appliquée. Vous obtenez ainsi une limite d'IOPS par To qui s'applique à toutes les LUN.

De même, si une base de données nécessitait 100 000 IOPS et occupait 10 LUN, il serait plus facile de définir une seule limite de 100 000 IOPS sur un seul volume que de définir 10 limites individuelles de 10 000 IOPS, une sur chaque LUN.

Dispositions multi-volumes

Dans certains cas, la distribution de LUN sur plusieurs volumes peut être avantageuse. La principale raison est la répartition des contrôleurs. Par exemple, un système de stockage haute disponibilité peut héberger une base de données unique dans laquelle chaque contrôleur a besoin du potentiel de traitement et de mise en cache complet. Dans ce cas, la conception type consisterait à placer la moitié des LUN dans un seul volume sur le contrôleur 1 et l'autre moitié des LUN dans un seul volume sur le contrôleur 2.

De même, la répartition des contrôleurs peut être utilisée pour l'équilibrage de la charge. Un système haute disponibilité hébergeant 100 bases de données de 10 LUN chacune peut être conçu où chaque base de données reçoit un volume de 5 LUN sur chacun des deux contrôleurs. Il en résulte une charge symétrique garantie de chaque contrôleur au fur et à mesure que des bases de données supplémentaires sont provisionnées.

Cependant, aucun de ces exemples ne correspond à un ratio volume/LUN de 1:1. L'objectif reste d'optimiser la gestion en co-localisant les LUN associées dans les volumes.

Par exemple, la conteneurisation est un rapport LUN/volume 1:1. Chaque LUN peut représenter une seule charge de travail et doit être gérée individuellement. Dans ce cas, un rapport de 1:1 peut être optimal.

Redimensionnement des LUN et LVM

Lorsqu'un système de fichiers SAN a atteint sa limite de capacité, il existe deux options pour augmenter l'espace disponible :

- Augmentez la taille des LUN
- Ajoutez une LUN à un groupe de volumes existant et développez le volume logique contenu

Bien que le redimensionnement des LUN soit une option d'augmentation de la capacité, il est généralement préférable d'utiliser un LVM, y compris Oracle ASM. L'une des principales raisons pour lesquelles les LVM existent est d'éviter la nécessité d'un redimensionnement des LUN. Avec une LVM, plusieurs LUN sont reliées entre elles dans un pool de stockage virtuel. Les volumes logiques extraits de ce pool sont gérés par le LVM et peuvent être facilement redimensionnés. Il est également possible d'éviter les points sensibles sur un disque en distribuant un volume logique donné à tous les LUN disponibles. Une migration transparente peut généralement être effectuée à l'aide du gestionnaire de volumes pour déplacer les extensions sous-jacentes d'un volume logique vers de nouvelles LUN.

Répartition LVM

La répartition des LVM consiste à distribuer les données entre plusieurs LUN. Les performances de nombreuses bases de données en sont ainsi considérablement améliorées.

Avant l'ère des disques Flash, la répartition était utilisée pour surmonter les limites de performances des disques rotatifs. Par exemple, si un système d'exploitation doit effectuer une opération de lecture de 1 Mo, la lecture de ce 1 Mo de données à partir d'un seul disque demande beaucoup de tête de lecture lorsque le transfert des 1 Mo est lent. Si ce 1 Mo de données a été réparti sur 8 LUN, le système d'exploitation pourrait exécuter huit opérations de lecture de 128 K en parallèle et réduire le temps nécessaire au transfert de 1 Mo.

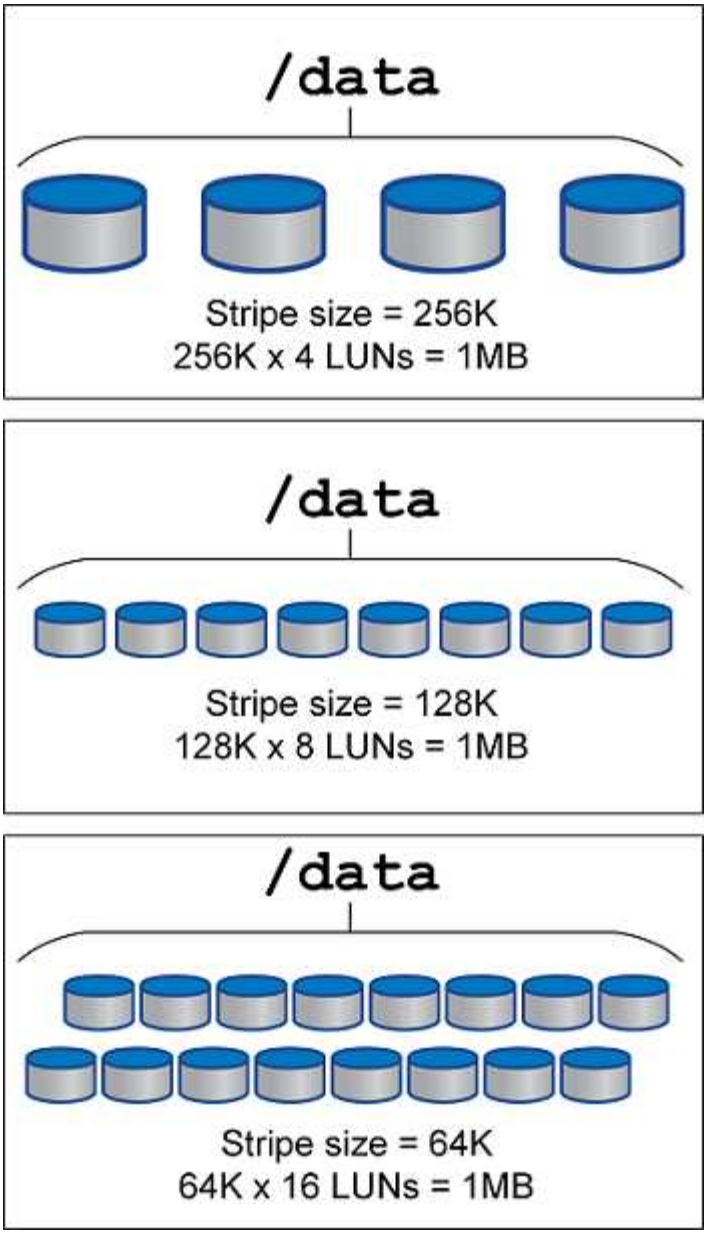
Le striping avec des disques rotatifs était plus difficile, car le modèle d'E/S devait être connu à l'avance. Si la répartition n'a pas été correctement réglée pour les véritables modèles d'E/S, les configurations à bandes risquent d'endommager les performances. Avec les bases de données Oracle, et en particulier les configurations 100 % Flash, le striping est beaucoup plus facile à configurer et a fait ses preuves pour améliorer considérablement les performances.

Par défaut, les gestionnaires de volumes logiques, tels que la bande Oracle ASM, ne le font pas pour le système d'exploitation natif LVM. Certaines lient plusieurs LUN ensemble en tant que périphérique concaténé. Résultat : des fichiers de données existent sur un seul périphérique LUN. Ceci provoque des points chauds. Les autres implémentations LVM prennent par défaut en charge les extensions distribuées. Cette méthode est similaire à la répartition, mais elle est plus grossière. Les LUN du groupe de volumes sont tranchées en grandes parties, appelées extensions et généralement mesurées en plusieurs mégaoctets. Ensuite, les volumes logiques sont distribués sur ces extensions. Il en résulte des E/S aléatoires sur un fichier qui doit être bien réparti entre les LUN, mais les opérations d'E/S séquentielles ne sont pas aussi efficaces qu'elles pourraient l'être.

Les E/S des applications exigeantes en performances sont presque toujours de (a) en unités de taille de bloc de base ou (b) d'un mégaoctet.

L'objectif principal d'une configuration à bandes est de s'assurer que les E/S de fichier unique peuvent être exécutées comme une seule unité, et que les E/S de plusieurs blocs, d'une taille de 1 Mo, peuvent être parallélisées de façon homogène sur toutes les LUN du volume réparti. Cela signifie que la taille de bande ne doit pas être inférieure à la taille du bloc de base de données, et que la taille de bande multipliée par le nombre de LUN doit être de 1 Mo.

La figure suivante présente trois options possibles pour le réglage de la taille et de la largeur des bandes. Le nombre de LUN est sélectionné pour répondre aux exigences de performances comme décrit ci-dessus, mais dans tous les cas, le total des données dans une seule bande est de 1 Mo.



NFS

Présentation

NetApp fournit un stockage NFS haute performance depuis plus de 30 ans et son utilisation se développe avec les infrastructures basées sur le cloud en raison de sa simplicité.

Le protocole NFS comprend plusieurs versions aux exigences variables. Pour une description complète de la configuration NFS avec ONTAP, reportez-vous à la section "Tr-4067 NFS sur les meilleures pratiques ONTAP". Les sections suivantes couvrent certaines des exigences les plus critiques et des erreurs utilisateur courantes.

Versions NFS

Le client NFS du système d'exploitation doit être pris en charge par NetApp.

- NFSv3 est pris en charge avec des systèmes d'exploitation conformes à la norme NFSv3.

- NFSv3 est pris en charge avec le client Oracle dNFS.
- NFSv4 est pris en charge avec tous les systèmes d'exploitation conformes à la norme NFSv4.
- NFSv4.1 et NFSv4.2 nécessitent une prise en charge spécifique du système d'exploitation. Consulter le "NetApp IMT" Pour les systèmes d'exploitation pris en charge.
- La prise en charge d'Oracle dNFS pour NFSv4.1 requiert Oracle 12.2.0.2 ou version supérieure.

> Le "Matrice de prise en charge de NetApp" Pour NFSv3 et NFSv4 n'incluent pas de systèmes d'exploitation spécifiques. Tous les systèmes d'exploitation conformes à la RFC sont généralement pris en charge. Lors d'une recherche dans la prise en charge en ligne de IMT pour NFSv3 ou NFSv4, ne sélectionnez pas de système d'exploitation spécifique, car aucune correspondance ne sera affichée. Tous les systèmes d'exploitation sont implicitement pris en charge par la politique générale.

Tables d'emplacements TCP Linux NFSv3

Les tables d'emplacements TCP sont l'équivalent NFSv3 de la profondeur de file d'attente de l'adaptateur de bus hôte (HBA). Ces tableaux contrôlent le nombre d'opérations NFS qui peuvent être en attente à la fois. La valeur par défaut est généralement 16, un chiffre bien trop faible pour assurer des performances optimales. Le problème inverse se produit sur les noyaux Linux plus récents : la limite de la table des emplacements TCP augmente automatiquement par envoi de demandes, jusqu'à atteindre le niveau de saturation du serveur NFS.

Pour des performances optimales et pour éviter les problèmes de performances, ajustez les paramètres du noyau qui contrôlent les tables d'emplacements TCP.

Exécutez le `sysctl -a | grep tcp.*.slot_table` et observez les paramètres suivants :

```
# sysctl -a | grep tcp.*.slot_table
sunrpc.tcp_max_slot_table_entries = 128
sunrpc.tcp_slot_table_entries = 128
```

Tous les systèmes Linux doivent inclure `sunrpc.tcp_slot_table_entries`, mais seulement certains incluent `sunrpc.tcp_max_slot_table_entries`. Ils doivent tous deux être réglés sur 128.

> Si vous ne définissez pas ces paramètres, vous risquez d'avoir des effets importants sur les performances. Dans certains cas, les performances sont limitées car le système d'exploitation linux n'émet pas suffisamment d'E/S. Dans d'autres cas, les latences d'E/S augmentent à mesure que le système d'exploitation linux tente d'émettre plus d'E/S que ce qui peut être traité.

ADR et NFS

Certains clients ont signalé des problèmes de performances liés à une quantité excessive d'E/S dans le `ADR` emplacement. Le problème ne se produit généralement pas tant qu'une grande quantité de données de performances ne s'est pas accumulée. La raison de cet excès d'E/S est inconnue, mais ce problème semble provenir des analyses répétées du répertoire cible par les processus Oracle pour détecter les modifications.

Dépose du `noac` et/ou `actimeo=0` Les options de montage permettent la mise en cache du système d'exploitation hôte et réduisent les niveaux d'E/S du stockage.

NetApp recommande de ne pas placer `ADR` données sur un système de fichiers avec `noac` ou `actimeo=0` parce que des problèmes de performances sont probables. Séparer `ADR` le cas échéant, les données vers un autre point de montage.

nfs-rootonly et mount-rootonly

ONTAP inclut une option NFS appelée `nfs-rootonly` Cela permet de contrôler si le serveur accepte les connexions de trafic NFS à partir des ports élevés. Par mesure de sécurité, seul l'utilisateur root est autorisé à ouvrir des connexions TCP/IP à l'aide d'un port source inférieur à 1024 car ces ports sont normalement réservés à l'utilisation du système d'exploitation, et non aux processus utilisateur. Cette restriction permet de s'assurer que le trafic NFS provient d'un client NFS du système d'exploitation et non d'un processus malveillant émulant un client NFS. Le client Oracle dNFS est un pilote d'espace utilisateur, mais le processus s'exécute en tant que root, il n'est donc généralement pas nécessaire de modifier la valeur de `nfs-rootonly`. Les connexions sont réalisées à partir de ports bas.

Le `mount-rootonly` Cette option s'applique uniquement à NFSv3. Il contrôle si l'appel de MONTAGE RPC est accepté à partir de ports supérieurs à 1024. Lorsque dNFS est utilisé, le client est de nouveau exécuté en tant que root, ce qui lui permet d'ouvrir des ports inférieurs à 1024. Ce paramètre n'a aucun effet.

Les processus ouvrant des connexions avec dNFS sur les versions 4.0 et supérieures de NFS ne s'exécutent pas en tant que root et nécessitent donc des ports supérieurs à 1024. Le `nfs-rootonly` Le paramètre doit être défini sur Désactivé pour dNFS pour terminer la connexion.

Si `nfs-rootonly` Est activé, le résultat est un blocage lors de la phase de montage ouvrant les connexions dNFS. La sortie sqlplus ressemble à ceci :

```
SQL>startup
ORACLE instance started.
Total System Global Area 4294963272 bytes
Fixed Size                  8904776 bytes
Variable Size             822083584 bytes
Database Buffers         3456106496 bytes
Redo Buffers                7868416 bytes
```

Le paramètre peut être modifié comme suit :

```
Cluster01::> nfs server modify -nfs-rootonly disabled
```

Dans de rares cas, vous devrez peut-être modifier nfs-rootonly et mount-rootonly sur Désactivé. Si un serveur gère un très grand nombre de connexions TCP, il est possible qu'aucun port inférieur à 1024 n'est disponible et que le système d'exploitation soit forcé d'utiliser des ports supérieurs. Ces deux paramètres ONTAP doivent être modifiés pour permettre la connexion.

Règles d'exportation NFS : superutilisateur et setuid

Si les binaires Oracle se trouvent sur un partage NFS, les règles d'export doivent inclure des autorisations de superutilisateur et de setuid.

Les exportations NFS partagées utilisées pour les services de fichiers génériques tels que les répertoires personnels des utilisateurs écraseront généralement l'utilisateur root. Cela signifie qu'une demande de l'utilisateur root sur un hôte qui a monté un système de fichiers est remappée en tant qu'utilisateur différent avec des privilèges inférieurs. Cela permet de sécuriser les données en empêchant un utilisateur root d'un serveur donné d'accéder aux données du serveur partagé. Le bit setuid peut également représenter un risque de sécurité dans un environnement partagé. Le bit setuid permet d'exécuter un processus en tant qu'utilisateur différent de celui qui appelle la commande. Par exemple, un script shell qui était détenu par root avec le bit setuid s'exécute en tant que root. Si ce script shell peut être modifié par d'autres utilisateurs, tout utilisateur non root peut émettre une commande en tant que root en mettant à jour le script.

Les binaires Oracle incluent les fichiers appartenant à root et utilisent le bit setuid. Si des binaires Oracle sont installés sur un partage NFS, les règles d'export doivent inclure les autorisations de superutilisateur et de setuid appropriées. Dans l'exemple ci-dessous, la règle inclut les deux `allow-suid` et permis `superuser` Accès (root) pour les clients NFS via l'authentification système.

```
Cluster01::> export-policy rule show -vserver vserver1 -policyname orabin
-fields allow-suid,superuser
vserver   policyname ruleindex superuser allow-suid
--------- ---------- --------- --------- ----------
vserver1  orabin     1         sys       true
```

Configuration NFSv4/4.1

Pour la plupart des applications, il y a très peu de différence entre NFSv3 et NFSv4. Les E/S applicatives sont généralement des E/S très simples et ne bénéficient pas énormément de certaines des fonctionnalités avancées de NFSv4. Les versions supérieures de NFS ne doivent pas être considérées comme une « mise à niveau » du point de vue du stockage de la base de données, mais plutôt comme des versions de NFS qui incluent des fonctionnalités supplémentaires. Par exemple, si la sécurité de bout en bout du mode de confidentialité kerberos (krb5p) est requise, NFSv4 est requis.

NetApp recommande d'utiliser NFSv4.1 si les fonctionnalités NFSv4 sont requises. Certaines améliorations fonctionnelles du protocole NFSv4 dans NFSv4.1 améliorent la résilience dans certains cas à la périphérie.

Le passage à NFSv4 est plus compliqué que de simplement changer les options de montage de vers=3 en vers=4.1. Pour une explication plus complète de la configuration de NFSv4 avec ONTAP, notamment des conseils sur la configuration du système d'exploitation, voir "Tr-4067 NFS sur les meilleures pratiques ONTAP". Les sections suivantes de ce TR expliquent certaines des exigences de base relatives à l'utilisation de NFSv4.

Domaine NFSv4

Une explication complète de la configuration NFSv4/4.1 dépasse le cadre de ce document, mais un problème couramment rencontré est une incohérence dans le mappage de domaine. Du point de vue de sysadmin, les systèmes de fichiers NFS semblent se comporter normalement, mais les applications signalent des erreurs concernant les autorisations et/ou le setuid sur certains fichiers. Dans certains cas, les administrateurs ont conclu à tort que les autorisations des binaires de l'application ont été endommagées et ont exécuté des commandes chown ou chmod lorsque le problème réel était le nom de domaine.

Le nom de domaine NFSv4 est défini sur le SVM ONTAP :

```
Cluster01::> nfs server show -fields v4-id-domain
vserver   v4-id-domain
--------- ------------
vserver1  my.lab
```

Le nom de domaine NFSv4 sur l'hôte est défini dans `/etc/idmap.cfg`

```
[root@host1 etc]# head /etc/idmapd.conf
[General]
#Verbosity = 0
# The following should be set to the local NFSv4 domain name
# The default is the host's DNS domain name.
Domain = my.lab
```

Les noms de domaine doivent correspondre. Si ce n'est pas le cas, des erreurs de mappage similaires à ce qui suit apparaissent dans `/var/log/messages`:

```
Apr 12 11:43:08 host1 nfsidmap[16298]: nss_getpwnam: name 'root@my.lab'
does not map into domain 'default.com'
```

Les binaires d'application, tels que les binaires de base de données Oracle, incluent les fichiers appartenant à root avec le bit setuid, ce qui signifie qu'une discordance dans les noms de domaine NFSv4 provoque des échecs avec le démarrage d'Oracle et un avertissement sur la propriété ou les autorisations d'un fichier appelé `oradism`, qui est situé dans le `$ORACLE_HOME/bin` répertoire. Elle doit apparaître comme suit :

```
[root@host1 etc]# ls -l /orabin/product/19.3.0.0/dbhome_1/bin/oradism
-rwsr-x--- 1 root oinstall 147848 Apr 17  2019
/orabin/product/19.3.0.0/dbhome_1/bin/oradism
```

Si ce fichier apparaît avec la propriété de personne, il peut y avoir un problème de mappage de domaine NFSv4.

```
[root@host1 bin]# ls -l oradism
-rwsr-x--- 1 nobody oinstall 147848 Apr 17  2019 oradism
```

Pour résoudre ce problème, vérifiez le `/etc/idmap.cfg` Comparez le paramètre v4-ID-domain sur ONTAP et assurez-vous qu'ils sont cohérents. Si ce n'est pas le cas, effectuez les modifications requises, exécutez `nfsidmap -c`, et attendez un moment pour que les modifications se propagent. La propriété du fichier doit alors être correctement reconnue en tant que racine. Si un utilisateur a tenté de s'exécuter `chown root` Sur ce fichier avant que la configuration des domaines NFS ne soit corrigée, il peut être nécessaire de l'exécuter `chown root` encore.

Oracle Direct NFS (dNFS)

Les bases de données Oracle peuvent utiliser NFS de deux manières.

Tout d'abord, il peut utiliser un système de fichiers monté à l'aide du client NFS natif qui fait partie du système d'exploitation. Il s'agit parfois de kernel NFS ou KNFS. Le système de fichiers NFS est monté et utilisé par la base de données Oracle exactement comme toute autre application utiliserait un système de fichiers NFS.

La deuxième méthode est Oracle Direct NFS (dNFS). Il s'agit d'une implémentation de la norme NFS dans le logiciel de base de données Oracle. Elle ne modifie pas la façon dont les bases de données Oracle sont configurées ou gérées par l'administrateur de base de données. Tant que les paramètres du système de stockage lui-même sont corrects, l'utilisation de dNFS doit être transparente pour l'équipe DBA et les utilisateurs finaux.

Les systèmes de fichiers NFS habituels sont toujours montés sur une base de données avec la fonction dNFS activée. Une fois la base de données ouverte, la base de données Oracle ouvre un ensemble de sessions TCP/IP et effectue directement des opérations NFS.

NFS direct

La valeur principale de Direct NFS d'Oracle est de contourner le client NFS hôte et d'effectuer des opérations de fichiers NFS directement sur un serveur NFS. Pour l'activer, il suffit de modifier la bibliothèque Oracle Disk Manager (ODM). Vous trouverez des instructions sur ce processus dans la documentation Oracle.

L'utilisation de dNFS entraîne une amélioration significative des performances d'E/S et réduit la charge sur l'hôte et le système de stockage, car les E/S sont effectuées de la manière la plus efficace possible.

En outre, Oracle dNFS inclut une **option** pour les chemins d'accès multiples et la tolérance aux pannes de l'interface réseau. Par exemple, il est possible de lier deux interfaces de 10 Gbits pour offrir 20 Go de bande passante. En cas de défaillance d'une interface, les E/S sont relancées sur l'autre interface. L'opération globale est très similaire aux chemins d'accès multiples FC. Les chemins d'accès multiples étaient courants il y a plusieurs années, alors que l'ethernet 1 Gbit était la norme la plus courante. Une carte réseau 10 Go suffit pour la plupart des charges de travail Oracle, mais si un nombre supérieur de cartes réseau 10 Go sont requises, elles peuvent être reliées.

Lorsque dNFS est utilisé, il est essentiel que tous les correctifs décrits dans Oracle Doc 1495104.1 soient installés. Si un correctif ne peut pas être installé, l'environnement doit être évalué pour s'assurer que les bugs décrits dans ce document ne causent pas de problèmes. Dans certains cas, une incapacité à installer les correctifs requis empêche l'utilisation de dNFS.

N'utilisez pas dNFS avec tout type de résolution de noms round-Robin, y compris DNS, DDNS, NIS ou toute autre méthode. Cela inclut la fonction d'équilibrage de la charge DNS disponible dans ONTAP. Lorsqu'une base de données Oracle utilisant dNFS résout un nom d'hôte en adresse IP, elle ne doit pas être modifiée lors des recherches ultérieures. Cela peut entraîner des pannes de la base de données Oracle et une corruption potentielle des données.

Activation de dNFS

Oracle dNFS peut fonctionner avec NFSv3 sans aucune configuration nécessaire au-delà de l'activation de la bibliothèque dNFS (voir la documentation Oracle pour la commande spécifique requise). Toutefois, si dNFS ne parvient pas à établir la connectivité, il peut revenir en arrière silencieux au client NFS du noyau. Dans ce cas, les performances peuvent être gravement affectées.

Si vous souhaitez utiliser le multiplexage dNFS sur plusieurs interfaces, avec NFSv4.X, ou utiliser le chiffrement, vous devez configurer un fichier oranfstab. La syntaxe est extrêmement stricte. De petites erreurs

dans le fichier peuvent entraîner l'affichage du démarrage ou le contournement du fichier orangfstab.

Au moment de la rédaction de ce rapport, les chemins d'accès multiples dNFS ne fonctionnent pas avec NFSv4.1 avec les versions récentes d'Oracle Database. Un fichier oranfstab qui spécifie NFSv4.1 comme protocole ne peut utiliser qu'une instruction de chemin unique pour une exportation donnée. La raison en est que ONTAP ne prend pas en charge l'agrégation ClientID. Les correctifs de bases de données Oracle permettant de résoudre cette limitation seront peut-être disponibles à l'avenir.

La seule façon d'être certain que dNFS fonctionne comme prévu est d'interroger les tables v$dnfs.

Vous trouverez ci-dessous un exemple de fichier oranfstab situé dans /etc Il s'agit de l'un des emplacements multiples où un fichier oranfstab peut être placé.

```
[root@jfs11 trace]# cat /etc/oranfstab
server: NFSv3test
path: jfs_svmdr-nfs1
path: jfs_svmdr-nfs2
export: /dbf mount: /oradata
export: /logs mount: /logs
nfs_version: NFSv3
```

La première étape consiste à vérifier que dNFS est opérationnel pour les systèmes de fichiers spécifiés :

```
SQL> select dirname,nfsversion from v$dnfs_servers;

DIRNAME
-----------------------------------
NFSVERSION
----------------
/logs
NFSv3.0

/dbf
NFSv3.0
```

Ce résultat indique que dNFS est utilisé avec ces deux systèmes de fichiers, mais que **pas** signifie que oranfstab est opérationnel. Si une erreur était présente, dNFS aurait détecté automatiquement les systèmes de fichiers NFS de l'hôte et il se peut que vous voyiez toujours la même sortie à partir de cette commande.

Les chemins d'accès multiples peuvent être vérifiés comme suit :

```
SQL> select svrname,path,ch_id from v$dnfs_channels;

SVRNAME
-----------------------------------
PATH
-----------------------------------
```

```
     CH_ID
----------
NFSv3test
jfs_svmdr-nfs1
         0

NFSv3test
jfs_svmdr-nfs2
         1

SVRNAME
-------------------------------------
PATH
-------------------------------------
     CH_ID
----------

NFSv3test
jfs_svmdr-nfs1
         0

NFSv3test
jfs_svmdr-nfs2

[output truncated]

SVRNAME
-------------------------------------
PATH
-------------------------------------
     CH_ID
----------
NFSv3test
jfs_svmdr-nfs2
         1

NFSv3test
jfs_svmdr-nfs1
         0

SVRNAME
-------------------------------------
PATH
-------------------------------------
     CH_ID
----------
```

```
NFSv3test
jfs_svmdr-nfs2
         1


66 rows selected.
```

Il s'agit des connexions que dNFS utilise. Deux chemins et canaux sont visibles pour chaque entrée SVRNAME. Cela signifie que les chemins d'accès multiples fonctionnent, ce qui signifie que le fichier oranfstab a été reconnu et traité.

Accès direct au NFS et au système de fichiers hôte

L'utilisation de dNFS peut parfois causer des problèmes pour les applications ou les activités des utilisateurs qui dépendent des systèmes de fichiers visibles montés sur l'hôte car le client dNFS accède au système de fichiers hors bande à partir du système d'exploitation hôte. Le client dNFS peut créer, supprimer et modifier des fichiers sans connaître le système d'exploitation.

Lorsque les options de montage des bases de données à instance unique sont utilisées, elles permettent la mise en cache des attributs de fichiers et de répertoires, ce qui signifie également que le contenu d'un répertoire est mis en cache. Par conséquent, dNFS peut créer un fichier, et il y a un court délai avant que le système d'exploitation ne relise le contenu du répertoire et que le fichier devienne visible pour l'utilisateur. Ce n'est généralement pas un problème, mais, dans de rares cas, des utilitaires tels que SAP BR*Tools peuvent présenter des problèmes. Si cela se produit, modifiez les options de montage pour utiliser les recommandations pour Oracle RAC. Ce changement entraîne la désactivation de l'ensemble de la mise en cache de l'hôte.

Ne modifiez les options de montage que si (a) dNFS est utilisé et (b) un problème résulte d'un décalage dans la visibilité des fichiers. Si dNFS n'est pas utilisé, les options de montage Oracle RAC sur une base de données à instance unique entraînent une dégradation des performances.

> Reportez-vous à la remarque à propos de `nosharecache` la "Options de montage NFS Linux" pour un problème dNFS spécifique à Linux qui peut produire des résultats inhabituels.

Locations et verrouillages NFS

NFSv3 est sans état. Cela signifie que le serveur NFS (ONTAP) ne suit pas les systèmes de fichiers montés, par qui ou quels verrous sont réellement en place.

ONTAP dispose de certaines fonctionnalités qui enregistreront les tentatives de montage. Vous savez donc quels clients accèdent aux données et il se peut que des verrous consultatifs soient présents, mais les informations ne sont pas 100 % complètes. Elle ne peut pas être terminée, car le suivi de l'état du client NFS ne fait pas partie de la norme NFSv3.

État NFSv4

En revanche, NFSv4 est avec état. Le serveur NFSv4 suit les clients qui utilisent les systèmes de fichiers, les fichiers existants, les fichiers et/ou les régions de fichiers verrouillés, etc Cela signifie qu'une communication régulière entre un serveur NFSv4 doit être établie pour maintenir les données d'état à jour.

Les États les plus importants gérés par le serveur NFS sont les verrous NFSv4 et les locations NFSv4, qui sont très étroitement liés. Vous devez comprendre comment chacun fonctionne par lui-même, et comment ils

se rapportent les uns aux autres.

Verrous NFSv4

Avec NFSv3, les verrous sont consultatifs. Un client NFS peut toujours modifier ou supprimer un fichier « verrouillé ». Un verrou NFSv3 n'expire pas de lui-même, il doit être supprimé. Cela crée des problèmes. Par exemple, si une application en cluster crée des verrous NFSv3 et que l'un des nœuds tombe en panne, que faire ? Vous pouvez coder l'application sur les nœuds survivants pour supprimer les verrous, mais comment savoir que c'est sûr ? Le nœud « en panne » est peut-être opérationnel, mais ne communique pas avec le reste du cluster ?

Avec NFSv4, les verrous ont une durée limitée. Tant que le client tenant les Locks continue à s'archiver avec le serveur NFSv4, aucun autre client n'est autorisé à acquérir ces Locks. Si un client ne parvient pas à s'archiver avec NFSv4, les verrous seront éventuellement révoqués par le serveur et d'autres clients pourront demander et obtenir des verrous.

Locations NFSv4

Les verrous NFSv4 sont associés à un bail NFSv4. Lorsqu'un client NFSv4 établit une connexion avec un serveur NFSv4, il obtient un bail. Si le client obtient un verrou (il existe plusieurs types de verrous), le verrou est associé au bail.

Ce bail a un délai défini. Par défaut, ONTAP définit la valeur de temporisation sur 30 secondes :

```
Cluster01::*> nfs server show -vserver vserver1 -fields v4-lease-seconds

vserver   v4-lease-seconds
--------- ----------------
vserver1  30
```

Cela signifie qu'un client NFSv4 doit vérifier avec le serveur NFSv4 toutes les 30 secondes pour renouveler ses baux.

Le bail est automatiquement renouvelé par n'importe quelle activité. Ainsi, si le client effectue des travaux, il n'est pas nécessaire d'effectuer des opérations supplémentaires. Si une application devient silencieuse et ne fait pas de véritable travail, elle devra effectuer une sorte d'opération de maintien en vie (appelée SÉQUENCE). Il s'agit essentiellement de dire « Je suis toujours là, veuillez actualiser mes contrats de location ».

```
 *Question:* What happens if you lose network connectivity for 31 seconds?
NFSv3 est sans état. Il ne s'attend pas à ce que les clients communiquent.
NFSv4 est avec état et une fois la période de location expirée, le bail
expire, et les verrous sont révoqués et les fichiers verrouillés sont mis
à disposition des autres clients.
```

Avec NFSv3, vous pouvez déplacer les câbles réseau, redémarrer les switchs réseau, modifier la configuration et être sûr qu'aucun problème ne se produirait. En général, les applications attendront patiemment le bon fonctionnement de la connexion réseau.

Avec NFSv4, vous disposez de 30 secondes (sauf si vous avez augmenté la valeur de ce paramètre dans

ONTAP) pour terminer votre travail. Si vous dépassez cette limite, vos contrats de location sont échus. Normalement, cela provoque des pannes d'application.

Par exemple, si vous disposez d'une base de données Oracle et que vous rencontrez une perte de connectivité réseau (parfois appelée « partition réseau ») qui dépasse le délai d'expiration du bail, vous plantez la base de données.

Voici un exemple de ce qui se passe dans le journal des alertes Oracle si cela se produit :

```
2022-10-11T15:52:55.206231-04:00
Errors in file /orabin/diag/rdbms/ntap/NTAP/trace/NTAP_ckpt_25444.trc:
ORA-00202: control file: '/redo0/NTAP/ctrl/control01.ctl'
ORA-27072: File I/O error
Linux-x86_64 Error: 5: Input/output error
Additional information: 4
Additional information: 1
Additional information: 4294967295
2022-10-11T15:52:59.842508-04:00
Errors in file /orabin/diag/rdbms/ntap/NTAP/trace/NTAP_ckpt_25444.trc:
ORA-00206: error in writing (block 3, # blocks 1) of control file
ORA-00202: control file: '/redo1/NTAP/ctrl/control02.ctl'
ORA-27061: waiting for async I/Os failed
```

Si vous examinez les syslog, vous devriez voir plusieurs de ces erreurs :

```
Oct 11 15:52:55 host1 kernel: NFS: nfs4_reclaim_open_state: Lock reclaim
failed!
Oct 11 15:52:55 host1 kernel: NFS: nfs4_reclaim_open_state: Lock reclaim
failed!
Oct 11 15:52:55 host1 kernel: NFS: nfs4_reclaim_open_state: Lock reclaim
failed!
```

Les messages du journal sont généralement le premier signe d'un problème, autre que le blocage de l'application. En général, vous ne voyez rien pendant la panne réseau, car les processus et le système d'exploitation lui-même sont bloqués et tentent d'accéder au système de fichiers NFS.

Les erreurs apparaissent une fois que le réseau est de nouveau opérationnel. Dans l'exemple ci-dessus, une fois la connectivité rétablie, le système d'exploitation a tenté de réacquérir les verrous, mais il était trop tard. Le bail avait expiré et les serrures ont été retirées. Cela entraîne une erreur qui se propage jusqu'à la couche Oracle et provoque le message dans le journal des alertes. Vous pouvez voir des variations sur ces modèles en fonction de la version et de la configuration de la base de données.

En résumé, NFSv3 tolère l'interruption du réseau, mais NFSv4 est plus sensible et impose une période de location définie.

Que se passe-t-il si un délai de 30 secondes n'est pas acceptable ? Que se passe-t-il si vous gérez un réseau changeant de façon dynamique où les commutateurs sont redémarrés ou les câbles sont déplacés et que le résultat est une interruption occasionnelle du réseau ? Vous pouvez choisir de prolonger la période de

location, mais pour savoir si vous voulez y parvenir, vous devez expliquer les périodes de grâce NFSv4.

Périodes de grâce NFSv4

Lorsqu'un serveur NFSv3 est redémarré, il est prêt à transmettre les E/S presque instantanément. Il ne maintenait aucune sorte d'état concernant les clients. Le résultat est qu'une opération de basculement ONTAP semble souvent proche de l'instantané. Dès qu'un contrôleur est prêt à commencer à transmettre des données, il envoie un ARP au réseau qui signale le changement de topologie. En règle générale, les clients le détectent presque instantanément et le flux des données reprend.

NFSv4, cependant, fera une courte pause. Cela fait partie du fonctionnement de NFSv4.

> Les sections suivantes sont à jour depuis ONTAP 9.15.1, mais le comportement de bail et de verrouillage ainsi que les options de réglage peuvent changer de version à version. Si vous avez besoin d'ajuster les délais de location/verrouillage de NFSv4, veuillez consulter le support NetApp pour obtenir les informations les plus récentes.

Les serveurs NFSv4 doivent suivre les baux, les verrous et les utilisateurs des données. Si un serveur NFS fonctionne de manière incohérente et redémarre, ou perd de l'alimentation pendant un moment, ou est redémarré pendant l'activité de maintenance, le résultat est le bail/verrouillage et d'autres informations client sont perdues. Le serveur doit déterminer quel client utilise les données avant de reprendre les opérations. C'est là que intervient le délai de grâce.

Si vous mettez soudainement votre serveur NFSv4 hors/sous tension. Lorsqu'il est rétabli, les clients qui tentent de reprendre l'E/S reçoivent une réponse qui dit essentiellement « J'ai perdu les informations de location/verrouillage. Voulez-vous réenregistrer vos verrous ? » C'est le début de la période de grâce. La valeur par défaut est 45 secondes sur ONTAP :

```
Cluster01::> nfs server show -vserver vserver1 -fields v4-grace-seconds

vserver   v4-grace-seconds
--------- ----------------
vserver1  45
```

Par conséquent, après un redémarrage, un contrôleur met en pause les E/S tandis que tous les clients récupèrent leurs baux et verrous. Une fois le délai de grâce terminé, le serveur reprend les opérations d'E/S.

Cette période de grâce contrôle la récupération de bail pendant les modifications de l'interface réseau, mais il existe une deuxième période de grâce qui contrôle la récupération pendant le basculement du stockage `locking.grace_lease_seconds`. Il s'agit d'une option au niveau du nœud.

```
cluster01::> node run [node names or *] options
locking.grace_lease_seconds
```

Par exemple, si vous avez fréquemment besoin d'effectuer des basculements LIF, et que vous devez réduire le délai de grâce, vous changiez . `v4-grace-seconds` Si vous souhaitez améliorer le temps de reprise des E/S pendant le basculement du contrôleur, vous devez le modifier `locking.grace_lease_seconds`.

Ne modifiez ces valeurs qu'avec prudence et après avoir parfaitement compris les risques et les conséquences. Les pauses E/S liées aux opérations de basculement et de migration avec NFSv4.X ne

peuvent pas être entièrement évitées. Les périodes de verrouillage, de bail et de grâce font partie de la RFC NFS. Pour de nombreux clients, NFSv3 est préférable, car les délais de basculement sont plus courts.

Délais de location par rapport aux délais de grâce

Le délai de grâce et la période de location sont connectés. Comme mentionné ci-dessus, le délai de bail par défaut est de 30 secondes, ce qui signifie que les clients NFSv4 doivent s'enregistrer auprès du serveur au moins toutes les 30 secondes, sinon ils perdent leur bail et, à leur tour, leurs verrous. Le délai de grâce existe pour permettre à un serveur NFS de reconstruire les données de bail/verrouillage, et il prend par défaut 45 secondes. Le délai de grâce doit être plus long que la période de location. Cela permet de s'assurer qu'un environnement client NFS conçu pour renouveler les contrats de location au moins toutes les 30 secondes aura la possibilité d'archiver avec le serveur après un redémarrage. Un délai de grâce de 45 secondes garantit que tous les clients qui s'attendent à renouveler leur contrat de location au moins toutes les 30 secondes ont certainement l'occasion de le faire.

Si un délai de 30 secondes n'est pas acceptable, vous pouvez choisir de prolonger la période de location.

Si vous souhaitez augmenter le délai de bail à 60 secondes pour résister à une panne réseau de 60 secondes, vous devrez également augmenter le délai de grâce. Une pause d'E/S plus longue sera donc nécessaire lors du basculement du contrôleur.

Ce ne devrait normalement pas être un problème. En général, les utilisateurs ne mettent à jour les contrôleurs ONTAP qu'une ou deux fois par an. En outre, les basculements non planifiés en raison de défaillances matérielles sont extrêmement rares. En outre, si vous aviez un réseau où une panne réseau de 60 secondes était possible, et que le délai de bail était de 60 secondes, vous n'auriez probablement pas à vous opposer à un basculement rare du système de stockage, ce qui aurait entraîné une pause de 61 secondes non plus. Vous avez déjà reconnu que vous disposez d'un réseau qui s'arrête pendant plus de 60 secondes plutôt fréquemment.

Mise en cache NFS

La présence de l'une des options de montage suivantes entraîne la désactivation de la mise en cache de l'hôte :

```
cio, actimeo=0, noac, forcedirectio
```

Ces paramètres peuvent avoir un effet négatif important sur la vitesse d'installation du logiciel, de correction et des opérations de sauvegarde/restauration. Dans certains cas, en particulier avec les applications en cluster, ces options sont obligatoires car elles doivent inévitablement assurer la cohérence du cache sur tous les nœuds du cluster. Dans d'autres cas, les clients utilisent ces paramètres par erreur, ce qui entraîne des dommages inutiles aux performances.

De nombreux clients suppriment temporairement ces options de montage lors de l'installation ou de l'application de correctifs binaires. Cette suppression peut être effectuée en toute sécurité si l'utilisateur vérifie qu'aucun autre processus n'utilise activement le répertoire cible pendant le processus d'installation ou de correction.

Tailles de transfert NFS

Par défaut, ONTAP limite la taille des E/S NFS à 64 Ko.

Les E/S aléatoires utilisent la plupart des applications et bases de données une taille de bloc bien inférieure à la taille maximale de 64 Ko. Les E/S de blocs volumineux sont généralement parallélisées de sorte que le

maximum de 64 Ko ne limite pas non plus l'obtention d'une bande passante maximale.

Dans certains cas, le maximum de 64 000 charges de travail entraîne une limitation. En particulier, les opérations à thread unique, telles que les opérations de sauvegarde ou de restauration, ou encore les analyses de table complète de base de données s'exécutent plus rapidement et plus efficacement si la base de données peut exécuter moins d'E/S, mais plus volumineuses. La taille optimale de gestion des E/S pour ONTAP est de 256 Ko.

La taille maximale de transfert pour un SVM ONTAP donné peut être modifiée comme suit :

```
Cluster01::> set advanced
Warning: These advanced commands are potentially dangerous; use them only
when directed to do so by NetApp personnel.
Do you want to continue? {y|n}: y
Cluster01::*> nfs server modify -vserver vserver1 -tcp-max-xfer-size
262144
Cluster01::*>
```

> Ne réduisez jamais la taille de transfert maximale autorisée sur ONTAP en dessous de la valeur de rsize/wsize des systèmes de fichiers NFS actuellement montés. Cela peut provoquer des blocages ou même une corruption des données avec certains systèmes d'exploitation. Par exemple, si les clients NFS sont actuellement définis sur une taille rsize/wsize de 65536, la taille maximale du transfert ONTAP peut être ajustée entre 65536 et 1048576 sans effet car les clients eux-mêmes sont limités. Réduire la taille de transfert maximale en dessous de 65536 peut endommager la disponibilité ou les données.

NVFAIL

NVFAIL est une fonctionnalité de ONTAP qui assure l'intégrité lors des scénarios de basculement catastrophiques.

En raison de la gestion de caches internes volumineux, les bases de données sont vulnérables à la corruption lors des événements de basculement du stockage. Si un événement catastrophique nécessite de forcer un basculement ONTAP ou de forcer le basculement MetroCluster, quel que soit l'état de santé de la configuration globale, les modifications qui ont été reconnues précédemment peuvent être supprimées. Le contenu de la matrice 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. Cette incohérence entraîne une corruption des données.

La mise en cache peut avoir lieu au niveau des applications ou des serveurs. Par exemple, une configuration Oracle Real application Cluster (RAC) avec des serveurs actifs sur un site principal et un site distant met en cache les données dans la SGA d'Oracle. Une opération de basculement forcé entraînant des pertes de données risque de corrompre la base de données, car les blocs stockés dans la mémoire SGA peuvent ne pas correspondre aux blocs du disque.

L'utilisation de la mise en cache est moins évidente au niveau du système de fichiers du système d'exploitation. Les blocs d'un système de fichiers NFS monté peuvent être mis en cache dans le système d'exploitation. Un système de fichiers en cluster basé sur des LUN situés sur le site principal peut également être monté sur des serveurs du site distant, et une fois encore, les données peuvent être mises en cache. Une défaillance de la mémoire NVRAM, un basculement forcé ou un basculement forcé dans ces situations peuvent entraîner une corruption du système de fichiers.

ONTAP protège les bases de données et les systèmes d'exploitation de ce scénario avec NVFAIL et ses paramètres associés.

Utilitaire de récupération ASM (ASMRU)

ONTAP supprime efficacement les blocs nuls écrits sur un fichier ou une LUN lorsque la compression à la volée est activée. Des utilitaires tels que l'utilitaire ASRU (Oracle ASM Reclamation Utility) sont utilisés en écrivant des zéros dans les extensions ASM inutilisées.

Cela permet aux administrateurs de bases de données de récupérer de l'espace sur la baie de stockage après la suppression des données. ONTAP intercepte les zéros et désalloue l'espace de la LUN. Le processus de récupération est extrêmement rapide, car aucune donnée n'est écrite dans le système de stockage.

Du point de vue de la base de données, le groupe de disques ASM contient des zéros et la lecture de ces régions des LUN entraîne un flux de zéros, mais ONTAP ne stocke pas les zéros sur les disques. Des modifications simples des métadonnées sont effectuées en interne pour marquer les régions mises à zéro de la LUN comme vides de toutes les données.

Pour des raisons similaires, le test de performance impliquant des données mises à zéro n'est pas valide, car les blocs de zéros ne sont pas réellement traités comme des écritures dans la baie de stockage.

Lorsque vous utilisez ASRU, assurez-vous que tous les correctifs recommandés par Oracle sont installés.

Configuration du stockage sur les systèmes ASA r2

SAN FC

Alignement de LUN

L'alignement des LUN fait référence à l'optimisation des E/S par rapport à la disposition du système de fichiers sous-jacent.

Les systèmes ASA r2 utilisent la même architecture ONTAP que AFF/ FAS mais avec un modèle de configuration simplifié. Les systèmes ASA r2 utilisent des zones de disponibilité de stockage (SAZ) au lieu d'agrégats, mais les principes d'alignement restent les mêmes car ONTAP gère la disposition des blocs de manière cohérente sur toutes les plateformes. Veuillez toutefois noter les points spécifiques à ASA suivants :

- Les systèmes ASA r2 fournissent des chemins symétriques actifs-actifs pour tous les LUN, ce qui élimine les problèmes d'asymétrie de chemin lors de l'alignement.
- Les unités de stockage (LUN) sont provisionnées en mode fin par défaut ; l'alignement ne change pas ce comportement.
- La réservation de snapshots et la suppression automatique des snapshots peuvent être configurées lors de la création du LUN (ONTAP 9.18.1 et versions ultérieures).

Sur un système ONTAP, le stockage est organisé en unités de 4 Ko. Un bloc de 8 Ko de base de données ou de système de fichiers doit être mappé à exactement deux blocs de 4 Ko. Si une erreur de configuration de LUN déplace l'alignement de 1 Ko dans les deux sens, chaque bloc de 8 Ko existerait sur trois blocs de stockage de 4 Ko différents au lieu de deux. Cette configuration entraîne une augmentation de la latence et des E/S supplémentaires au sein du système de stockage.

L'alignement affecte également les architectures LVM. Si un volume physique au sein d'un groupe de volumes logiques est défini sur l'unité entière (aucune partition n'est créée), le premier bloc de 4 Ko de la LUN s'aligne sur le premier bloc de 4 Ko du système de stockage. Il s'agit d'un alignement correct. Des problèmes surviennent avec les partitions car elles déplacent l'emplacement de départ où le système d'exploitation utilise le LUN. Tant que le décalage est décalé en unités entières de 4 Ko, la LUN est alignée.

Dans les environnements Linux, créez des groupes de volumes logiques sur l'ensemble de l'unité de disque. Lorsqu'une partition est requise, vérifiez l'alignement en exécutant et en `fdisk -u` vérifiant que le début de chaque partition est un multiple de huit. Cela signifie que la partition démarre à un multiple de huit secteurs de 512 octets, soit 4 Ko.

Voir également la discussion sur l'alignement des blocs de compression dans la section "Efficacité". Toute disposition alignée avec les limites des blocs de compression de 8 Ko est également alignée avec les limites de 4 Ko.

Avertissements de mauvais alignement

La journalisation des opérations de reprise et des transactions de la base de données génère normalement des E/S non alignées qui peuvent entraîner des avertissements erronés concernant les LUN mal alignées sur ONTAP.

La journalisation effectue une écriture séquentielle du fichier journal avec des écritures de taille variable. Une opération d'écriture de journal qui ne s'aligne pas sur les limites de 4 Ko ne provoque généralement pas de problèmes de performances, car l'opération d'écriture de journal suivante termine le bloc. ONTAP est ainsi en mesure de traiter la quasi-totalité des écritures sous forme de blocs complets de 4 Ko, même si les données de blocs de 4 Ko ont été écrites dans deux opérations distinctes.

Vérifiez l'alignement à l'aide d'utilitaires tels que `sio` ou `dd` qui peut générer des E/S à une taille de bloc définie. Les statistiques d'alignement des E/S sur le système de stockage peuvent être consultées avec le `stats` commande. Voir "Vérification de l'alignement WAFL" pour plus d'informations.

L'alignement dans les environnements Solaris est plus compliqué. Reportez-vous à la section "Configuration de l'hôte SAN ONTAP" pour en savoir plus.

> Dans les environnements Solaris x86, prenez davantage soin de l'alignement approprié car la plupart des configurations comportent plusieurs couches de partitions. Les tranches de partition Solaris x86 existent généralement au-dessus d'une table de partition d'enregistrement d'amorçage maître standard.

Autres bonnes pratiques :

- Vérifiez les paramètres du micrologiciel HBA et du système d'exploitation par rapport à l'outil de matrice d'interopérabilité NetApp (IMT).
- Utilisez les utilitaires sanlun pour vérifier l'état et l'alignement du chemin.
- Pour Oracle ASM et LVM, assurez-vous que les fichiers de configuration (/etc/lvm/lvm.conf, /etc/sysconfig/oracleasm) sont correctement configurés afin d'éviter les problèmes d'alignement.

Dimensionnement des LUN et nombre de LUN

Il est essentiel de sélectionner la taille de LUN optimale et le nombre de LUN à utiliser pour optimiser les performances et la gestion des bases de données Oracle.

Un LUN est un objet virtualisé sur ONTAP qui existe sur tous les disques de la zone de disponibilité de

stockage (SAZ) hôte sur les systèmes ASA r2. Par conséquent, les performances du LUN ne sont pas affectées par sa taille car le LUN exploite pleinement le potentiel de performance du SAZ, quelle que soit la taille choisie.

À titre de commodité, les clients peuvent souhaiter utiliser un LUN de taille spécifique. Par exemple, si une base de données est construite sur un groupe de disques LVM ou Oracle ASM composé de deux LUN de 1 To chacune, ce groupe de disques doit être développé par incréments de 1 To. Il peut être préférable de créer le groupe de disques à partir de huit LUN de 500 Go chacune, de sorte que le groupe de disques puisse être augmenté par incréments plus petits.

Il n'est pas recommandé d'établir une taille de LUN standard universelle, car cela peut compliquer la gestion. Par exemple, une taille de LUN standard de 100 Go peut fonctionner correctement lorsqu'une base de données ou un datastore se situe entre 1 et 2 To, mais qu'une base de données ou un datastore de 20 To nécessite 200 LUN. Cela signifie que les délais de redémarrage du serveur sont plus longs, que les différents utilisateurs doivent gérer davantage d'objets et que des produits tels que SnapCenter doivent effectuer des recherches sur de nombreux objets. L'utilisation d'un nombre inférieur de LUN de plus grande taille permet d'éviter de tels problèmes.

- Considérations relatives à ASA r2 :*
- La taille maximale d'un LUN pour ASA r2 est de 128 To, ce qui permet d'utiliser moins de LUN, mais plus volumineux, sans impact sur les performances.
- ASA r2 utilise des zones de disponibilité de stockage (SAZ) au lieu d'agrégats, mais cela ne modifie pas la logique de dimensionnement des LUN pour les charges de travail Oracle.
- Le provisionnement fin est activé par défaut ; le redimensionnement des LUN est non perturbateur et ne nécessite pas leur mise hors ligne.

Nombre de LUN

Contrairement à la taille de LUN, le nombre de LUN affecte les performances. La performance des applications dépend souvent de la capacité à réaliser des E/S parallèles via la couche SCSI. Ainsi, deux LUN offrent de meilleures performances qu'une seule LUN. L'utilisation d'un LVM tel que Veritas VxVM, Linux LVM2 ou Oracle ASM est la méthode la plus simple pour augmenter le parallélisme.

Avec ASA r2, les principes de comptage des LUN restent les mêmes qu'avec AFF/ FAS car ONTAP gère les E/S parallèles de manière similaire sur toutes les plateformes. Cependant, l'architecture exclusivement SAN et les chemins symétriques actifs-actifs de ASA r2 garantissent des performances constantes sur tous les LUN.

Les clients NetApp n'ont généralement pas eu l'avantage d'augmenter le nombre de LUN au-delà de seize. Toutefois, le test d'environnements 100 % SSD avec des E/S aléatoires très lourdes a permis d'améliorer encore jusqu'à 64 LUN.

NetApp recommande ce qui suit :

En général, quatre à seize LUN suffisent pour prendre en charge les besoins d'E/S de toute charge de travail de base de données Oracle donnée. L'utilisation de moins de quatre LUN peut entraîner des limitations de performances en raison des limitations des implémentations SCSI de l'hôte. Augmenter le nombre de LUN au-delà de seize améliore rarement les performances, sauf dans des cas extrêmes (comme des charges de travail SSD d'E/S aléatoires très élevées).

Placement des LUN

Le placement optimal des LUN de base de données au sein des systèmes ASA r2 dépend principalement de la manière dont les différentes fonctionnalités ONTAP seront

utilisées.

Dans les systèmes ASA r2, les unités de stockage (LUN ou espaces de noms NVMe) sont créées à partir d'une couche de stockage simplifiée appelée zones de disponibilité de stockage (SAZ), qui servent de pools de stockage communs pour une paire HA.

Il n'y a généralement qu'une seule zone de disponibilité de stockage (SAZ) par paire HA.

Zones de disponibilité de stockage (ZDS)

Dans les systèmes ASA r2, les volumes sont toujours présents, mais ils sont créés automatiquement lors de la création des unités de stockage. Les unités de stockage (LUN ou espaces de noms NVMe) sont provisionnées directement dans les volumes créés automatiquement dans les zones de disponibilité de stockage (SAZ). Cette conception élimine le besoin de gestion manuelle des volumes et rend le provisionnement plus direct et rationalisé pour les charges de travail par blocs telles que les bases de données Oracle.

SAZ et unités de stockage

Les unités de stockage associées (LUN ou espaces de noms NVMe) sont normalement colocalisées dans une seule zone de disponibilité de stockage (SAZ). Par exemple, une base de données qui nécessite 10 unités de stockage (LUN) aurait généralement toutes les 10 unités placées dans la même SAZ pour des raisons de simplicité et de performance.

- L'utilisation d'un ratio de 1:1 entre les unités de stockage et les volumes, c'est-à-dire une unité de stockage (LUN) par volume, est le comportement par défaut de ASA r2.
- En cas de présence de plusieurs paires HA dans le système ASA r2, les unités de stockage (LUN) d'une base de données donnée peuvent être réparties sur plusieurs SAZ afin d'optimiser l'utilisation et les performances du contrôleur.

Dans le contexte d'un SAN FC, l'unité de stockage fait ici référence à un LUN.

Groupes de cohérence (CG), LUN et instantanés

Dans ASA r2, les politiques et les planifications de snapshots sont appliquées au niveau du groupe de cohérence, qui est une construction logique regroupant plusieurs LUN ou espaces de noms NVMe pour une protection coordonnée des données. Un ensemble de données composé de 10 LUN ne nécessiterait qu'une seule stratégie de snapshot si ces LUN font partie du même groupe de cohérence.

Les groupes de cohérence garantissent des opérations d'instantané atomiques sur tous les LUN inclus. Par exemple, une base de données résidant sur 10 LUN, ou un environnement d'application basé sur VMware composé de 10 systèmes d'exploitation différents, peut être protégé comme un seul objet cohérent si les LUN sous-jacents sont regroupés dans le même groupe de cohérence. Si elles sont placées dans des groupes de cohérence différents, les instantanés peuvent ne pas être parfaitement synchronisés, même s'ils sont planifiés en même temps.

Dans certains cas, un ensemble de LUN apparentées peut devoir être divisé en deux groupes de cohérence différents en raison des exigences de récupération. Par exemple, une base de données peut comporter quatre LUN pour les fichiers de données et deux LUN pour les journaux. Dans ce cas, un groupe de cohérence de fichiers de données avec 4 LUN et un groupe de cohérence de journal avec 2 LUN pourraient être la meilleure option. La raison est la récupération indépendante : le groupe de cohérence des fichiers de données pourrait être restauré sélectivement à un état antérieur, ce qui signifie que les quatre LUN seraient ramenés à l'état de l'instantané, tandis que le groupe de cohérence des journaux avec ses données critiques resterait intact.

CG, LUN et SnapMirror

Les politiques et opérations SnapMirror , tout comme les opérations de snapshot, sont exécutées sur le groupe de cohérence et non sur le LUN.

Le regroupement des LUN apparentées dans un seul groupe de cohérence vous permet de créer une seule relation SnapMirror et de mettre à jour toutes les données contenues en une seule mise à jour. Comme pour les instantanés, la mise à jour sera également une opération atomique. La destination SnapMirror garantirait la présence d'une réplique unique à un instant donné des LUN sources. Si les LUN étaient répartis sur plusieurs groupes de cohérence, les répliques pourraient être cohérentes ou non entre elles.

La réplication SnapMirror sur les systèmes ASA r2 présente les limitations suivantes :

- La réplication synchrone SnapMirror n'est pas prise en charge.
- La synchronisation active SnapMirror est prise en charge uniquement entre deux systèmes ASA r2.
- La réplication asynchrone SnapMirror est prise en charge uniquement entre deux systèmes ASA r2.
- La réplication asynchrone SnapMirror n'est pas prise en charge entre un système ASA r2 et un système ASA, AFF ou FAS ou le cloud.

Apprenez-en davantage sur "Les politiques de réplication SnapMirror sont prises en charge sur les systèmes ASA r2.".

CG, LUN et QoS

Bien que la QoS puisse être appliquée sélectivement à des LUN individuels, il est généralement plus facile de la configurer au niveau du groupe de cohérence. Par exemple, tous les LUN utilisés par les invités dans un serveur ESX donné pourraient être placés dans un seul groupe de cohérence, puis une politique QoS adaptative ONTAP pourrait être appliquée. Il en résulte une limite d'IOPS par Tio à auto-ajustement qui s'applique à tous les LUN.

De même, si une base de données nécessitait 100 000 IOPS et occupait 10 LUN, il serait plus facile de définir une seule limite de 100 000 IOPS sur un seul groupe de cohérence que de définir 10 limites individuelles de 10 000 IOPS, une sur chaque LUN.

Plusieurs mises en page CG

Il existe certains cas où la répartition des LUN sur plusieurs groupes de cohérence peut s'avérer bénéfique. La raison principale est le découpage des contrôleurs. Par exemple, un système de stockage HA ASA r2 peut héberger une seule base de données Oracle nécessitant la pleine capacité de traitement et de mise en cache de chaque contrôleur. Dans ce cas, une conception typique consisterait à placer la moitié des LUN dans un seul groupe de cohérence sur le contrôleur 1, et l'autre moitié des LUN dans un seul groupe de cohérence sur le contrôleur 2.

De même, pour les environnements hébergeant de nombreuses bases de données, la répartition des LUN sur plusieurs groupes de cohérence peut garantir une utilisation équilibrée du contrôleur. Par exemple, un système HA hébergeant 100 bases de données de 10 LUN chacune pourrait attribuer 5 LUN à un groupe de cohérence sur le contrôleur 1 et 5 LUN à un groupe de cohérence sur le contrôleur 2 par base de données. Cela garantit une charge symétrique lors de la mise en service de bases de données supplémentaires.

Aucun de ces exemples n'implique cependant un ratio LUN/groupe de consistance de 1:1. L'objectif reste d'optimiser la gestion en regroupant logiquement les LUN apparentées dans un groupe de cohérence.

Un exemple où un ratio LUN/groupe de cohérence de 1:1 est pertinent est celui des charges de travail conteneurisées, où chaque LUN peut en réalité représenter une seule charge de travail nécessitant des politiques de snapshot et de réplication distinctes et doit donc être gérée individuellement. Dans de tels cas, un ratio de 1:1 peut être optimal.

Redimensionnement des LUN et LVM

Lorsqu'un système de fichiers SAN ou un groupe de disques Oracle ASM atteint sa limite de capacité sur ASA r2, deux options permettent d'augmenter l'espace disponible :

- Augmenter la taille des LUN (unités de stockage) existantes
- Ajoutez un nouveau LUN à un groupe de disques ASM ou à un groupe de volumes LVM existant et étendez le volume logique qu'il contient.

Bien que le redimensionnement des LUN soit pris en charge sur ASA r2, il est généralement préférable d'utiliser un gestionnaire de volumes logiques (LVM) tel qu'Oracle ASM. L'une des principales raisons d'être des LVM est d'éviter la nécessité de redimensionner fréquemment les LUN. Avec LVM, plusieurs LUN sont combinés en un pool de stockage virtuel. Les volumes logiques extraits de ce pool peuvent être facilement redimensionnés sans impacter la configuration de stockage sous-jacente.

L'utilisation de LVM ou d'ASM présente également les avantages suivants :

- Optimisation des performances : répartit les E/S sur plusieurs LUN, réduisant ainsi les points chauds.
- Flexibilité : Ajoutez de nouveaux LUN sans perturber les charges de travail existantes.
- Migration transparente : ASM ou LVM peuvent déplacer des étendues vers de nouveaux LUN à des fins d'équilibrage ou de hiérarchisation sans interruption de service de l'hôte.

Principaux points à prendre en compte concernant ASA r2 :

- Le redimensionnement LUN est effectué au niveau de l'unité de stockage au sein d'une machine virtuelle de stockage (SVM) en utilisant la capacité de la zone de disponibilité de stockage (SAZ).
- Pour Oracle, la meilleure pratique consiste à ajouter des LUN aux groupes de disques ASM plutôt qu'à redimensionner les LUN existants, afin de maintenir le striping et le parallélisme.

Répartition LVM

La répartition des LVM consiste à distribuer les données entre plusieurs LUN. Les performances de nombreuses bases de données en sont ainsi considérablement améliorées.

Avant l'ère des disques Flash, la répartition était utilisée pour surmonter les limites de performances des disques rotatifs. Par exemple, si un système d'exploitation doit effectuer une opération de lecture de 1 Mo, la lecture de ce 1 Mo de données à partir d'un seul disque demande beaucoup de tête de lecture lorsque le transfert des 1 Mo est lent. Si ce 1 Mo de données a été réparti sur 8 LUN, le système d'exploitation pourrait exécuter huit opérations de lecture de 128 K en parallèle et réduire le temps nécessaire au transfert de 1 Mo.

Le découpage avec des disques durs rotatifs était plus difficile car le schéma d'E/S devait être connu à l'avance. Si le découpage en bandes n'était pas correctement paramétré pour les véritables configurations d'E/S, les configurations découpées en bandes pouvaient nuire aux performances. Avec les bases de données Oracle, et en particulier avec les configurations de stockage tout flash, le striping est beaucoup plus facile à

configurer et il a été prouvé qu'il améliore considérablement les performances.

Par défaut, les gestionnaires de volumes logiques, tels que la bande Oracle ASM, ne le font pas pour le système d'exploitation natif LVM. Certaines lient plusieurs LUN ensemble en tant que périphérique concaténé. Résultat : des fichiers de données existent sur un seul périphérique LUN. Ceci provoque des points chauds. Les autres implémentations LVM prennent par défaut en charge les extensions distribuées. Cette méthode est similaire à la répartition, mais elle est plus grossière. Les LUN du groupe de volumes sont tranchées en grandes parties, appelées extensions et généralement mesurées en plusieurs mégaoctets. Ensuite, les volumes logiques sont distribués sur ces extensions. Il en résulte des E/S aléatoires sur un fichier qui doit être bien réparti entre les LUN, mais les opérations d'E/S séquentielles ne sont pas aussi efficaces qu'elles pourraient l'être.

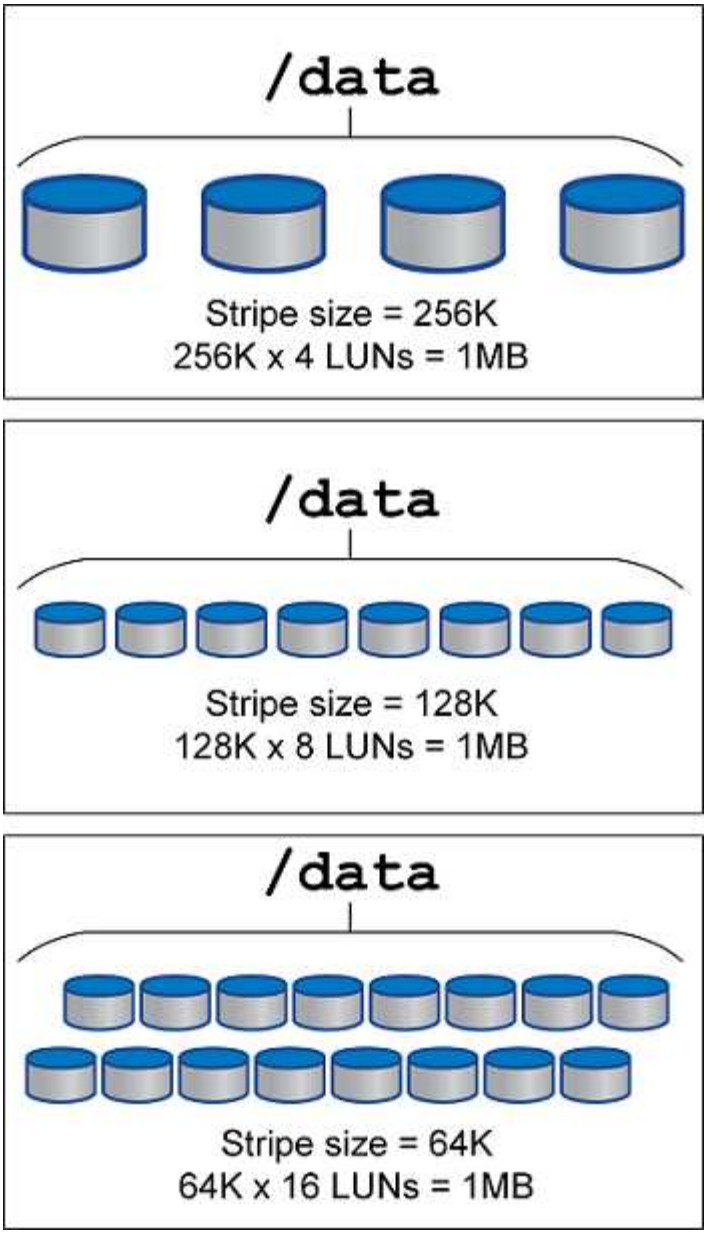
Les E/S des applications exigeantes en performances sont presque toujours de (a) en unités de taille de bloc de base ou (b) d'un mégaoctet.

L'objectif principal d'une configuration à bandes est de s'assurer que les E/S de fichier unique peuvent être exécutées comme une seule unité, et que les E/S de plusieurs blocs, d'une taille de 1 Mo, peuvent être parallélisées de façon homogène sur toutes les LUN du volume réparti. Cela signifie que la taille de bande ne doit pas être inférieure à la taille du bloc de base de données, et que la taille de bande multipliée par le nombre de LUN doit être de 1 Mo.

Meilleures pratiques pour le striping LVM avec une base de données Oracle :

- Taille de la bande ≥ taille du bloc de base de données.
- Taille de la bande * nombre de LUN ≈ 1 Mo pour un parallélisme optimal.
- Utilisez plusieurs LUN par groupe de disques ASM pour maximiser le débit et éviter les points chauds.

La figure suivante présente trois options possibles pour le réglage de la taille et de la largeur des bandes. Le nombre de LUN est sélectionné pour répondre aux exigences de performances comme décrit ci-dessus, mais dans tous les cas, le total des données dans une seule bande est de 1 Mo.



NVFAIL

NVFAIL est une fonctionnalité ONTAP qui garantit l'intégrité des données lors de scénarios de basculement catastrophiques.

Cette fonctionnalité reste applicable sur les systèmes ASA r2, même si ASA r2 utilise une architecture SAN simplifiée (SAZ et unités de stockage au lieu de volumes).

Les bases de données sont vulnérables à la corruption lors des basculements de stockage car elles conservent d'importants caches internes. Si un événement catastrophique nécessite de forcer un basculement ONTAP , indépendamment de l'état de la configuration globale, il en résulte que des modifications précédemment reconnues peuvent être effectivement abandonnées. Le contenu de la baie de stockage remonte dans le temps, et l'état du cache de la base de données ne reflète plus l'état des données sur le disque. Cette incohérence entraîne une corruption des données.

La mise en cache peut avoir lieu au niveau de l'application ou du serveur. Par exemple, une configuration Oracle Real Application Cluster (RAC) avec des serveurs actifs à la fois sur un site principal et un site distant

met en cache les données dans l'Oracle SGA. Une opération de basculement forcé entraînant une perte de données exposerait la base de données à un risque de corruption, car les blocs stockés dans la SGA pourraient ne pas correspondre aux blocs sur le disque.

Une utilisation moins évidente de la mise en cache se situe au niveau du système de fichiers du système d'exploitation. Un système de fichiers en cluster basé sur des LUN situés sur le site principal pourrait être monté sur des serveurs du site distant, et les données pourraient à nouveau être mises en cache. Une défaillance de la NVRAM ou une prise de contrôle forcée dans ces situations pourrait entraîner une corruption du système de fichiers.

ONTAP protège les bases de données et les systèmes d'exploitation contre ce scénario grâce à NVFAIL et à ses paramètres associés, qui signalent à l'hôte d'invalider les données mises en cache et de remonter les systèmes de fichiers affectés après le basculement. Ce mécanisme s'applique aux LUN et aux espaces de noms ASA r2 de la même manière qu'à AFF/ FAS.

Principaux points à prendre en compte concernant ASA r2 :

- NVFAIL fonctionne au niveau LUN (unité de stockage), et non au niveau SAZ.
- Pour les bases de données Oracle, NVFAIL doit être activé sur tous les LUN hébergeant des composants critiques (fichiers de données, journaux de restauration, fichiers de contrôle).
- MetroCluster n'est pas pris en charge sur ASA r2, NVFAIL s'applique donc principalement aux scénarios de basculement HA locaux.
- NFS n'est pas pris en charge sur ASA r2, les considérations NVFAIL ne s'appliquent donc qu'aux charges de travail basées sur SAN (FC/iSCSI/NVMe).

Utilitaire de récupération ASM (ASRU)

ONTAP sur ASA r2 supprime efficacement les blocs de zéros écrits sur un LUN (unité de stockage) lorsque la compression en ligne est activée. Les utilitaires tels que l'utilitaire de récupération Oracle ASM (ASRU) fonctionnent en écrivant des zéros dans les étendues ASM inutilisées.

Cela permet aux administrateurs de bases de données de récupérer de l'espace sur la baie de stockage après la suppression de données. ONTAP intercepte les zéros et libère l'espace du LUN. Le processus de récupération est extrêmement rapide car aucune donnée n'est réellement écrite dans le système de stockage.

Du point de vue de la base de données, le groupe de disques ASM contient des zéros et la lecture de ces régions des LUN entraîne un flux de zéros, mais ONTAP ne stocke pas les zéros sur les disques. Des modifications simples des métadonnées sont effectuées en interne pour marquer les régions mises à zéro de la LUN comme vides de toutes les données.

Pour des raisons similaires, le test de performance impliquant des données mises à zéro n'est pas valide, car les blocs de zéros ne sont pas réellement traités comme des écritures dans la baie de stockage.

Principales considérations relatives à l'ASRU avec ASA r2 ONTAP:

- Fonctionne de la même manière AFF/ FAS pour les charges de travail SAN car ASA r2 est uniquement de type bloc.
- S'applique aux LUN et aux espaces de noms NVMe provisionnés dans les SAZ.
- Il n'existe pas de volumes FlexVol , mais le comportement de récupération des blocs nuls est identique.

Lorsque vous utilisez ASRU, assurez-vous que tous les correctifs recommandés par Oracle sont installés.

Virtualisation

La virtualisation des bases de données avec VMware, Oracle OLVM ou KVM est un choix de plus en plus courant pour les clients NetApp qui ont choisi la virtualisation, même pour leurs bases de données les plus stratégiques.

Prise en charge

Il existe de nombreuses idées fausses sur les politiques de prise en charge d'Oracle pour la virtualisation, en particulier pour les produits VMware. Il n'est pas rare d'entendre qu'Oracle ne prend pas en charge la virtualisation. Cette notion est incorrecte et ne permet pas de bénéficier de la virtualisation. Oracle Doc ID 249212.1 traite des besoins réels et est rarement considéré par les clients comme un problème.

Si un problème se produit sur un serveur virtualisé et que ce problème n'est pas encore connu du support Oracle, le client peut être invité à reproduire le problème sur du matériel physique. Si un client Oracle exécute une version de pointe d'un produit, il est possible qu'il ne souhaite pas utiliser la virtualisation en raison de problèmes de prise en charge potentiels, mais cette situation n'est pas réelle pour les clients qui utilisent des versions de produits Oracle généralement disponibles.

Présentation du stockage

Les clients qui envisagent de virtualiser leurs bases de données doivent baser leurs décisions de stockage sur leurs besoins métier. Bien qu'il s'agisse généralement d'une véritable déclaration pour toutes les décisions IT, elle est particulièrement importante pour les projets de base de données, car la taille et le champ d'application des exigences varient considérablement.

Il existe trois options de base pour la présentation du stockage :

- LUN virtualisées sur les datastores de l'hyperviseur
- LUN iSCSI gérées par l'initiateur iSCSI sur la machine virtuelle, pas par l'hyperviseur
- Systèmes de fichiers NFS montés par la machine virtuelle (pas à partir d'un datastore basé sur NFS)
- Mappages directs de périphériques. Les clients ne préfèrent pas les RDM VMware, mais les périphériques physiques sont souvent mappés directement avec la virtualisation KVM et OLVM.

Performance

La méthode de présentation du stockage à un invité virtualisé n'a généralement pas d'incidence sur les performances. Les systèmes d'exploitation hôtes, les pilotes réseau virtualisés et les implémentations de datastores d'hyperviseurs sont tous optimisés et peuvent généralement utiliser toute la bande passante réseau FC ou IP disponible entre l'hyperviseur et le système de stockage, dans la mesure où les meilleures pratiques de base sont respectées. Dans certains cas, il peut être légèrement plus facile d'obtenir des performances optimales en utilisant une approche de présentation du stockage par rapport à une autre, mais le résultat final devrait être comparable.

Gestion aisée

Le facteur clé dans la décision de présenter le stockage à un invité virtualisé est la mangeabilité. Il n'y a pas de bonne ou de mauvaise méthode. La meilleure approche dépend des besoins, des compétences et des

préférences opérationnels de l'IT.

Les facteurs à prendre en compte sont les suivants :

- **Transparence.** lorsqu'une machine virtuelle gère ses systèmes de fichiers, il est plus facile pour un administrateur de base de données ou un administrateur système d'identifier la source des systèmes de fichiers pour leurs données. L'accès aux systèmes de fichiers et aux LUN est différent de celui d'un serveur physique.
- **Cohérence.** lorsqu'une machine virtuelle possède ses systèmes de fichiers, l'utilisation ou la non-utilisation d'une couche hyperviseur affecte la gestion. Les mêmes procédures de provisionnement, de surveillance, de protection des données, etc. Peuvent être utilisées dans l'ensemble du parc, y compris dans les environnements virtualisés et non virtualisés.

  Par contre, dans un data Center 100 % virtualisé, il peut être préférable d'utiliser un stockage basé sur un datastore pour l'ensemble de l'encombrement, selon la même logique que celle mentionnée ci-dessus. Cohérence : possibilité d'utiliser les mêmes procédures de provisionnement, de protection, de regroupement et de protection des données.

- **Stabilité et dépannage.** lorsqu'une machine virtuelle possède ses systèmes de fichiers, il est plus simple de fournir des performances stables et de résoudre les problèmes car la pile de stockage complète est présente sur la machine virtuelle. Le seul rôle de l'hyperviseur est de transporter des trames FC ou IP. Lorsqu'un datastore est inclus dans une configuration, il complique la configuration en introduisant un autre ensemble d'expirations de délai, de paramètres, de fichiers journaux et de bogues potentiels.
- **Portabilité.** lorsqu'une machine virtuelle possède ses systèmes de fichiers, le processus de déplacement d'un environnement Oracle devient beaucoup plus simple. Les systèmes de fichiers peuvent facilement être déplacés entre des invités virtualisés et non virtualisés.
- **Dépendance vis-à-vis d'un fournisseur.** une fois les données placées dans un datastore, il devient difficile d'utiliser un hyperviseur différent ou de retirer les données de l'environnement virtualisé.
- **Activation de Snapshot.** les procédures de sauvegarde traditionnelles dans un environnement virtualisé peuvent devenir problématiques en raison de la bande passante relativement limitée. Par exemple, un agrégat 10 GbE à quatre ports peut suffire pour répondre aux besoins quotidiens en performances de nombreuses bases de données virtualisées, mais ce trunk ne permet pas d'effectuer des sauvegardes à l'aide de RMAN ou d'autres produits de sauvegarde nécessitant le streaming d'une copie complète des données. Résultat : un environnement virtualisé de plus en plus consolidé doit effectuer des sauvegardes via des snapshots de stockage. Ainsi, il n'est pas nécessaire de surconstruire la configuration de l'hyperviseur uniquement pour prendre en charge les besoins en bande passante et en CPU dans la fenêtre de sauvegarde.

  L'utilisation de systèmes de fichiers détenus par les clients facilite parfois l'exploitation des sauvegardes et des restaurations basées sur des snapshots, car les objets de stockage nécessitant une protection peuvent être ciblés plus facilement. Cependant, de plus en plus de produits de protection des données de virtualisation s'intègrent bien aux datastores et aux snapshots. La stratégie de sauvegarde doit être totalement adoptée avant de décider comment présenter le stockage à un hôte virtualisé.

Pilotes paravirtualisés

Pour des performances optimales, l'utilisation de pilotes de réseau paravirtualisés est essentielle. Lorsqu'un datastore est utilisé, un pilote SCSI paravirtualisé est requis. Un pilote de périphérique paravirtualisé permet à un invité de s'intégrer plus profondément dans l'hyperviseur, au lieu d'un pilote émulé dans lequel l'hyperviseur passe plus de temps CPU à imiter le comportement du matériel physique.

Saturation de la mémoire RAM

La saturation de la mémoire RAM implique la configuration d'une quantité de mémoire RAM virtualisée supérieure à celle qui existe sur le matériel physique sur différents hôtes. Cela peut entraîner des problèmes de performances inattendus. Lors de la virtualisation d'une base de données, les blocs sous-jacents de la SGA d'Oracle ne doivent pas être remplacés par l'hyperviseur vers le stockage. Cela entraîne des résultats de performances très instables.

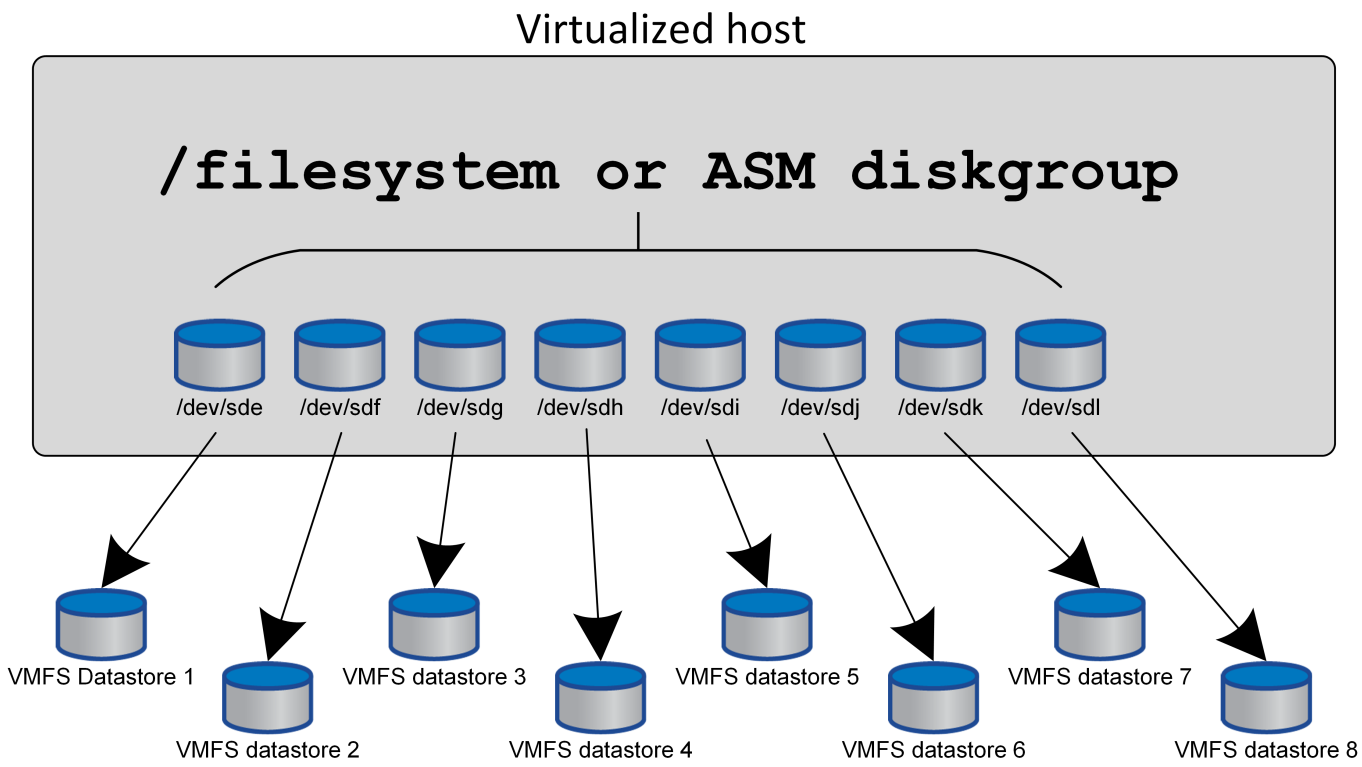
Répartition des datastores

Lorsque vous utilisez des bases de données avec des datastores, un facteur critique est à prendre en compte en ce qui concerne la répartition des performances.

Les technologies de datastore, telles que VMFS, peuvent couvrir plusieurs LUN, mais ne sont pas des périphériques répartis. Les LUN sont concaténées. Il peut en résulter des points sensibles de la LUN. Par exemple, une base de données Oracle standard peut disposer d'un groupe de disques ASM à 8 LUN. Les 8 LUN virtualisées pourraient être provisionnées sur un datastore VMFS de 8 LUN, mais il n'y a aucune garantie sur les LUN sur lesquelles les données résideront. La configuration résultante peut être les 8 LUN virtualisées occupant une seule LUN au sein du datastore VMFS. Cela risque d'engorgement des performances.

La répartition est généralement requise. Avec certains hyperviseurs, dont KVM, il est possible de créer un datastore à l'aide de la répartition LVM, comme décrit ci-dessous "ici". Avec VMware, l'architecture semble un peu différente. Chaque LUN virtualisée doit être placée sur un datastore VMFS différent.

Par exemple :



Le facteur principal de cette approche n'est pas le ONTAP. En raison de la limitation inhérente du nombre d'opérations qu'une seule machine virtuelle ou LUN d'hyperviseur peut traiter en parallèle, En règle générale, un LUN ONTAP peut prendre en charge beaucoup plus d'IOPS qu'un hôte ne peut en demander. La limite de performances d'une seule LUN est presque universellement due au système d'exploitation hôte. Ainsi, la plupart des bases de données ont besoin de 4 à 8 LUN pour répondre à leurs besoins de performance.

Les architectures VMware doivent planifier soigneusement leurs architectures pour s'assurer que cette approche ne permet pas d'optimiser le datastore et/ou le chemin LUN. Par ailleurs, il n'est pas nécessaire de disposer d'un ensemble unique de datastores VMFS pour chaque base de données. Le principal besoin est de s'assurer que chaque hôte dispose d'un ensemble propre de 4-8 chemins d'E/S entre les LUN virtualisées et les LUN back-end sur le système de stockage lui-même. Dans de rares cas, des exigences de performances vraiment extrêmes peuvent se révéler bénéfiques pour encore plus de données, mais 4-8 LUN suffisent généralement pour 95 % de toutes les bases de données. Un volume ONTAP unique contenant 8 LUN peut prendre en charge jusqu'à 250,000 000 IOPS de bloc Oracle aléatoires avec une configuration type OS/ONTAP/réseau.

Tiering

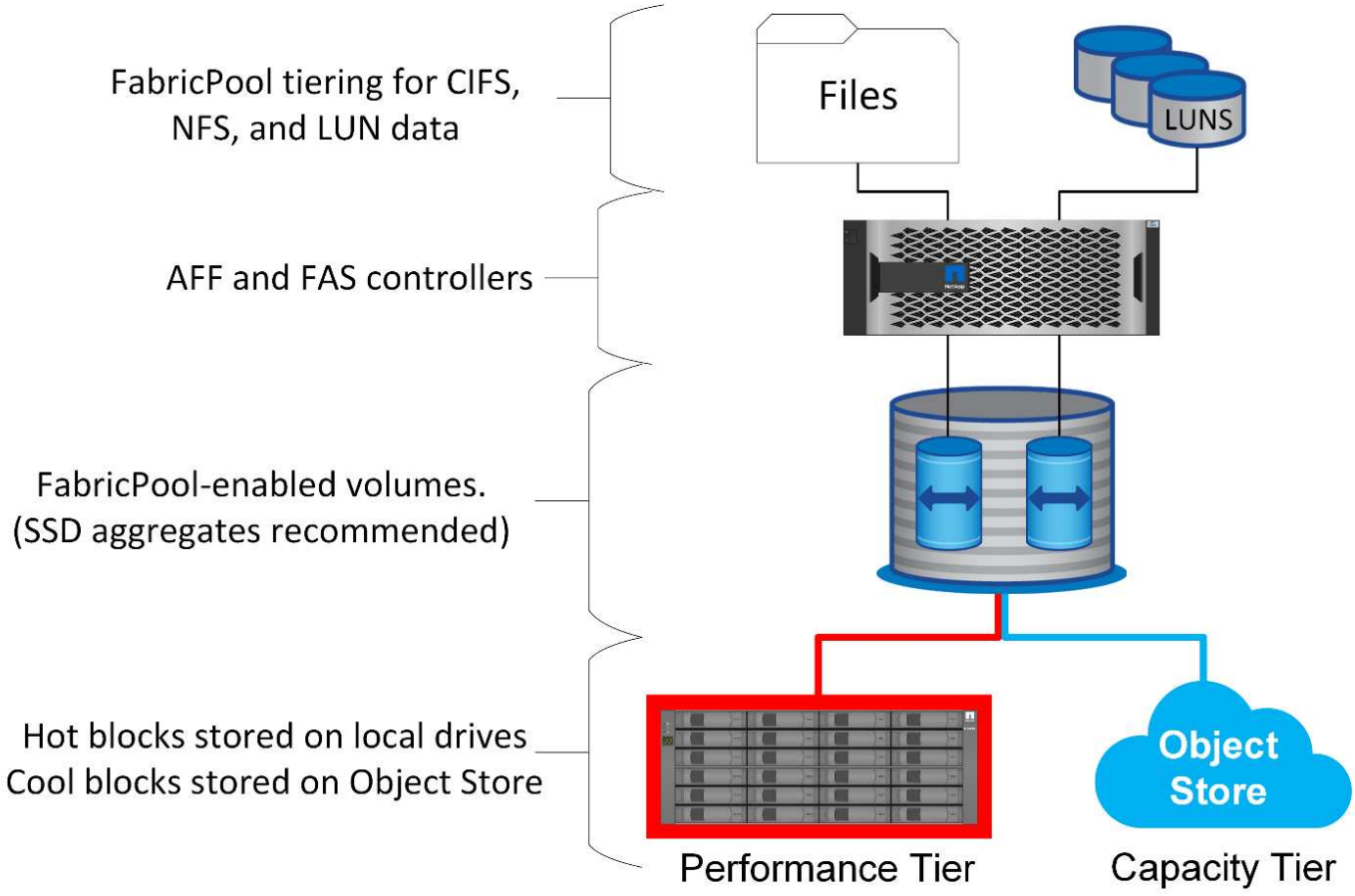
Présentation

Pour comprendre l'impact du Tiering FabricPool sur Oracle et d'autres bases de données, il est nécessaire de connaître l'architecture FabricPool de bas niveau.

Architecture

FabricPool est une technologie de hiérarchisation qui classe les blocs « actifs » ou « froids » et les place dans le Tier de stockage le plus approprié. Le Tier de performance se trouve le plus souvent sur un stockage SSD et héberge les blocs de données fortement sollicités. Le Tier de capacité se trouve dans un magasin d'objets et héberge les blocs de données utiles. Elle prend en charge le stockage objet, notamment NetApp StorageGRID, ONTAP S3, Microsoft Azure Blob Storage, le service de stockage objet Alibaba Cloud, IBM Cloud Object Storage, Google Cloud Storage et Amazon AWS S3.

Plusieurs règles de Tiering sont disponibles pour contrôler la façon dont les blocs sont classés comme actifs ou froids. Il est également possible de définir des règles par volume et de les modifier selon les besoins. Seuls les blocs de données sont déplacés entre les tiers de performance et de capacité. Les métadonnées qui définissent la structure des LUN et du système de fichiers restent toujours sur le Tier de performance. La gestion est ainsi centralisée sous ONTAP. Les fichiers et les LUN n'apparaissent pas différents des données stockées dans une autre configuration ONTAP. Le contrôleur NetApp AFF ou FAS applique les règles définies pour déplacer les données vers le Tier approprié.



Fournisseurs de magasins d'objets

Les protocoles de stockage objet utilisent de simples requêtes HTTP ou HTTPS pour stocker un grand nombre d'objets de données. L'accès au stockage objet doit être fiable, car l'accès aux données depuis ONTAP dépend du traitement rapide des demandes. Notamment Amazon S3 Standard et Infrequent Access, Microsoft Azure Hot Blob Storage, IBM Cloud et Google Cloud. Les options d'archivage telles qu'Amazon Glacier et Amazon Archive ne sont pas prises en charge, car le temps nécessaire à la récupération des données peut dépasser les tolérances des systèmes d'exploitation et des applications hôtes.

NetApp StorageGRID est également pris en charge et constitue une solution optimale. C'est un système de stockage objet haute performance, évolutif et hautement sécurisé qui assure une redondance géographique pour les données FabricPool ainsi que pour les autres applications de magasin d'objets qui font de plus en plus partie des environnements applicatifs d'entreprise.

StorageGRID peut également réduire les coûts en évitant les frais de sortie imposés par de nombreux fournisseurs de cloud public pour la lecture des données de leurs services.

Données et métadonnées

Notez que le terme « données » s'applique ici aux blocs de données réels, et non aux métadonnées. Seuls les blocs de données sont hiérarchisés, tandis que les métadonnées restent dans le Tier de performance. En outre, l'état d'un bloc en tant que bloc chaud ou froid n'est affecté que par la lecture du bloc de données réel. La simple lecture du nom, de l'horodatage ou des métadonnées de propriété d'un fichier n'affecte pas l'emplacement des blocs de données sous-jacents.

Sauvegardes

Même si FabricPool permet de réduire considérablement l'encombrement du stockage, il ne s'agit pas à lui seul d'une solution de sauvegarde. Les métadonnées NetApp WAFL restent toujours sur le Tier de performance. Si un incident catastrophique détruit le Tier de performance, il est impossible de créer un nouvel environnement à l'aide des données du Tier de capacité, car il ne contient pas de métadonnées WAFL.

FabricPool peut cependant faire partie d'une stratégie de sauvegarde. Par exemple, FabricPool peut être configuré avec la technologie de réplication NetApp SnapMirror. Chaque moitié du miroir peut avoir sa propre connexion à une cible de stockage objet. Vous obtenez ainsi deux copies indépendantes des données. La copie principale se compose des blocs du niveau de performance et des blocs associés du niveau de capacité, tandis que la réplique constitue un second ensemble de blocs de performance et de capacité.

Règles de hiérarchisation

Règles de hiérarchisation

Quatre règles sont disponibles dans ONTAP, qui contrôlent la façon dont les données Oracle du niveau de performance deviennent candidates à la relocalisation vers le niveau de capacité.

Copies Snapshot uniquement

Le `snapshot-only` `tiering-policy` s'applique uniquement aux blocs qui ne sont pas partagés avec le système de fichiers actif. Elle entraîne essentiellement une hiérarchisation des sauvegardes de bases de données. Les blocs deviennent candidats au Tiering après la création d'une copie Snapshot et l'écrasement du bloc, ce qui entraîne l'affichage d'un bloc uniquement dans la copie Snapshot. Le délai avant un `snapshot-only` le bloc est considéré comme froid est contrôlé par le `tiering-minimum-cooling-days` réglage du volume. La plage à partir de ONTAP 9.8 est de 2 à 183 jours.

De nombreux jeux de données ont des taux de modification faibles, ce qui permet de réduire au minimum les économies réalisées grâce à cette règle. Par exemple, un taux de modification hebdomadaire d'une base de données type observée sur ONTAP est inférieur à 5 %. Les journaux d'archivage de base de données peuvent occuper un espace important, mais ils continuent généralement d'exister dans le système de fichiers actif et ne sont donc pas candidats à la hiérarchisation dans le cadre de cette règle.

Auto

Le `auto` la règle de tiering étend le tiering aux blocs spécifiques de snapshot et aux blocs dans le système de fichiers actif. Le délai avant qu'un bloc soit considéré comme froid est contrôlé par le `tiering-minimum-cooling-days` réglage du volume. La plage à partir de ONTAP 9.8 est de 2 à 183 jours.

Cette approche permet d'activer des options de hiérarchisation qui ne sont pas disponibles avec le `snapshot-only` politique. Par exemple, une règle de protection des données peut nécessiter la conservation de 90 jours de certains fichiers journaux. Si vous définissez une période de refroidissement de 3 jours, tous les fichiers journaux de plus de 3 jours doivent être placés hors de la couche de performances. Cela libère un espace considérable sur le Tier de performance tout en vous permettant de consulter et de gérer l'ensemble des 90 jours de données.

Aucune

Le `none` la règle de tiering empêche tout bloc supplémentaire d'être hiérarchisé de la couche de stockage, mais toutes les données qui se trouvent toujours dans le tier de capacité restent dans le tier de capacité jusqu'à ce qu'elles soient lues. Si le bloc est ensuite lu, il est retiré et placé sur le Tier de performance.

La principale raison d'utiliser le `none` la règle de tiering consiste à empêcher les blocs d'être hiérarchisés, mais elle peut s'avérer utile pour modifier les règles au fil du temps. Par exemple, imaginons qu'un dataset spécifique soit beaucoup hiérarchisé vers la couche de capacité, mais qu'un besoin inattendu de fonctionnalités de performance complètes se produit. La règle peut être modifiée pour éviter tout Tiering supplémentaire et confirmer que tous les blocs lus en cas d'augmentation des E/S restent dans le Tier de performance.

Tout

Le `all` la règle de tiering remplace la `backup` Politique à partir de ONTAP 9.6. Le `backup` Règle appliquée uniquement aux volumes de protection des données, c'est-à-dire une destination SnapMirror ou NetApp SnapVault. Le `all` les règles fonctionnent de même, mais ne se limitent pas aux volumes de protection des données.

Avec cette règle, les blocs sont immédiatement considérés comme « cool » et peuvent être immédiatement hiérarchisés jusqu'à la couche de capacité.

Cette règle est particulièrement appropriée pour les sauvegardes à long terme. Il peut également être utilisé comme une forme de gestion hiérarchique du stockage (HSM). Auparavant, HSM était couramment utilisé pour classer les blocs de données d'un fichier sur bande tout en gardant le fichier lui-même visible sur le système de fichiers. Un volume FabricPool avec `all` cette stratégie vous permet de stocker des fichiers dans un espace visible et gérable, tout en ne consommant quasiment aucun espace sur le niveau de stockage local.

Stratégies de récupération

Les règles de Tiering contrôlent quels blocs de base de données Oracle sont hiérarchisés du niveau de performance au niveau de capacité. Les règles de récupération contrôlent ce qui se passe lorsqu'un bloc qui a été hiérarchisé est lu.

Valeur par défaut

Tous les volumes FabricPool sont initialement définis sur `default`, ce qui signifie que le comportement est contrôlé par la `cloud-retrieval-policy. `le comportement exact dépend de la règle de hiérarchisation utilisée.

- `auto`- récupérer uniquement les données lues de façon aléatoire
- `snapshot-only`- récupérer toutes les données lues de manière séquentielle ou aléatoire
- `none`- récupérer toutes les données lues de manière séquentielle ou aléatoire
- `all`- ne pas récupérer les données du niveau de capacité

En lecture

Réglage `cloud-retrieval-policy` la lecture remplace le comportement par défaut, de sorte qu'une lecture de toutes les données hiérarchisées entraîne le renvoi de ces données vers le niveau de performance.

Par exemple, un volume peut avoir été légèrement utilisé pendant une longue période sous le `auto` la règle de tiering et la plupart des blocs sont désormais hiérarchisés.

Si une modification inattendue des besoins de l'entreprise nécessitait l'analyse répétée de certaines données pour préparer un rapport spécifique, il peut être souhaitable de modifier le `cloud-retrieval-policy` à `on-read` pour garantir que toutes les données lues sont renvoyées au niveau de performances, y compris les données lues de manière séquentielle et aléatoire. Cela améliorerait les performances des E/S séquentielles par rapport au volume.

Promouvoir

Le comportement de la règle de promotion dépend de la règle de hiérarchisation. Si la règle de hiérarchisation est `auto`, puis réglage du `cloud-retrieval-policy `to `promote` ramène tous les blocs du tier de capacité à l'analyse de tiering suivante.

Si la règle de hiérarchisation est `snapshot-only`, les seuls blocs renvoyés sont les blocs associés au système de fichiers actif. Normalement, cela n'aurait aucun effet car les seuls blocs placés sous le sont `snapshot-only` la règle serait les blocs associés exclusivement aux snapshots. Il n'y aurait pas de blocs hiérarchisés dans le système de fichiers actif.

Toutefois, si une SnapRestore de volume ou une opération de clonage de fichiers a été effectuée pour restaurer les données d'un volume à partir d'un snapshot, le système de fichiers actif peut désormais avoir besoin de certains blocs qui ont été hiérarchisés, car ils n'étaient associés qu'à des snapshots. Il peut être souhaitable de modifier temporairement le `cloud-retrieval-policy` règle à `promote` pour récupérer rapidement tous les blocs localement requis.

Jamais

Ne récupérez pas les blocs du niveau de capacité.

Stratégies de Tiering

Tiering complet de fichiers

Bien que le Tiering FabricPool fonctionne au niveau des blocs, il peut dans certains cas servir à fournir un Tiering au niveau des fichiers.

De nombreux jeux de données d'applications sont organisés par date, et ces données sont généralement moins susceptibles d'être accessibles au fur et à mesure du vieillissement. Par exemple, une banque peut disposer d'un référentiel de fichiers PDF contenant cinq années de relevés clients, mais seuls les derniers mois sont actifs. FabricPool peut être utilisé pour déplacer d'anciens fichiers de données vers le Tier de capacité. Une période de refroidissement de 14 jours permettrait de conserver les fichiers PDF de 14 jours les plus récents sur le niveau de performance. En outre, les fichiers lus au moins tous les 14 jours resteraient fortement sollicités et resteraient donc sur le Tier de performance.

Stratégies

Pour mettre en œuvre une approche de hiérarchisation basée sur des fichiers, vous devez avoir des fichiers écrits et non modifiés par la suite. Le `tiering-minimum-cooling-days` la règle doit être définie suffisamment haut pour que les fichiers dont vous avez besoin restent sur le tier de performance. Par exemple, un jeu de données pour lequel les 60 derniers jours de données sont requis avec des performances optimales garantit le paramétrage du `tiering-minimum-cooling-days` période jusqu'en 60. Des résultats similaires peuvent également être obtenus en fonction des modèles d'accès aux fichiers. Par exemple, si les 90 derniers jours de données sont requis et que l'application accède à cette période de 90 jours, les données restent sur le Tier de performance. En réglant le `tiering-minimum-cooling-days` sur 2, le tiering s'affiche rapidement une fois les données moins actives.

Le `auto` la règle est requise pour la hiérarchisation de ces blocs, car uniquement le système `auto` la règle affecte les blocs qui se trouvent dans le système de fichiers actif.

Tout type d'accès aux données réinitialise les données de la carte thermique. L'analyse antivirus, l'indexation et même l'activité de sauvegarde qui lit les fichiers source empêchent le Tiering, car les besoins sont importants `tiering-minimum-cooling-days` le seuil n'est jamais atteint.

Tiering partiel des fichiers

Comme FabricPool fonctionne au niveau des blocs, les fichiers susceptibles d'être modifiés peuvent être partiellement hiérarchisés vers un stockage objet tout en restant partiellement sur le Tier de performance.

Ceci est courant avec les bases de données. Les bases de données qui contiennent des blocs inactifs sont également candidates au Tiering FabricPool. Par exemple, une base de données de gestion de la chaîne logistique peut contenir des informations historiques qui doivent être disponibles si nécessaire, mais qui ne sont pas accessibles pendant les opérations normales. FabricPool peut être utilisé pour déplacer de manière sélective les blocs inactifs.

Par exemple, les fichiers de données s'exécutant sur un volume FabricPool avec un `tiering-minimum-cooling-days` la période de 90 jours permet de conserver les blocs auxquels le tier de performance accède au cours des 90 jours précédents. Toutefois, tout élément non utilisé pendant 90 jours est transféré vers le niveau de capacité. Dans d'autres cas, l'activité normale de l'application préserve les blocs corrects du niveau approprié. Par exemple, si une base de données est normalement utilisée pour traiter les 60 jours précédents de données sur une base régulière, c'est beaucoup moins `tiering-minimum-cooling-days` la période peut être définie car l'activité naturelle de l'application s'assure que les blocs ne sont pas déplacés prématurément.

Le `auto` la politique doit être utilisée avec soin pour les bases de données. De nombreuses bases de données ont des activités périodiques, comme le processus de fin de trimestre ou les opérations de réindexation. Si la période de ces opérations est supérieure à `tiering-minimum-cooling-days` des problèmes de performances peuvent se produire. Par exemple, si le traitement de fin de trimestre nécessite 1 To de données qui n'étaient pas modifiées, ces données peuvent maintenant être présentes sur le niveau de capacité. Les lectures à partir du niveau de capacité sont souvent extrêmement rapides et ne provoquent pas de problèmes de performance, mais les résultats exacts dépendent de la configuration du magasin d'objets.

Stratégies

Le `tiering-minimum-cooling-days` la règle doit être définie de manière suffisamment élevée pour conserver les fichiers qui peuvent être requis sur le niveau de performance. Par exemple, une base de données dans laquelle les 60 derniers jours de données peuvent être requis avec des performances optimales justifierait de définir le `tiering-minimum-cooling-days` période à 60 jours. Des résultats similaires pourraient également être obtenus en fonction des modèles d'accès aux fichiers. Par exemple, si les 90 derniers jours de données sont requis et que l'application accède à cette période de 90 jours, les données restent sur le Tier de performance. Réglage du `tiering-minimum-cooling-days` une période de 2 jours permettrait de hiérarchiser les données rapidement lorsque celles-ci deviennent moins actives.

Le `auto` la règle est requise pour la hiérarchisation de ces blocs, car uniquement le système `auto` la règle affecte les blocs qui se trouvent dans le système de fichiers actif.

Tout type d'accès aux données réinitialise les données de la carte thermique. Par conséquent, les analyses de la table complète des bases de données, et même les opérations de sauvegarde qui lisent les fichiers source, empêchent le Tiering, car nécessaire `tiering-minimum-cooling-days` le seuil n'est jamais atteint.

Tiering des journaux d'archivage

L'utilisation la plus importante pour FabricPool est peut-être l'amélioration de l'efficacité des données inactives connues, telles que les journaux de transactions de base de données.

La plupart des bases de données relationnelles opèrent en mode d'archivage du journal de transactions pour assurer une restauration instantanée. Les modifications apportées aux bases de données sont validées en enregistrant les modifications dans les journaux de transactions et le journal de transactions est conservé sans être écrasé. Il peut donc s'avérer nécessaire de conserver un énorme volume de journaux de transactions archivés. De nombreux autres workflows applicatifs génèrent des données qui doivent être conservées, mais il est très peu probable qu'elles soient accessibles.

Pour résoudre ces problèmes, FabricPool propose une solution unique avec hiérarchisation intégrée. Les fichiers sont stockés et restent accessibles à leur emplacement habituel, mais ne prennent pratiquement pas d'espace sur la baie principale.

Stratégies

Utiliser un `tiering-minimum-cooling-days` la règle de quelques jours permet de conserver les blocs dans les fichiers récemment créés (les fichiers les plus susceptibles d'être requis à court terme) sur le niveau de performance. Les blocs de données des anciens fichiers sont ensuite déplacés vers le niveau de capacité.

Le `auto` applique la hiérarchisation des invites lorsque le seuil de refroidissement a été atteint, que les journaux aient été supprimés ou qu'ils continuent d'exister dans le système de fichiers principal. Le stockage de tous les journaux potentiellement requis dans un seul emplacement du système de fichiers actif simplifie également la gestion. Il n'y a aucune raison de rechercher un fichier à restaurer à l'aide de snapshots.

Certaines applications, telles que Microsoft SQL Server, tronquent les fichiers journaux de transactions pendant les opérations de sauvegarde afin que les journaux ne soient plus dans le système de fichiers actif. Il est possible d'économiser de la capacité à l'aide de `snapshot-only` la règle de tiering, mais la règle `auto` la règle n'est pas utile pour les données de journal car il devrait rarement y avoir des données de journal refroidies dans le système de fichiers actif.

Tiering Snapshot

La version initiale de FabricPool a ciblé le cas d'utilisation de la sauvegarde. Les seuls types de blocs qui ont pu être hiérarchisés sont les blocs qui n'étaient plus associés aux données dans le système de fichiers actif. Par conséquent, seuls les blocs de données des snapshots peuvent être déplacés vers le niveau de capacité. Il s'agit là de l'une des options de hiérarchisation les plus sécurisées lorsque vous devez vous assurer que les performances ne sont jamais affectées.

Règles - snapshots locaux

Deux options sont disponibles pour le Tiering des blocs de snapshots inactifs vers le niveau de capacité. Tout d'abord, le `snapshot-only` la règle cible uniquement les blocs de snapshot. Bien que le `auto` la politique

inclut le `snapshot-only` et tiering des blocs à partir du système de fichiers actif. Ce n'est peut-être pas souhaitable.

Le `tiering-minimum-cooling-days` value doit être défini sur une période qui met à disposition les données éventuellement requises lors d'une restauration sur le tier de performance. Par exemple, la plupart des scénarios de restauration d'une base de données de production stratégique incluent un point de restauration à un moment donné au cours des jours précédents. Réglage a `tiering-minimum-cooling-days` la valeur 3 garantit que toute restauration du fichier entraîne un fichier qui offre immédiatement des performances maximales. Tous les blocs des fichiers actifs sont toujours présents sur un système de stockage rapide sans avoir à les restaurer à partir du niveau de capacité.

Règles - snapshots répliqués

Les snapshots répliqués avec SnapMirror ou SnapVault, uniquement utilisés pour la restauration, doivent généralement utiliser FabricPool `all` politique. Avec cette règle, les métadonnées sont répliquées, mais tous les blocs de données sont immédiatement envoyés au niveau de capacité pour des performances maximales. La plupart des processus de restauration impliquent des E/S séquentielles, ce qui est intrinsèquement efficace. Le délai de restauration à partir de la destination du magasin d'objets doit être évalué, mais dans une architecture bien conçue, ce processus de restauration ne doit pas nécessairement être beaucoup plus lent que la restauration à partir de données locales.

Si les données répliquées sont également destinées à être utilisées pour le clonage, le `auto` la politique est plus appropriée, avec un `tiering-minimum-cooling-days` valeur qui englobe les données qui doivent être utilisées régulièrement dans un environnement de clonage. Par exemple, le jeu de travail actif d'une base de données peut inclure des données lues ou écrites au cours des trois jours précédents, mais il peut également inclure 6 mois de données historiques supplémentaires. Si oui, alors le `auto` La règle appliquée à la destination SnapMirror met à disposition le jeu de travail sur le Tier de performance.

Tiering de sauvegarde

Les sauvegardes d'applications traditionnelles incluent des produits tels qu'Oracle Recovery Manager, qui créent des sauvegardes basées sur des fichiers en dehors de l'emplacement de la base de données d'origine.

```
 `tiering-minimum-cooling-days` policy of a few days preserves the most
recent backups, and therefore the backups most likely to be required for
an urgent recovery situation, on the performance tier. The data blocks of
the older files are then moved to the capacity tier.
Le `auto` la règle est la règle la plus appropriée pour les données de
sauvegarde. Cela garantit une hiérarchisation rapide lorsque le seuil de
refroidissement a été atteint, que les fichiers aient été supprimés ou
qu'ils continuent d'exister dans le système de fichiers principal. Le
stockage de tous les fichiers potentiellement requis dans un emplacement
unique du système de fichiers actif simplifie également la gestion. Il n'y
a aucune raison de rechercher un fichier à restaurer à l'aide de
snapshots.
```

Le `snapshot-only` la stratégie peut être mise en œuvre, mais elle s'applique uniquement aux blocs qui ne sont plus dans le système de fichiers actif. Par conséquent, les fichiers d'un partage NFS ou SMB doivent d'abord être supprimés avant de pouvoir placer les données dans un Tier.

Cette règle serait encore moins efficace avec une configuration de LUN, car la suppression d'un fichier d'une LUN supprime uniquement les références de fichier des métadonnées du système de fichiers. Les blocs réels des LUN restent en place jusqu'à ce qu'ils soient remplacés. Cette situation peut entraîner un délai long entre la suppression d'un fichier et l'écrasement des blocs et leur candidature à la hiérarchisation. Il est avantageux de déplacer le `snapshot-only` Bloque le niveau de capacité, mais, dans l'ensemble, la gestion FabricPool des données de sauvegarde fonctionne mieux avec le `auto` politique.

> Cette approche permet aux utilisateurs de gérer plus efficacement l'espace requis pour les sauvegardes, mais FabricPool lui-même n'est pas une technologie de sauvegarde. Le Tiering des fichiers de sauvegarde vers un magasin d'objets simplifie la gestion, car les fichiers restent visibles sur le système de stockage d'origine. Cependant, les blocs de données de destination du magasin d'objets dépendent du système de stockage d'origine. En cas de perte du volume source, les données du magasin d'objets ne sont plus utilisables.

Interruptions d'accès au magasin d'objets

Le Tiering d'un dataset avec FabricPool entraîne une dépendance entre la baie de stockage primaire et le Tier de magasin d'objets. De nombreuses options de stockage objet offrent différents niveaux de disponibilité. Il est important de comprendre l'impact d'une éventuelle perte de connectivité entre la baie de stockage primaire et le niveau de stockage objet.

Si une E/S émise par ONTAP nécessite des données du niveau de capacité et que les ONTAP ne peuvent pas atteindre le niveau de capacité pour récupérer des blocs, les E/S finissent par être sorties. L'effet de ce délai dépend du protocole utilisé. Dans un environnement NFS, ONTAP répond par une réponse EJUKEBOX ou EDELAY, selon le protocole. Certains systèmes d'exploitation plus anciens peuvent interpréter cela comme une erreur, mais les systèmes d'exploitation actuels et les niveaux de correctifs actuels du client Oracle Direct NFS traitent cette erreur comme une nouvelle tentative et continuent d'attendre la fin des E/S.

Un délai plus court s'applique aux environnements SAN. Si un bloc de l'environnement de magasin d'objets est requis et reste inaccessible pendant deux minutes, une erreur de lecture est renvoyée à l'hôte. Le volume ONTAP et les LUN restent en ligne, mais le système d'exploitation hôte peut signaler le système de fichiers comme étant dans un état d'erreur.

Les problèmes de connectivité du stockage objet `snapshot-only` la politique est moins préoccupante, car seules les données de sauvegarde sont hiérarchisées. Les problèmes de communication ralentiraient la récupération des données, mais n'affecteraient pas les données utilisées activement. Le `auto` et `all` Les règles permettent le Tiering des données inactives de la LUN active, ce qui signifie qu'une erreur lors de la récupération des données du magasin d'objets peut affecter la disponibilité de la base de données. Un déploiement SAN doté de ces règles doit uniquement être utilisé avec un stockage objet de grande qualité et des connexions réseau conçues pour une haute disponibilité. NetApp StorageGRID est la meilleure option.

Protection des données Oracle

Protection des données avec ONTAP

NetApp sait que les bases de données contient les données les plus stratégiques.

Une entreprise ne peut pas fonctionner sans accéder à ses données, et parfois l'activité repose sur les données. Ces données doivent être protégées, mais la protection ne se limite pas à garantir une sauvegarde utilisable. Elle consiste également à effectuer des sauvegardes rapidement et de manière fiable en plus de les

stocker en toute sécurité.

L'autre côté de la protection des données est la restauration des données. Lorsque les données ne sont pas accessibles, l'entreprise est affectée et peut ne pas fonctionner tant qu'elle n'est pas restaurée. Ce processus doit être rapide et fiable. Enfin, la plupart des bases de données doivent être protégées contre les incidents, ce qui signifie maintenir une réplique de la base de données. La réplique doit être suffisamment à jour. Il doit également être rapide et simple de faire de la réplique une base de données entièrement opérationnelle.

Cette documentation remplace le rapport technique *TR-4591 : protection des données Oracle : sauvegarde, restauration et réplication.*

Planification

Une architecture de protection des données d'entreprise adaptée dépend des exigences de l'entreprise concernant la conservation des données, la restauration et la tolérance aux perturbations à divers moments.

Prenons l'exemple du nombre d'applications, de bases de données et de datasets importants pris en compte. Il est relativement simple d'élaborer une stratégie de sauvegarde pour un seul dataset afin d'assurer la conformité aux SLA standard, car la gestion ne comporte pas beaucoup d'objets. À mesure que le nombre de jeux de données augmente, la surveillance devient plus complexe et les administrateurs peuvent être obligés de consacrer de plus en plus de temps aux pannes de sauvegarde. Dès qu'un environnement évolue, il faut adopter une approche totalement différente.

La taille des datasets affecte également la stratégie. Par exemple, le jeu de données étant si petit, de nombreuses options sont possibles pour la sauvegarde et la restauration avec une base de données de 100 Go. En général, la simple copie des données à partir du support de sauvegarde avec des outils classiques permet d'atteindre un RTO suffisant pour la restauration. Une base de données de 100 To a généralement besoin d'une stratégie totalement différente, sauf si le RTO autorise une panne de plusieurs jours. Dans ce cas, une procédure classique de sauvegarde et de restauration basée sur des copies peut être acceptable.

Enfin, il y a des facteurs en dehors du processus de sauvegarde et de restauration lui-même. Par exemple, existe-t-il des bases de données qui prennent en charge les activités de production stratégiques, faisant de la restauration un événement rare uniquement effectué par des administrateurs de bases de données qualifiés ? Ou bien, les bases de données font-elles partie d'un vaste environnement de développement dans lequel la restauration est fréquente et gérée par une équipe INFORMATIQUE généraliste ?

Planification des objectifs de durée de restauration, de point de récupération et des SLA

ONTAP vous permet d'adapter facilement une stratégie de protection des données des bases de données Oracle aux besoins de votre entreprise.

Ces exigences comprennent des facteurs tels que la vitesse de restauration, la perte de données maximale autorisée et les besoins de conservation des sauvegardes. Le plan de protection des données doit également tenir compte de diverses exigences réglementaires en matière de conservation et de restauration des données. Enfin, différents scénarios de restauration des données doivent être pris en compte, allant de la restauration classique et prévisible résultant d'erreurs d'utilisateurs ou d'applications à des scénarios de reprise sur incident incluant la perte complète d'un site.

Les modifications mineures apportées aux règles de protection et de restauration des données peuvent avoir un impact significatif sur l'architecture globale du stockage, de la sauvegarde et de la restauration. Il est essentiel de définir et de documenter des normes avant de commencer le travail de conception afin d'éviter de compliquer une architecture de protection des données. Des fonctions ou des niveaux de protection inutiles entraînent des coûts et des frais de gestion inutiles. Par ailleurs, une exigence initialement négligée peut

conduire un projet dans la mauvaise direction ou nécessiter des modifications de conception de dernière minute.

Objectif de délai de restauration

L'objectif de délai de restauration (RTO) définit le temps maximal autorisé pour la restauration d'un service. Par exemple, une base de données de ressources humaines peut atteindre un objectif de délai de restauration de 24 heures. En effet, même s'il ne serait pas très pratique de perdre l'accès à ces données pendant les jours de travail, l'entreprise peut tout de même fonctionner. En revanche, une base de données prenant en charge le grand livre d'une banque aurait un RTO mesuré en minutes, voire en secondes. Un objectif RTO de zéro n'est pas possible, car il doit y avoir un moyen de faire la différence entre une panne de service réelle et un événement de routine tel qu'un paquet réseau perdu. Toutefois, un objectif RTO quasi nul est généralement requis.

Objectif de point de récupération

L'objectif de point de récupération (RPO) définit la perte de données maximale tolérable. Dans de nombreux cas, l'objectif de point de récupération est uniquement déterminé par la fréquence des copies Snapshot ou des mises à jour snapmirror.

Dans certains cas, le RPO peut être rendu plus agressif, car il permet de protéger certaines données de manière sélective plus fréquemment. Dans un contexte de base de données, le RPO correspond généralement à la quantité de données perdues dans un journal spécifique. Dans un scénario de restauration typique dans lequel une base de données est endommagée en raison d'un bogue de produit ou d'une erreur utilisateur, le RPO doit être égal à zéro, ce qui signifie qu'il ne doit pas y avoir de perte de données. La procédure de restauration implique la restauration d'une copie antérieure des fichiers de base de données, puis la relecture des fichiers journaux pour ramener l'état de la base de données au point dans le temps souhaité. Les fichiers journaux requis pour cette opération doivent déjà être en place à l'emplacement d'origine.

Dans des scénarios inhabituels, les données des journaux peuvent être perdues. Par exemple, un accident ou un acte malveillant `rm -rf *` des fichiers de base de données peuvent entraîner la suppression de toutes les données. La seule option serait de restaurer des données à partir de sauvegardes, y compris des fichiers journaux, et certaines seraient inévitablement perdues. Dans un environnement de sauvegarde classique, la seule option permettant d'améliorer le RPO consiste à effectuer des sauvegardes répétées des données du journal. Cela a toutefois ses limites en raison du déplacement constant des données et de la difficulté à maintenir un système de sauvegarde en tant que service en continu. L'un des avantages des systèmes de stockage avancés est la capacité à protéger les données contre les dommages accidentels ou malveillants aux fichiers et à fournir ainsi un meilleur RPO sans déplacement des données.

Reprise après incident

La reprise après incident comprend l'architecture INFORMATIQUE, les règles et les procédures requises pour restaurer un service en cas d'incident physique. Cela peut inclure les inondations, les incendies ou les personnes agissant avec une intention malveillante ou négligente.

La reprise sur incident est bien plus qu'un ensemble de procédures de restauration. Il s'agit du processus complet d'identification des différents risques, de définition des exigences en matière de restauration des données et de continuité des services, et de mise à disposition de l'architecture appropriée avec les procédures associées.

Lors de l'établissement des exigences de protection des données, il est essentiel de faire la différence entre les objectifs RPO et RTO types et les exigences RPO et RTO requises pour la reprise après incident. Pour les situations de perte de données, allant d'une erreur utilisateur relativement normale à un incendie qui détruit un data Center, certains environnements applicatifs nécessitent un RPO nul et un RTO quasi nul. Cependant, il y

a des conséquences administratives et des coûts pour ces niveaux élevés de protection.

En général, les exigences de restauration des données non liées aux incidents doivent être strictes pour deux raisons. Tout d'abord, les bogues d'application et les erreurs d'utilisateur qui endommagent les données sont prévisibles au point qu'ils sont presque inévitables. Deuxièmement, il n'est pas difficile de concevoir une stratégie de sauvegarde capable de fournir un RPO nul et un RTO faible tant que le système de stockage n'est pas détruit. Il n'y a aucune raison de ne pas traiter un risque important facilement résolu. C'est pourquoi les objectifs RPO et RTO pour la reprise locale doivent être agressifs.

Les exigences en termes de RTO et de RPO pour la reprise d'activité varient plus largement en fonction du risque d'incident et des conséquences de la perte de données ou de l'interruption pour une entreprise. Les exigences en matière de RPO et de RTO doivent être basées sur les besoins réels de l'entreprise et non sur des principes généraux. Ils doivent prendre en compte plusieurs scénarios de catastrophe physique et logique.

Incidents logiques

Les incidents logiques incluent la corruption des données provoquée par les utilisateurs, les bogues des applications ou du système d'exploitation et les dysfonctionnements logiciels. Les incidents logiques peuvent également inclure des attaques malveillantes de tiers contenant des virus ou des vers, ou encore en exploitant les vulnérabilités des applications. Dans ces cas, l'infrastructure physique n'est pas endommagée, mais les données sous-jacentes ne sont plus valides.

Les ransomwares sont un type de catastrophe logique de plus en plus courant qui sert à chiffrer les données à l'aide d'un vecteur d'attaque. Le chiffrement n'endommage pas les données, mais il les rend indisponibles jusqu'à ce que le paiement soit effectué à un tiers. De plus en plus d'entreprises sont spécifiquement la cible de piratage. Face à cette menace, NetApp propose des snapshots inviolables où même l'administrateur du stockage ne peut pas modifier les données protégées avant la date d'expiration configurée.

Incidents physiques

Les incidents physiques incluent la défaillance de composants d'une infrastructure qui dépasse ses capacités de redondance et entraînent une perte de données ou une perte de service prolongée. Par exemple, la protection RAID assure la redondance des disques durs et l'utilisation de HBA assure la redondance des ports FC et des câbles FC. Les pannes matérielles de ces composants sont prévisibles et n'ont pas d'incidence sur la disponibilité.

Dans un environnement d'entreprise, il est généralement possible de protéger l'infrastructure d'un site entier avec des composants redondants au point où le seul scénario de catastrophe physique prévisible est la perte complète du site. La planification de la reprise d'activité dépend alors de la réplication de site à site.

Protection des données synchrone et asynchrone

Dans l'idéal, toutes les données seraient répliquées de manière synchrone sur des sites dispersés géographiquement. Une telle réplication n'est pas toujours possible, voire possible pour plusieurs raisons :

- La réplication synchrone entraîne inévitablement une augmentation de la latence d'écriture, car toutes les modifications doivent être répliquées vers les deux emplacements avant que l'application/la base de données ne puisse poursuivre le traitement. L'effet de performance qui en résulte est parfois inacceptable, excluant l'utilisation de la mise en miroir synchrone.
- En raison de l'adoption accrue de 100 % de stockage SSD, il est plus probable que l'on remarque une latence d'écriture supplémentaire, car les attentes en termes de performances comprennent des centaines de milliers d'IOPS et une latence inférieure à la milliseconde. Pour tirer pleinement parti de l'utilisation de 100 % des SSD, il peut être nécessaire de revoir la stratégie de reprise sur incident.
- La croissance des datasets en octets continue, ce qui engendre des défis en garantissant une bande

passante suffisante pour soutenir la réplication synchrone.

- La croissance des datasets s'accompagne également de défis liés à la gestion de la réplication synchrone à grande échelle.
- Les stratégies basées sur le cloud impliquent souvent des distances de réplication et une latence plus importantes, ce qui exclut davantage l'utilisation de la mise en miroir synchrone.

NetApp propose des solutions qui incluent à la fois la réplication synchrone pour satisfaire les besoins les plus exigeants en matière de restauration des données et des solutions asynchrones qui assurent des performances et une flexibilité accrues. De plus, la technologie NetApp s'intègre en toute transparence à de nombreuses solutions de réplication tierces, telles qu'Oracle DataGuard

Durée de conservation

Le dernier aspect d'une stratégie de protection des données est la durée de conservation des données, qui peut varier considérablement.

- Il est généralement nécessaire d'effectuer 14 jours de sauvegardes nocturnes sur le site principal et 90 jours de sauvegardes sur un site secondaire.
- De nombreux clients créent des archives trimestrielles autonomes stockées sur différents supports.
- Une base de données constamment mise à jour n'a peut-être pas besoin de données historiques, et les sauvegardes ne doivent être conservées que pendant quelques jours.
- Pour des raisons réglementaires, une capacité de restauration peut être nécessaire au point de toute transaction arbitraire dans une fenêtre de 365 jours.

Disponibilité de la base de données

ONTAP est conçu pour offrir une disponibilité maximale des bases de données Oracle. Ce document ne contient pas de description complète des fonctionnalités de haute disponibilité de ONTAP. Cependant, comme pour la protection des données, il est important de bien comprendre cette fonctionnalité lors de la conception d'une infrastructure de base de données.

Paires HA

L'unité de base de la haute disponibilité est la paire haute disponibilité. Chaque paire contient des liens redondants pour prendre en charge la réplication des données vers la mémoire NVRAM. La NVRAM n'est pas un cache d'écriture. La RAM à l'intérieur du contrôleur sert de cache d'écriture. L'objectif de la mémoire NVRAM est de journaliser temporairement les données afin de prévenir toute panne système inattendue. À cet égard, il est similaire à un fichier redo log de base de données.

La mémoire NVRAM et le journal de reprise de base de données sont utilisés pour stocker des données rapidement, ce qui permet d'y apporter les modifications le plus rapidement possible. La mise à jour des données persistantes sur les disques (ou fichiers de données) n'a lieu qu'une fois plus tard lors d'un processus appelé point de contrôle sur ONTAP et la plupart des plateformes de bases de données. Les données NVRAM et les redo logs de base de données ne sont pas lus pendant les opérations normales.

Si un contrôleur tombe en panne brusquement, des modifications sont susceptibles d'être en attente de stockage dans la mémoire NVRAM qui n'ont pas encore été écrites sur les disques. Le contrôleur partenaire détecte la panne, prend le contrôle des disques et applique les modifications requises qui ont été stockées dans la mémoire NVRAM.

Takeover et Giveback

Le basculement et le rétablissement font référence au processus de transfert de la responsabilité des ressources de stockage entre les nœuds d'une paire HA. Le basculement et le rétablissement sont deux aspects :

- Gestion de la connectivité réseau permettant l'accès aux lecteurs
- Gestion des disques eux-mêmes

Les interfaces réseau prenant en charge le trafic CIFS et NFS sont configurées avec un emplacement de home et de basculement. Il inclut le déplacement des interfaces réseau vers leur domicile temporaire sur une interface physique située sur le(s) même(s) sous-réseau que l'emplacement d'origine. Le rétablissement inclut le déplacement des interfaces réseau vers leurs emplacements d'origine. Le comportement exact peut être réglé selon les besoins.

Les interfaces réseau prenant en charge les protocoles de bloc SAN, tels que iSCSI et FC, ne sont pas déplacées pendant le basculement et le rétablissement. Les LUN doivent plutôt être provisionnées avec des chemins qui incluent une paire HA complète entraînant un chemin principal et un chemin secondaire.

Des chemins d'accès supplémentaires vers des contrôleurs supplémentaires peuvent également être configurés pour prendre en charge le déplacement des données entre les nœuds d'un cluster plus grand, mais cela ne fait pas partie du processus de haute disponibilité.

Le deuxième aspect du Takeover et Giveback est le transfert de la propriété de disque. Le processus exact dépend de plusieurs facteurs, notamment la raison du Takeover/Giveback et les options de ligne de commande émises. L'objectif est de réaliser l'opération aussi efficacement que possible. Bien que le processus global puisse sembler durer plusieurs minutes, le moment réel où la propriété du disque est transférée d'un nœud à un autre peut généralement se mesurer en secondes.

Temps de reprise

Les E/S de l'hôte font l'objet d'une courte pause au niveau des E/S lors des opérations de basculement et de rétablissement. Cependant, la configuration de l'environnement ne doit pas provoquer d'interruption des applications. Le processus de transition réel dans lequel les E/S sont retardées se mesure généralement en secondes, mais l'hôte peut avoir besoin de plus de temps pour reconnaître la modification des chemins de données et renvoyer les opérations d'E/S.

La nature de la perturbation dépend du protocole :

- Une interface réseau prenant en charge le trafic NFS et CIFS émet une requête ARP (Address Resolution Protocol) vers le réseau après la transition vers un nouvel emplacement physique. Les commutateurs réseau mettent ainsi à jour leurs tables d'adresses MAC (Media Access Control) et reprennent le traitement des E/S. L'interruption dans le cas d'un basculement et d'un rétablissement planifiés se mesure généralement en secondes et, dans la plupart des cas, elle n'est pas détectable. Certains réseaux peuvent être plus lents à reconnaître pleinement le changement de chemin réseau et certains systèmes d'exploitation peuvent mettre en file d'attente beaucoup d'E/S dans un délai très court qui doit être réessayé. Cela peut prolonger le temps nécessaire pour reprendre les E/S.
- Une interface réseau prenant en charge les protocoles SAN ne peut pas être mise à niveau vers un nouvel emplacement. Un système d'exploitation hôte doit modifier le ou les chemins utilisés. La pause des E/S observée par l'hôte dépend de plusieurs facteurs. Du point de vue du système de stockage, la période pendant laquelle les E/S ne peuvent pas être servies ne prend que quelques secondes. Cependant, des systèmes d'exploitation hôtes différents peuvent nécessiter plus de temps pour permettre à une E/S de se déconnecter avant de réessayer. Les systèmes d'exploitation les plus récents sont mieux à même de reconnaître un changement de chemin beaucoup plus rapidement, mais les systèmes d'exploitation plus

anciens nécessitent généralement jusqu'à 30 secondes pour reconnaître un changement.

Les délais de basculement attendus lors desquels le système de stockage ne peut pas transmettre de données à un environnement applicatif sont indiqués dans le tableau ci-dessous. Aucun environnement applicatif ne doit contenir d'erreurs ; le basculement doit alors apparaître sous forme de courte pause dans le traitement des E/S.

	NFS	AFF	ASA
Basculement planifié	15 s	6-10 s	2-3 s
Basculement non planifié	30 s	6-10 s	2-3 s

Checksums et intégrité des données

ONTAP et les protocoles qu'il prend en charge incluent de nombreuses fonctionnalités qui protègent l'intégrité des bases de données Oracle, notamment les données au repos et les données transmises sur le réseau.

La protection logique des données dans ONTAP comprend trois exigences clés :

- Les données doivent être protégées contre la corruption.
- Les données doivent être protégées contre les pannes disques.
- Les modifications de données doivent être protégées contre la perte.

Ces trois besoins sont abordés dans les sections suivantes.

Corruption du réseau : checksums

Le niveau de protection de données le plus élémentaire est la somme de contrôle, qui est un code spécial de détection d'erreur stocké avec les données. La corruption des données lors de la transmission du réseau est détectée grâce à l'utilisation d'un checksum et, dans certains cas, de multiples checksums.

Par exemple, une trame FC inclut une forme de somme de contrôle appelée contrôle de redondance cyclique (CRC) pour s'assurer que la charge utile n'est pas corrompue en transit. L'émetteur envoie les données et le CRC des données. Le récepteur d'une trame FC recalcule le CRC des données reçues pour s'assurer qu'il correspond au CRC transmis. Si le nouveau CRC calculé ne correspond pas au CRC joint à la trame, les données sont corrompues et la trame FC est supprimée ou rejetée. Une opération d'E/S iSCSI comprend des checksums au niveau des couches TCP/IP et Ethernet. Pour une protection supplémentaire, elle peut également inclure la protection CRC facultative au niveau de la couche SCSI. Toute corruption de bit sur le fil est détectée par la couche TCP ou la couche IP, ce qui entraîne la retransmission du paquet. Comme avec FC, les erreurs dans le CRC SCSI entraînent une suppression ou un rejet de l'opération.

Corruption de disque : checksums

Des checksums sont également utilisés pour vérifier l'intégrité des données stockées sur les disques. Les blocs de données écrits sur les disques sont stockés avec une fonction de checksum qui génère un nombre imprévisible lié aux données d'origine. Lorsque les données sont lues à partir du lecteur, la somme de contrôle est recalculée et comparée à la somme de contrôle stockée. Si elle ne correspond pas, les données sont corrompues et doivent être restaurées par la couche RAID.

Corruption des données : écritures perdues

L'un des types de corruption les plus difficiles à détecter est une écriture perdue ou mal placée. Lorsqu'une écriture est reconnue, elle doit être écrite sur le support à l'emplacement correct. La corruption des données sur place est relativement facile à détecter à l'aide d'une simple somme de contrôle stockée avec les données. Cependant, si l'écriture est simplement perdue, alors la version précédente des données peut toujours exister et le total de contrôle serait correct. Si l'écriture est placée au mauvais emplacement physique, la somme de contrôle associée sera à nouveau valide pour les données stockées, même si l'écriture a détruit d'autres données.

La solution à ce défi est la suivante :

- Une opération d'écriture doit inclure des métadonnées indiquant l'emplacement où l'écriture est attendue.
- Une opération d'écriture doit inclure une sorte d'identifiant de version.

Lorsque ONTAP écrit un bloc, il inclut les données à l'emplacement où ce bloc appartient. Si une lecture ultérieure identifie un bloc, mais que les métadonnées indiquent qu'il appartient à l'emplacement 123 lorsqu'il a été trouvé à l'emplacement 456, l'écriture a été déplacée.

Il est plus difficile de détecter une écriture entièrement perdue. L'explication est très complexe, mais ONTAP stocke les métadonnées de façon à ce qu'une opération d'écriture entraîne des mises à jour vers deux emplacements différents sur les disques. En cas de perte d'une écriture, une lecture ultérieure des données et des métadonnées associées affiche deux identités de version différentes. Cela indique que l'écriture n'a pas été effectuée par le lecteur.

La corruption des écritures perdues ou déplacées est extrêmement rare. Cependant, avec la croissance continue des disques et l'expansion des jeux de données en exaoctets, le risque augmente. La détection des pertes en écriture doit être incluse dans tout système de stockage prenant en charge les charges de travail de la base de données.

Panne de disque : RAID, RAID DP et RAID-TEC

Si un bloc de données sur un disque est détecté comme étant corrompu, ou si l'ensemble du disque tombe en panne et est totalement indisponible, les données doivent être reconstituées. Cette opération est réalisée dans ONTAP à l'aide de disques de parité. Les données sont réparties sur plusieurs disques, puis des données de parité sont générées. Ces données sont stockées séparément des données d'origine.

ONTAP utilisait à l'origine RAID 4, qui utilise un seul lecteur de parité pour chaque groupe de lecteurs de données. Le résultat a été qu'un disque du groupe pouvait tomber en panne sans entraîner de perte de données. En cas de panne du disque de parité, aucune donnée n'a été endommagée et un nouveau disque de parité a pu être construit. En cas de panne d'un seul lecteur de données, les lecteurs restants peuvent être utilisés avec le lecteur de parité pour régénérer les données manquantes.

Lorsque les disques étaient petits, le risque statistique de défaillance simultanée de deux disques était négligeable. Avec l'augmentation des capacités des disques, la reconstruction des données suite à une panne disque s'est également accompagnée d'un temps considérable. Cela a augmenté la fenêtre au cours de laquelle une panne de second disque entraînerait la perte de données. De plus, le processus de reconstruction crée une grande quantité d'E/S supplémentaires sur les disques survivants. Au fur et à mesure du vieillissement des disques, le risque d'une charge supplémentaire entraînant une panne de second disque augmente également. Enfin, même si le risque de perte de données n'augmente pas avec l'utilisation continue de RAID 4, les conséquences de la perte de données deviendront plus graves. Plus la perte de données en cas de panne d'un groupe RAID est importante, plus la restauration des données est longue, ce qui entraîne une interruption de l'activité prolongée.

Ces problèmes ont conduit NetApp à développer la technologie NetApp RAID DP, une variante de RAID 6.

Cette solution comprend deux disques de parité, ce qui signifie que deux disques d'un groupe RAID peuvent tomber en panne sans générer de perte de données. Les disques ont continué de croître en taille, ce qui a conduit NetApp à développer la technologie NetApp RAID-TEC, qui introduit un troisième disque de parité.

Certaines meilleures pratiques en matière de bases de données historiques recommandent l'utilisation de RAID-10, également appelée mise en miroir par bandes. Cela offre une protection des données inférieure à celle de RAID DP, car il existe plusieurs scénarios de défaillance de deux disques, alors que dans RAID DP, il n'en existe aucune.

Par ailleurs, certaines bonnes pratiques en matière d'historique de bases de données indiquent que RAID-10 est préféré aux options RAID-4/5/6 en raison de problèmes de performances. Ces recommandations font parfois référence à une pénalité RAID. Bien que ces recommandations soient généralement correctes, elles ne s'appliquent pas aux implémentations de RAID dans ONTAP. Le problème de performances est lié à la régénération de parité. Dans les implémentations RAID traditionnelles, le traitement des écritures aléatoires de routine effectuées par une base de données nécessite plusieurs lectures de disque pour régénérer les données de parité et terminer l'écriture. La pénalité est définie comme les IOPS de lecture supplémentaires requises pour exécuter les opérations d'écriture.

ONTAP n'engendre pas de pénalité RAID, car les écritures sont placées dans la mémoire où la parité est générée, puis écrites sur le disque sous la forme d'une seule bande RAID. Aucune lecture n'est requise pour terminer l'opération d'écriture.

En résumé, par rapport à RAID 10, les systèmes RAID DP et RAID-TEC fournissent une capacité utilisable nettement plus importante, une meilleure protection contre les pannes disque et sans sacrifier les performances.

Protection contre les pannes matérielles : NVRAM

Toute baie de stockage servant de charge de travail de base de données doit traiter les opérations d'écriture le plus rapidement possible. En outre, une opération d'écriture doit être protégée contre la perte d'un événement inattendu tel qu'une coupure de courant. Cela signifie que toute opération d'écriture doit être stockée en toute sécurité dans au moins deux emplacements.

Les systèmes AFF et FAS utilisent la mémoire NVRAM pour répondre à ces exigences. Le processus d'écriture fonctionne comme suit :

1. Les données d'écriture entrantes sont stockées dans la mémoire RAM.
2. Les modifications à apporter aux données du disque sont journalisées dans la mémoire NVRAM sur le nœud local et le nœud partenaire. La mémoire NVRAM n'est pas un cache d'écriture. Il s'agit plutôt d'un journal similaire à un redo log de base de données. Dans des conditions normales, il n'est pas lu. Il est utilisé uniquement pour la restauration, par exemple après une coupure de courant pendant le traitement des E/S.
3. L'écriture est alors validée par l'hôte.

À ce stade, le processus d'écriture est complet du point de vue de l'application. Les données sont protégées contre les pertes, car elles sont stockées dans deux emplacements différents. Finalement, les modifications sont écrites sur le disque, mais ce processus est hors bande du point de vue de l'application, car il se produit après l'acquittement de l'écriture et n'affecte donc pas la latence. Ce processus est une fois de plus similaire à la journalisation de la base de données. Une modification de la base de données est enregistrée dans les journaux de reprise aussi rapidement que possible, et la modification est alors reconnue comme validée. Les mises à jour des fichiers de données sont effectuées beaucoup plus tard et n'affectent pas directement la vitesse de traitement.

En cas de panne de contrôleur, le contrôleur partenaire prend possession des disques requis et lit à nouveau

les données consignées dans la mémoire NVRAM pour récupérer toutes les opérations d'E/S en cours de fonctionnement au moment de la défaillance.

Protection contre les défaillances matérielles : 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 au partenaire de haute disponibilité, ces données à la volée 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 de haute disponibilité. Aucune donnée n'est perdue parce que le contrôleur qui connaît la défaillance n'a pas acquitté l'opération d'écriture.

ONTAP n'autorise pas le basculement lorsque les données sont désynchronisées, sauf si 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 en cas de basculement forcé, car elles conservent de grands caches internes de données sur disque. En cas 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.

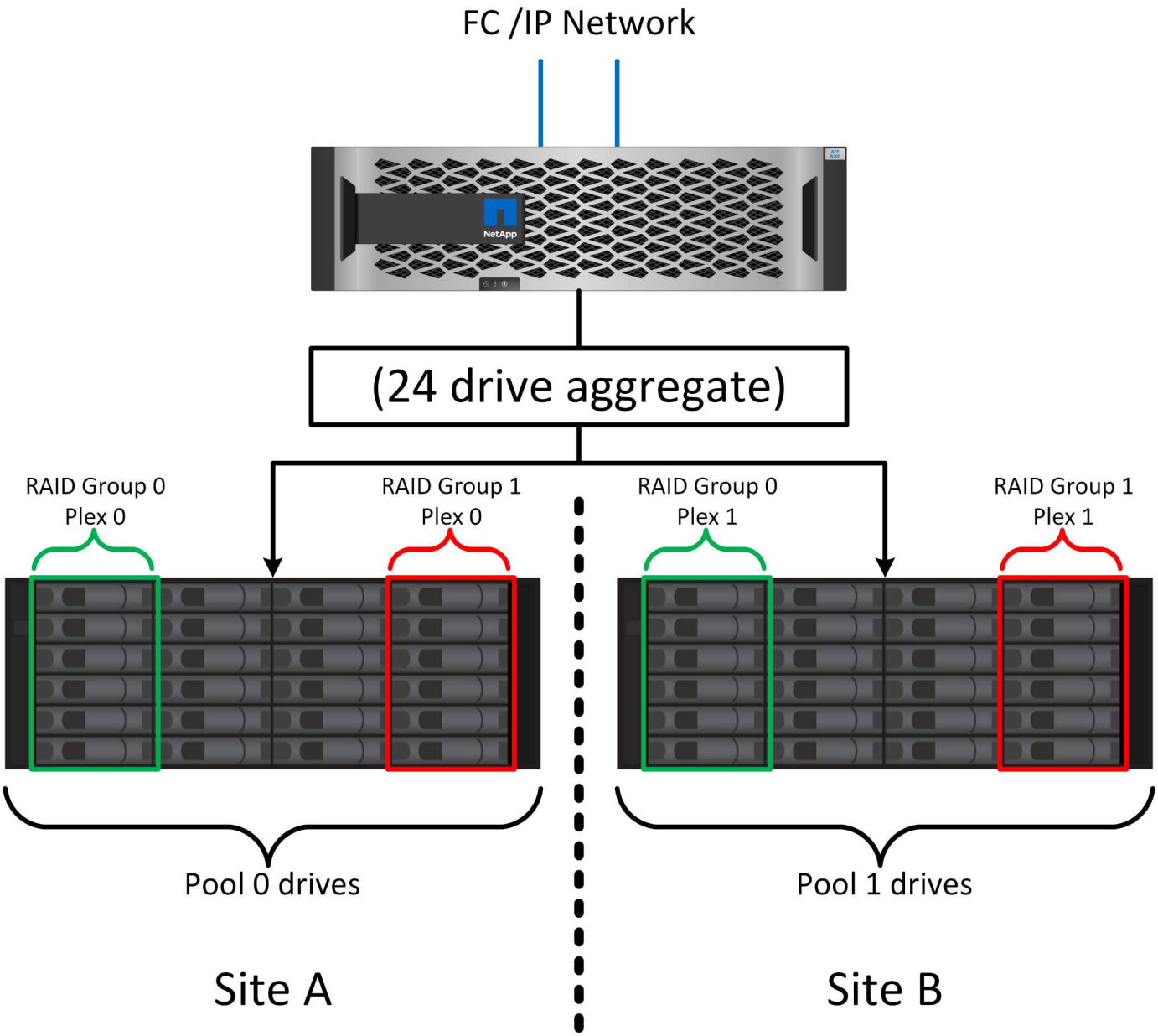
Afin de protéger les données de cette situation, ONTAP permet de configurer les volumes pour une protection spéciale contre les défaillances de mémoire NVRAM. Lorsqu'il est déclenché, ce mécanisme de protection entraîne l'entrée d'un volume dans un état appelé NVFAIL. Cet état entraîne des erreurs d'E/S qui entraînent l'arrêt d'une application et n'utilisent donc pas de données obsolètes. Les données ne doivent pas être perdues car une écriture reconnue doit être présente sur la matrice de stockage.

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.

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 :

- 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.
- 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.
- 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. Les données sont ainsi protégées en permanence, 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, RAID Group 1 inclut 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 alimentés par des alimentations indépendantes provenant de systèmes UPS indépendants.

Checksums

Le thème des checksums est particulièrement intéressant pour les administrateurs de bases de données habitués à l'utilisation de sauvegardes en continu Oracle RMAN qui migrent vers des sauvegardes basées sur des snapshots. RMAN permet notamment de procéder à des contrôles d'intégrité lors des opérations de sauvegarde. Bien que cette fonctionnalité présente un certain intérêt, son principal avantage est une base de données qui n'est pas utilisée sur une baie de stockage moderne. Lorsque des disques physiques sont utilisés pour une base de données Oracle, il est presque certain que la corruption finit par se produire lorsque les disques vieillissent, un problème qui est résolu par les checksums basés sur les baies dans les baies de stockage réelles.

Avec une baie de stockage réelle, l'intégrité des données est protégée par des checksums à plusieurs niveaux. Si les données sont corrompues dans un réseau IP, la couche TCP (transmission Control Protocol) rejette les données de paquets et demande la retransmission. Le protocole FC inclut des checksums, tout comme les données SCSI encapsulées. Une fois sur la matrice, ONTAP dispose d'une protection RAID et checksum. Une corruption peut se produire, mais, comme dans la plupart des baies d'entreprise, elle est détectée et corrigée. En général, un disque entier tombe en panne, ce qui invite à une reconstruction RAID et l'intégrité de la base de données n'est pas affectée. Il est toujours possible que des octets individuels sur un disque soient endommagés par le rayonnement cosmique ou par des cellules flash défectueuses. Si cela se produit, la vérification de parité échoue, le disque est mis hors service et la reconstruction RAID démarre. Là encore, l'intégrité des données n'est pas affectée. La dernière ligne de défense est l'utilisation de checksums. Si, par exemple, une erreur de micrologiciel catastrophique sur un disque a corrompu des données d'une manière qui n'a pas été détectée par un contrôle de parité RAID, le checksum ne correspond pas et ONTAP empêche le transfert d'un bloc corrompu avant que la base de données Oracle puisse le recevoir.

L'architecture des fichiers de données et des redo log Oracle est également conçue pour offrir le plus haut niveau possible d'intégrité des données, même dans des circonstances extrêmes. Au niveau le plus élémentaire, les blocs Oracle incluent un checksum et des contrôles logiques de base avec presque toutes les E/S. Si Oracle ne s'est pas écrasé ou n'a pas mis un tablespace hors ligne, les données sont intactes. Le degré de vérification de l'intégrité des données est réglable et Oracle peut également être configuré pour confirmer les écritures. Par conséquent, la quasi-totalité des scénarios de panne et de panne peuvent être restaurés, et dans le cas extrêmement rare d'une situation irrécupérable, la corruption est rapidement détectée.

La plupart des clients NetApp qui utilisent des bases de données Oracle cessent d'utiliser RMAN et d'autres produits de sauvegarde après la migration vers des sauvegardes snapshot. Il existe encore des options permettant d'utiliser RMAN pour effectuer une restauration au niveau des blocs avec SnapCenter. Toutefois, au quotidien, RMAN, NetBackup et d'autres produits ne sont utilisés qu'occasionnellement pour créer des copies d'archivage mensuelles ou trimestrielles.

Certains clients choisissent d'exécuter `dbv` périodiquement pour effectuer des contrôles d'intégrité sur leurs bases de données existantes. NetApp déconseille cette pratique, car elle entraîne une charge d'E/S inutile. Comme indiqué ci-dessus, si la base de données ne rencontrait pas de problèmes auparavant, le risque de `dbv` La détection d'un problème est proche de zéro et cet utilitaire entraîne une charge d'E/S séquentielles très élevée sur le réseau et le système de stockage. À moins qu'il n'y ait de raison de croire qu'il existe une corruption, comme l'exposition à un bogue connu d'Oracle, il n'y a aucune raison de s'exécuter `dbv`.

Notions de base sur la sauvegarde et la restauration

Sauvegardes basées sur des snapshots

La technologie Snapshot de NetApp constitue le socle de la protection des données des bases de données Oracle sur ONTAP.

Les valeurs clés sont les suivantes :

- **Simplicité.** Un instantané est une copie en lecture seule du contenu d'un conteneur de données à un moment donné.
- **Efficacité.** les instantanés ne nécessitent pas d'espace au moment de la création. L'espace n'est consommé que lorsque des données sont modifiées.
- **Gérabilité.** Une stratégie de sauvegarde basée sur les snapshots est facile à configurer et à gérer car les snapshots font partie intégrante du système d'exploitation du stockage. Si le système de stockage est sous tension, il est prêt à créer des sauvegardes.
- **Évolutivité.** vous pouvez conserver jusqu'à 1024 sauvegardes d'un seul conteneur de fichiers et de LUN. Dans le cas de jeux de données complexes, plusieurs conteneurs de données peuvent être protégés par un ensemble unique et cohérent de snapshots.
- Les performances ne sont pas affectées, qu'un volume contienne ou non 1024 snapshots.

Bien que de nombreux fournisseurs de stockage proposent la technologie Snapshot, la technologie Snapshot de ONTAP est unique et offre des avantages significatifs pour les environnements applicatifs et de bases de données d'entreprise :

- Les copies Snapshot font partie de la WAFL (Write-Anywhere File Layout) sous-jacente. Il ne s'agit pas d'une technologie complémentaire ou externe. La gestion est donc simplifiée, car le système de stockage est le système de sauvegarde.
- Les copies Snapshot n'affectent pas les performances, sauf dans certains cas en périphérie, par exemple lorsque le volume de données est stocké dans des snapshots que le système de stockage sous-jacent se remplit.
- Le terme « groupe de cohérence » fait souvent référence à un regroupement d'objets de stockage gérés comme un ensemble cohérent de données. La copie Snapshot d'un volume ONTAP donné constitue une sauvegarde de groupe de cohérence.

Les copies Snapshot ONTAP ont également une meilleure évolutivité que la technologie concurrente. Les clients peuvent stocker 5, 50 ou 500 copies Snapshot sans affecter les performances. Le nombre maximal de snapshots actuellement autorisés dans un volume est de 1024. Si une conservation supplémentaire des snapshots est nécessaire, il existe des options pour les transmettre en cascade à des volumes supplémentaires.

Par conséquent, la protection d'un dataset hébergé sur ONTAP est simple et hautement évolutive. Les sauvegardes ne nécessitent pas de déplacement de données. Par conséquent, une stratégie de sauvegarde peut être adaptée aux besoins de l'entreprise plutôt qu'aux limites des taux de transfert réseau, du grand nombre de lecteurs de bande ou des zones de transfert de disque.

Un snapshot est-il une sauvegarde ?

La question couramment posée sur l'utilisation des snapshots en tant que stratégie de protection des données est le fait que les données « réelles » et les données de snapshot se trouvent sur les mêmes disques. La perte de ces disques entraînerait la perte des données primaires et de la sauvegarde.

Ce problème est valide. Les snapshots locaux sont utilisés pour les besoins quotidiens de sauvegarde et de restauration, et dans ce sens, le snapshot est une sauvegarde. Dans les environnements NetApp, près de 99 % des scénarios de restauration s'appuient sur des copies Snapshot pour répondre aux exigences de RTO les plus strictes.

Toutefois, les snapshots locaux ne doivent jamais être la seule stratégie de sauvegarde. C'est pourquoi NetApp propose des technologies telles que la réplication SnapMirror et SnapVault pour répliquer rapidement

et efficacement des copies Snapshot sur un ensemble indépendant de disques. Dans une solution bien conçue avec des snapshots et une réplication Snapshot, l'utilisation des bandes peut être réduite au minimum, voire même à une archive trimestrielle, ou totalement éliminée.

Sauvegardes basées sur des snapshots

Vous pouvez utiliser les copies Snapshot ONTAP pour protéger vos données, et les copies Snapshot sont la base de nombreuses autres fonctionnalités ONTAP, notamment la réplication, la reprise d'activité et le clonage. Une description complète de la technologie Snapshot ne fait pas partie du présent document, mais les sections suivantes offrent un aperçu général.

Il existe deux approches principales pour créer un snapshot d'un dataset :

- Sauvegardes cohérentes après panne
- Sauvegardes cohérentes au niveau des applications

Une sauvegarde cohérente après panne d'un dataset fait référence à la capture de l'ensemble de la structure du dataset à un point dans le temps. Si le dataset est stocké dans un seul volume, le processus est simple ; il est possible de créer une copie Snapshot à tout moment. Si un dataset s'étend sur plusieurs volumes, un snapshot de groupe de cohérence doit être créé. Plusieurs options sont disponibles pour la création des snapshots de groupe de cohérence, notamment le logiciel NetApp SnapCenter, les fonctionnalités natives de groupe de cohérence ONTAP et les scripts gérés par l'utilisateur.

Les sauvegardes cohérentes après panne sont principalement utilisées lorsque la restauration au point de sauvegarde est suffisante. Lorsqu'une restauration plus granulaire est nécessaire, des sauvegardes cohérentes au niveau des applications sont généralement nécessaires.

Le mot "cohérent" dans "application-cohérente" est souvent un mal nommer. Par exemple, le placement d'une base de données Oracle en mode de sauvegarde est appelé sauvegarde cohérente au niveau des applications, mais les données ne sont en aucun cas rendues cohérentes ou suspendues. Les données continuent de changer tout au long de la sauvegarde. En revanche, la plupart des sauvegardes MySQL et Microsoft SQL Server ont effectivement mis les données au repos avant d'exécuter la sauvegarde. VMware peut rendre certains fichiers cohérents ou non.

Groupes de cohérence

Le terme « groupe de cohérence » fait référence à la capacité d'une baie de stockage à gérer plusieurs ressources de stockage comme une seule image. Par exemple, une base de données peut comprendre 10 LUN. La baie doit pouvoir sauvegarder, restaurer et répliquer ces 10 LUN de manière cohérente. La restauration n'est pas possible si les images des LUN n'étaient pas cohérentes au point de sauvegarde. La réplication de ces 10 LUN nécessite que tous les réplicas soient parfaitement synchronisés.

Le terme « groupe de cohérence » n'est pas souvent utilisé lors des discussions sur ONTAP, car la cohérence a toujours été une fonction de base de l'architecture de volumes et d'agrégats au sein de ONTAP. De nombreuses autres baies de stockage gèrent des LUN ou des systèmes de fichiers en tant qu'unités individuelles. Ils peuvent ensuite être configurés en tant que « groupe de cohérence » pour la protection des données, mais cette étape supplémentaire est nécessaire dans la configuration.

ONTAP a toujours pu capturer des images locales et répliquées cohérentes de données. Bien que les différents volumes d'un système ONTAP ne soient généralement pas officiellement décrits comme des groupes de cohérence, c'est ce qu'ils sont. Une copie Snapshot de ce volume est une image de groupe de cohérence. La restauration de ce Snapshot correspond à une restauration de groupe de cohérence. SnapMirror et SnapVault proposent tous deux une réplication de groupe de cohérence.

Snapshots de groupes de cohérence

Les copies Snapshot de groupe de cohérence (cg-snapshots) sont une extension de la technologie Snapshot ONTAP de base. Une opération de snapshot standard crée une image cohérente de toutes les données d'un même volume, mais il est parfois nécessaire de créer un ensemble cohérent de snapshots sur plusieurs volumes et même sur plusieurs systèmes de stockage. Il en résulte un ensemble de snapshots qui peuvent être utilisés de la même manière qu'un snapshot d'un seul volume individuel. Elles peuvent être utilisées pour la restauration des données locales, répliquées à des fins de reprise après incident ou clonées sous la forme d'une unité cohérente unique.

L'utilisation la plus connue des cg-snapshots concerne un environnement de base de données d'environ 1 po de capacité couvrant 12 contrôleurs. Les snapshots de groupe de cohérence créés sur ce système ont été utilisés pour la sauvegarde, la restauration et le clonage.

La plupart du temps, lorsqu'un dataset s'étend sur des volumes et que l'ordre d'écriture doit être préservé, le logiciel de gestion choisi utilise automatiquement un snapshot de groupe de cohérence. Dans ce cas, il n'est pas nécessaire de comprendre les détails techniques des cg-snapshots. Toutefois, les exigences complexes en matière de protection des données nécessitent un contrôle détaillé du processus de protection et de réplication des données. Certains workflows d'automatisation ou scripts personnalisés permettent d'appeler les API cg-Snapshot. Pour comprendre la meilleure option et le rôle de cg-snapshot, vous devez fournir une explication plus détaillée de la technologie.

La création d'un ensemble de snapshots des groupes de cohérence s'effectue en deux étapes :

1. Établir une clôture d'écriture sur tous les volumes cibles.
2. Créez des instantanés de ces volumes à l'état clôturé.

L'escrime d'écriture est établi en série. Cela signifie que lorsque le processus de recel est configuré sur plusieurs volumes, les E/S d'écriture sont bloquées sur le premier volume de la séquence au fur et à mesure qu'elles continuent d'être validées sur les volumes qui apparaissent plus tard. Cela peut sembler initialement contraire à l'exigence de préservation de l'ordre d'écriture, mais cela s'applique uniquement aux E/S émises de manière asynchrone sur l'hôte et ne dépend pas d'autres écritures.

Par exemple, une base de données peut émettre de nombreuses mises à jour asynchrones des fichiers de données et permettre au système d'exploitation de réorganiser les E/S et de les compléter selon sa propre configuration de planificateur. L'ordre de ce type d'E/S ne peut pas être garanti car l'application et le système d'exploitation ont déjà libéré l'obligation de conserver l'ordre d'écriture.

Par exemple, la plupart des activités de journalisation de la base de données sont synchrones. La base de données ne procède pas à d'autres écritures de journal tant que les E/S n'ont pas été acquittées et que l'ordre de ces écritures doit être conservé. Si une E/S de journal arrive sur un volume clôturé, elle n'est pas validée et l'application se bloque lors d'écritures ultérieures. De même, les E/S des métadonnées du système de fichiers sont généralement synchrones. Par exemple, une opération de suppression de fichier ne doit pas être perdue. Si un système d'exploitation doté d'un système de fichiers xfs supprime un fichier et que les E/S qui ont mis à jour les métadonnées du système de fichiers xfs pour supprimer la référence à ce fichier ont été reçues sur un volume isolé, l'activité du système de fichiers est alors interrompue. Cela garantit l'intégrité du système de fichiers pendant les opérations cg-Snapshot.

Une fois l'isolation d'écriture configurée sur les volumes cibles, ils sont prêts pour la création d'instantanés. Les snapshots n'ont pas besoin d'être créés précisément en même temps, car l'état des volumes est figé du point de vue de l'écriture dépendant. Pour éviter toute faille dans l'application qui crée les instantanés cg, l'escrime d'écriture initiale inclut un délai configurable dans lequel ONTAP libère automatiquement l'escrime et reprend le traitement d'écriture après un nombre défini de secondes. Si tous les snapshots sont créés avant l'expiration du délai, le jeu de snapshots résultant est un groupe de cohérence valide.

Ordre d'écriture dépendant

Du point de vue technique, la préservation de l'ordre d'écriture et, plus particulièrement, de l'ordre d'écriture dépendant constitue la clé d'un groupe de cohérence. Par exemple, une base de données qui écrit 10 LUN écrit simultanément sur toutes ces LUN. De nombreuses écritures sont émises de manière asynchrone, ce qui signifie que l'ordre dans lequel elles sont effectuées n'est pas important et que l'ordre dans lequel elles sont effectuées varie en fonction du système d'exploitation et du comportement du réseau.

Certaines opérations d'écriture doivent être présentes sur le disque avant que la base de données puisse procéder à des écritures supplémentaires. Ces opérations d'écriture critiques sont appelées écritures dépendantes. Les E/S d'écriture suivantes dépendent de la présence de ces écritures sur le disque. Tout snapshot, restauration ou réplication de ces 10 LUN doit garantir l'ordre d'écriture dépendant. Les mises à jour du système de fichiers sont un autre exemple d'écritures dépendantes de l'ordre d'écriture. L'ordre dans lequel les modifications du système de fichiers sont effectuées doit être conservé, sinon l'ensemble du système de fichiers pourrait être corrompu.

Stratégies

Il existe deux approches principales des sauvegardes basées sur des snapshots :

- Sauvegardes cohérentes après panne
- Sauvegardes à chaud protégées pour les snapshots

Une sauvegarde cohérente après panne d'une base de données fait référence à la capture à un moment précis de l'ensemble de la structure de la base de données, y compris les fichiers de données, les journaux de reprise et les fichiers de contrôle. Si la base de données est stockée sur un seul volume, le processus est simple ; il est possible de créer un Snapshot à tout moment. Si la base de données s'étend sur plusieurs volumes, un snapshot de groupe de cohérence doit être créé. Plusieurs options sont disponibles pour la création des snapshots de groupe de cohérence, notamment le logiciel NetApp SnapCenter, les fonctionnalités natives de groupe de cohérence ONTAP et les scripts gérés par l'utilisateur.

Les sauvegardes Snapshot cohérentes après panne sont principalement utilisées lorsque la restauration au point de sauvegarde est suffisante. Les journaux d'archivage peuvent être appliqués dans certains cas, mais lorsqu'une restauration granulaire à un point dans le temps est nécessaire, il est préférable d'effectuer une sauvegarde en ligne.

La procédure de base pour une sauvegarde en ligne basée sur un snapshot est la suivante :

1. Placez la base de données dans `backup` mode.
2. Créez un Snapshot de tous les volumes qui hébergent les fichiers de données.
3. Quitter `backup` mode.
4. Lancer la commande `alter system archive log current` pour forcer l'archivage des journaux.
5. Créer des instantanés de tous les volumes hébergeant les journaux d'archivage.

Cette procédure permet d'obtenir un ensemble de snapshots contenant les fichiers de données en mode de sauvegarde et les journaux d'archivage critiques générés en mode de sauvegarde. Il s'agit des deux conditions requises pour restaurer une base de données. Il est également conseillé de protéger les fichiers tels que les fichiers de contrôle, mais la seule condition absolue est la protection des fichiers de données et des journaux d'archivage.

Même si différents clients peuvent avoir des stratégies très différentes, la quasi-totalité de ces stratégies s'appuient sur les mêmes principes que ceux décrits ci-dessous.

Restauration basée sur des snapshots

Lors de la conception d'infrastructures de volumes pour les bases de données Oracle, la première décision est d'utiliser ou non la technologie VBSR (Volume-Based NetApp SnapRestore).

La fonction SnapRestore basée sur les volumes permet de rétablir quasi instantanément un volume à un point antérieur. Toutes les données du volume étant rétablies, VBSR peut ne pas convenir à toutes les utilisations. Par exemple, si l'intégralité d'une base de données, y compris les fichiers de données, les journaux de reprise et les journaux d'archivage, est stockée sur un seul volume restauré avec VBSR, les données sont perdues, car les nouveaux journaux d'archivage et les données de reprise sont supprimés.

La technologie VBSR n'est pas requise pour la restauration. De nombreuses bases de données peuvent être restaurées avec SFSR (Single File SnapRestore) ou en copiant simplement les fichiers du snapshot vers le système de fichiers actif.

La technologie VBSR est recommandée pour les bases de données très volumineuses ou si une restauration doit être effectuée le plus rapidement possible et que l'utilisation de VBSR nécessite l'isolement des fichiers de données. Dans un environnement NFS, les fichiers de données d'une base de données doivent être stockés sur des volumes dédiés non endommagés par d'autres types de fichiers. Dans un environnement SAN, les fichiers de données doivent être stockés sur des LUN dédiés sur des volumes dédiés. Si un gestionnaire de volumes est utilisé (y compris Oracle Automatic Storage Management (ASM)), le groupe de disques doit également être dédié aux fichiers de données.

Cette méthode d'isolement des fichiers de données permet de rétablir leur état antérieur sans endommager d'autres systèmes de fichiers.

Réserve Snapshot

Pour chaque volume contenant des données Oracle dans un environnement SAN, le `percent-snapshot-space` Doit être défini sur zéro car il n'est pas utile de réserver de l'espace pour un snapshot dans un environnement LUN. Si la réserve fractionnaire est définie sur 100, un snapshot d'un volume avec des LUN nécessite suffisamment d'espace libre dans le volume, à l'exception de la réserve Snapshot, pour absorber 100 % de CA de toutes les données. Si la réserve fractionnaire est définie sur une valeur inférieure, une quantité d'espace libre correspondante est nécessaire, mais elle exclut toujours la réserve snapshot. Cela signifie que l'espace de réserve du snapshot dans un environnement de LUN est gaspillé.

Dans un environnement NFS, deux options sont possibles :

- Réglez le `percent-snapshot-space` basé sur la consommation d'espace prévue du snapshot.
- Réglez le `percent-snapshot-space` pour zéro et gérer collectivement l'espace utilisé actif et snapshot.

Avec la première option, `percent-snapshot-space` est défini sur une valeur différente de zéro, généralement autour de 20 %. Cet espace est alors masqué par l'utilisateur. Toutefois, cette valeur ne crée pas de limite d'utilisation. Si une base de données avec une réservation de 20 % connaît un chiffre d'affaires de 30 %, l'espace snapshot peut dépasser les limites de la réserve de 20 % et occuper un espace non réservé.

Le principal avantage de la définition d'une réserve sur une valeur telle que 20 % est de vérifier qu'un peu d'espace est toujours disponible pour les snapshots. Par exemple, un volume de 1 To avec une réserve de 20 % permettrait uniquement à un administrateur de base de données (DBA) de stocker 800 Go de données. Cette configuration garantit au moins 200 Go d'espace pour la consommation de snapshots.

Quand `percent-snapshot-space` est défini sur zéro, tout l'espace du volume est disponible pour l'utilisateur final, ce qui offre une meilleure visibilité. L'administrateur de base de données doit comprendre que,

s'il constate qu'un volume de 1 To exploite les snapshots, cet espace de 1 To est partagé entre les données actives et le renouvellement du Snapshot.

Il n'existe pas de préférence claire entre l'option 1 et l'option 2 parmi les utilisateurs finaux.

ONTAP et snapshots tiers

Oracle Doc ID 604683.1 décrit les conditions requises pour la prise en charge des snapshots tiers et les nombreuses options disponibles pour les opérations de sauvegarde et de restauration.

Les fournisseurs tiers doivent garantir la conformité de leurs snapshots à plusieurs exigences :

- Les snapshots doivent intégrer les opérations de restauration et de reprise recommandées par Oracle.
- Les snapshots doivent être cohérents après panne de la base de données au point du Snapshot.
- L'ordre d'écriture est conservé pour chaque fichier d'un snapshot.

Les produits de gestion Oracle de ONTAP et NetApp sont conformes à ces exigences.

SnapRestore

La technologie NetApp SnapRestore assure la restauration rapide des données dans ONTAP à partir d'une copie Snapshot.

Lorsqu'un dataset stratégique n'est pas disponible, les opérations stratégiques de l'entreprise ne sont pas disponibles. Les bandes peuvent se rompre, et même les restaurations à partir de sauvegardes sur disque peuvent être lentes à transférer sur le réseau. SnapRestore évite ces problèmes en offrant une restauration quasi instantanée des datasets. Même les bases de données de plusieurs pétaoctets peuvent être entièrement restaurées en quelques minutes à peine.

Il existe deux types d'SnapRestore : basés sur les fichiers/LUN et sur les volumes.

- Il est possible de restaurer des fichiers individuels ou des LUN en quelques secondes, qu'il s'agisse d'un LUN de 2 To ou d'un fichier de 4 Ko.
- Le conteneur de fichiers ou de LUN peut être restauré en quelques secondes, qu'il s'agisse de 10 Go ou 100 To de données.

Un « conteneur de fichiers ou de LUN » fait généralement référence à un volume FlexVol. Par exemple, vous pouvez avoir 10 LUN qui composent un groupe de disques LVM dans un seul volume, ou un volume peut stocker les home directories NFS de 1000 utilisateurs. Au lieu d'exécuter une opération de restauration pour chaque fichier ou LUN individuel, vous pouvez restaurer le volume entier en une seule opération. Ce processus fonctionne également avec des conteneurs scale-out qui incluent plusieurs volumes, tels qu'un FlexGroup ou un groupe de cohérence ONTAP.

La rapidité et l'efficacité de SnapRestore sont dues à la nature d'une copie Snapshot, qui offre essentiellement une vue en lecture seule parallèle du contenu d'un volume à un moment donné. Les blocs actifs sont les blocs réels qui peuvent être modifiés, tandis que le snapshot offre une vue en lecture seule de l'état des blocs qui constituent les fichiers et les LUN au moment de la création du snapshot.

ONTAP permet uniquement un accès en lecture seule aux données instantanées, mais les données peuvent être réactivées avec SnapRestore. L'instantané est réactivé en tant que vue en lecture-écriture des données, renvoyant les données à leur état précédent. SnapRestore peut fonctionner au niveau du volume ou du fichier. La technologie est essentiellement la même avec quelques différences mineures de comportement.

SnapRestore du volume

La fonction SnapRestore basée sur les volumes renvoie la totalité du volume de données à un état antérieur. Cette opération ne nécessite pas de déplacement de données. Le processus de restauration est donc pratiquement instantané, bien que le traitement des opérations via l'API ou l'interface de ligne de commande puisse prendre quelques secondes. La restauration de 1 Go de données n'est pas plus compliquée et chronophage que la restauration de 1 po de données. Cette fonctionnalité est la principale raison pour laquelle de nombreux clients grands comptes migrent vers des systèmes de stockage ONTAP. Il assure un RTO se mesure en quelques secondes, même pour les datasets les plus volumineux.

L'un des inconvénients des SnapRestore sur volume est le fait que les modifications au sein d'un volume sont cumulées dans le temps. Par conséquent, chaque snapshot et les données de fichier actives dépendent des modifications apportées jusqu'à ce point. Le rétablissement d'un volume à un état antérieur implique la suppression de toutes les modifications ultérieures apportées aux données. Ce qui est moins évident, cependant, c'est qu'il s'agit d'instantanés créés par la suite. Ce n'est pas toujours souhaitable.

Par exemple, un SLA de conservation des données peut spécifier 30 jours de sauvegardes nocturnes. La restauration d'un dataset sur un snapshot créé il y a cinq jours avec SnapRestore du volume abandonnerait tous les snapshots créés les cinq jours précédents, en violation du SLA.

Un certain nombre d'options sont disponibles pour résoudre cette limitation :

1. Les données peuvent être copiées à partir d'un instantané précédent, au lieu d'effectuer une SnapRestore du volume entier. Cette méthode fonctionne mieux avec les jeux de données plus petits.
2. Un snapshot peut être cloné plutôt que restauré. La limitation à cette approche est que le snapshot source dépend du clone. Par conséquent, elle ne peut pas être supprimée si le clone n'est pas également supprimé ou s'il est divisé en volume indépendant.
3. Utilisation d'un SnapRestore basé sur des fichiers.

Fichier SnapRestore

SnapRestore basé sur les fichiers est un processus de restauration plus granulaire basé sur des snapshots. Au lieu de rétablir l'état d'un volume entier, l'état d'un fichier ou d'une LUN individuel est rétabli. Il n'est pas nécessaire de supprimer des snapshots et cette opération ne crée aucune dépendance vis-à-vis d'un instantané précédent. Le fichier ou la LUN est immédiatement disponible dans le volume actif.

Aucun déplacement des données n'est nécessaire lors de la restauration d'un fichier ou d'une LUN par SnapRestore. Cependant, des mises à jour internes des métadonnées sont nécessaires pour refléter le fait que les blocs sous-jacents d'un fichier ou d'une LUN existent désormais à la fois dans un snapshot et dans le volume actif. Les performances ne doivent pas être affectées, mais ce processus bloque la création de snapshots jusqu'à ce qu'elle soit terminée. Le taux de traitement est d'environ 5 Gbit/s (18 To/heure) en fonction de la taille totale des fichiers restaurés.

Sauvegardes en ligne

Deux datasets sont nécessaires pour protéger et restaurer une base de données Oracle en mode de sauvegarde. Notez qu'il ne s'agit pas de la seule option de sauvegarde Oracle, mais qu'elle est la plus courante.

- Un Snapshot des fichiers de données en mode de sauvegarde
- Les journaux d'archivage créés pendant que les fichiers de données étaient en mode de sauvegarde

Si une récupération complète incluant toutes les transactions validées est requise, un troisième élément est

requis :

- Les journaux de reprise en cours

Il existe plusieurs façons de restaurer une sauvegarde en ligne. De nombreux clients restaurent les snapshots à l'aide de l'interface de ligne de commande ONTAP, puis à l'aide d'Oracle RMAN ou de sqlplus pour terminer la restauration. Cette approche est particulièrement fréquente dans les environnements de production de grande taille. En effet, la probabilité et la fréquence des restaurations de bases de données sont extrêmement faibles et les restaurations sont gérées par un administrateur de bases de données qualifié. Pour une automatisation totale, des solutions telles que NetApp SnapCenter intègrent un plug-in Oracle avec une ligne de commande et des interfaces graphiques.

Certains grands clients ont adopté une approche plus simple en configurant des scripts de base sur les hôtes afin de placer les bases de données en mode de sauvegarde à un moment spécifique en préparation d'un snapshot planifié. Par exemple, planifiez la commande `alter database begin backup` à 23:58, `alter database end backup` à 00:02, puis planifiez les snapshots directement sur le système de stockage à minuit. Résultat : une stratégie de sauvegarde simple et hautement évolutive ne nécessite aucun logiciel ni licence externe.

Disposition des données

La disposition la plus simple consiste à isoler les fichiers de données dans un ou plusieurs volumes dédiés. Ils doivent être non contaminés par tout autre type de fichier. Cela permet de s'assurer que les volumes de fichiers de données peuvent être rapidement restaurés via une opération SnapRestore sans détruire un journal de reprise, un fichier de contrôle ou un journal d'archivage important.

LE SYSTÈME SAN présente des exigences similaires en matière d'isolation des fichiers de données dans des volumes dédiés. Avec un système d'exploitation tel que Microsoft Windows, un seul volume peut contenir plusieurs LUN de fichiers de données, chacune avec un système de fichiers NTFS. Avec d'autres systèmes d'exploitation, il existe généralement un gestionnaire de volumes logiques. Par exemple, avec Oracle ASM, l'option la plus simple consiste à limiter les LUN d'un groupe de disques ASM à un seul volume pouvant être sauvegardé et restauré en tant qu'unité. Si des volumes supplémentaires sont nécessaires pour des raisons de performance ou de gestion de la capacité, la création d'un groupe de disques supplémentaire sur le nouveau volume simplifie la gestion.

Si ces instructions sont respectées, les snapshots peuvent être planifiés directement sur le système de stockage sans avoir à créer de snapshot de groupe de cohérence. En effet, les sauvegardes Oracle ne nécessitent pas la sauvegarde simultanée de fichiers de données. La procédure de sauvegarde en ligne a été conçue pour assurer la mise à jour des fichiers de données, qui seront ensuite transmis progressivement sur bande en quelques heures.

Une complication se produit dans des situations telles que l'utilisation d'un groupe de disques ASM distribué sur des volumes. Dans ce cas, un snapshot de groupe de cohérence doit être réalisé pour s'assurer que les métadonnées ASM sont cohérentes sur tous les volumes constitutifs.

Attention : Vérifiez que l'ASM `spfile` et `passwd` les fichiers ne se trouvent pas dans le groupe de disques hébergeant les fichiers de données. Cela interfère avec la capacité à restaurer de manière sélective les fichiers de données et uniquement les fichiers de données.

Procédure de restauration locale : NFS

Cette procédure peut être conduite manuellement ou via une application telle que SnapCenter. La procédure de base est la suivante :

1. Arrêtez la base de données.

2. Restaurez le ou les volumes de fichiers de données sur l'instantané immédiatement avant le point de restauration souhaité.
3. Réexécutez les journaux d'archivage au point souhaité.
4. Relire les journaux de reprise en cours si vous souhaitez effectuer une restauration complète.

Cette procédure suppose que les journaux d'archive souhaités sont toujours présents dans le système de fichiers actif. Si ce n'est pas le cas, les journaux d'archivage doivent être restaurés ou rman/sqlplus peut être dirigé vers les données du répertoire d'instantanés.

En outre, dans le cas de bases de données plus petites, l'utilisateur peut restaurer les fichiers de données directement à partir du système `.snapshot` répertoire n'ayant pas besoin des outils d'automatisation ou des administrateurs de stockage pour exécuter une `snaprestore` commande.

Procédure de restauration locale—SAN

Cette procédure peut être conduite manuellement ou via une application telle que SnapCenter. La procédure de base est la suivante :

1. Arrêtez la base de données.
2. Arrêter le ou les groupes de disques hébergeant les fichiers de données. La procédure varie en fonction du gestionnaire de volumes logiques choisi. Avec ASM, le processus nécessite de démonter le groupe de disques. Sous Linux, les systèmes de fichiers doivent être démontés et les volumes logiques et les groupes de volumes doivent être désactivés. L'objectif est d'arrêter toutes les mises à jour du groupe de volumes cible à restaurer.
3. Restaurez les groupes de disques de fichiers de données sur l'instantané immédiatement avant le point de restauration souhaité.
4. Réactivez les groupes de disques récemment restaurés.
5. Réexécutez les journaux d'archivage au point souhaité.
6. Relire tous les journaux de reprise si vous souhaitez procéder à une restauration complète.

Cette procédure suppose que les journaux d'archive souhaités sont toujours présents dans le système de fichiers actif. Si ce n'est pas le cas, les journaux d'archivage doivent être restaurés en mettant les LUN du journal d'archivage hors ligne et en effectuant une restauration. Il s'agit également d'un exemple dans lequel il est utile de diviser les journaux d'archivage en volumes dédiés. Si les journaux d'archivage partagent un groupe de volumes avec les journaux de reprise, les journaux de reprise doivent être copiés ailleurs avant la restauration de l'ensemble global des LUN. Cette étape empêche la perte de ces transactions finales enregistrées.

Sauvegardes optimisées pour les snapshots de stockage

La sauvegarde et la restauration basées sur des snapshots sont devenues encore plus simples au moment du lancement d'Oracle 12c. En effet, il n'est pas nécessaire de placer une base de données en mode de sauvegarde à chaud. Il est possible de planifier des sauvegardes Snapshot directement sur un système de stockage et d'effectuer des restaurations complètes ou à un point dans le temps.

Les administrateurs de bases de données maîtrisent mieux la procédure de restauration à partir d'une sauvegarde à chaud, mais il est depuis longtemps possible d'utiliser des snapshots qui n'ont pas été créés pendant que la base de données était en mode de sauvegarde à chaud. Pour assurer la cohérence de la base de données, des étapes manuelles supplémentaires ont été nécessaires avec Oracle 10g et 11g. Avec Oracle

12c, `sqlplus` et `rman` contiennent la logique supplémentaire permettant de relire les journaux d'archivage sur des sauvegardes de fichiers de données qui n'étaient pas en mode de sauvegarde à chaud.

Comme nous l'avons vu précédemment, la restauration d'une sauvegarde à chaud basée sur des snapshots nécessite deux jeux de données :

- Un Snapshot des fichiers de données créés en mode de sauvegarde
- Les journaux d'archivage générés pendant que les fichiers de données étaient en mode de sauvegarde à chaud

Lors de la restauration, la base de données lit les métadonnées à partir des fichiers de données pour sélectionner les journaux d'archivage requis à des fins de restauration.

La restauration optimisée pour les snapshots de stockage nécessite des jeux de données légèrement différents pour obtenir les mêmes résultats :

- Un Snapshot des fichiers de données et une méthode d'identification de l'heure de création du Snapshot
- Archiver les journaux à partir de l'heure du point de contrôle du fichier de données le plus récent jusqu'à l'heure exacte du snapshot

Lors de la restauration, la base de données lit les métadonnées à partir des fichiers de données pour identifier le premier journal d'archivage requis. Il est possible d'effectuer une restauration complète ou instantanée. Lors de l'exécution d'une restauration à un point dans le temps, il est essentiel d'connaître l'heure du Snapshot des fichiers de données. Le point de restauration spécifié doit être après l'heure de création des snapshots. NetApp recommande d'ajouter au moins quelques minutes à l'heure du snapshot pour tenir compte des variations d'horloge.

Pour plus de détails, consultez la documentation d'Oracle sur la rubrique « Restauration à l'aide de l'optimisation des snapshots de stockage » disponible dans les différentes versions de la documentation d'Oracle 12c. Consultez également le document Oracle document ID Doc ID 604683.1 concernant la prise en charge des snapshots tiers par Oracle.

Disposition des données

La disposition la plus simple consiste à isoler les fichiers de données dans un ou plusieurs volumes dédiés. Ils doivent être non contaminés par tout autre type de fichier. Cela permet de s'assurer que les volumes de fichiers de données peuvent être rapidement restaurés lors d'une opération SnapRestore sans détruire un journal de reprise, un fichier de contrôle ou un journal d'archivage important.

LE SYSTÈME SAN présente des exigences similaires en matière d'isolation des fichiers de données dans des volumes dédiés. Avec un système d'exploitation tel que Microsoft Windows, un seul volume peut contenir plusieurs LUN de fichiers de données, chacune avec un système de fichiers NTFS. Avec d'autres systèmes d'exploitation, il existe généralement un gestionnaire de volumes logiques. Par exemple, avec Oracle ASM, l'option la plus simple consiste à limiter les groupes de disques à un volume unique pouvant être sauvegardé et restauré comme une unité. Si des volumes supplémentaires sont nécessaires pour des raisons de performance ou de gestion de la capacité, la création d'un groupe de disques supplémentaire sur le nouveau volume simplifie la gestion.

Si ces instructions sont respectées, les snapshots peuvent être planifiés directement sur ONTAP sans avoir à créer de snapshot de groupe de cohérence. En effet, les sauvegardes optimisées pour les snapshots ne nécessitent pas la sauvegarde simultanée de fichiers de données.

Une complication se produit dans des situations telles qu'un groupe de disques ASM distribué sur des volumes. Dans ce cas, un snapshot de groupe de cohérence doit être réalisé pour s'assurer que les

métadonnées ASM sont cohérentes sur tous les volumes constitutifs.

[Remarque]Vérifiez que les fichiers spfile et passwd ASM ne se trouvent pas dans le groupe de disques hébergeant les fichiers de données. Cela interfère avec la capacité à restaurer de manière sélective les fichiers de données et uniquement les fichiers de données.

Procédure de restauration locale : NFS

Cette procédure peut être conduite manuellement ou via une application telle que SnapCenter. La procédure de base est la suivante :

1. Arrêtez la base de données.
2. Restaurez le ou les volumes de fichiers de données sur l'instantané immédiatement avant le point de restauration souhaité.
3. Réexécutez les journaux d'archivage au point souhaité.

Cette procédure suppose que les journaux d'archive souhaités sont toujours présents dans le système de fichiers actif. Si ce n'est pas le cas, les journaux d'archive doivent être restaurés, ou `rman` ou `sqlplus` peut être dirigé vers les données dans le `.snapshot` répertoire.

En outre, dans le cas de bases de données plus petites, l'utilisateur peut restaurer les fichiers de données directement à partir du système `.snapshot` Répertoire n'ayant pas besoin des outils d'automatisation ou d'un administrateur du stockage pour exécuter une commande SnapRestore.

Procédure de restauration locale—SAN

Cette procédure peut être conduite manuellement ou via une application telle que SnapCenter. La procédure de base est la suivante :

1. Arrêtez la base de données.
2. Arrêter le ou les groupes de disques hébergeant les fichiers de données. La procédure varie en fonction du gestionnaire de volumes logiques choisi. Avec ASM, le processus nécessite de démonter le groupe de disques. Sous Linux, les systèmes de fichiers doivent être démontés et les volumes logiques et les groupes de volumes désactivés. L'objectif est d'arrêter toutes les mises à jour du groupe de volumes cible à restaurer.
3. Restaurez les groupes de disques de fichiers de données sur l'instantané immédiatement avant le point de restauration souhaité.
4. Réactivez les groupes de disques récemment restaurés.
5. Réexécutez les journaux d'archivage au point souhaité.

Cette procédure suppose que les journaux d'archive souhaités sont toujours présents dans le système de fichiers actif. Si ce n'est pas le cas, les journaux d'archivage doivent être restaurés en mettant les LUN du journal d'archivage hors ligne et en effectuant une restauration. Il s'agit également d'un exemple dans lequel il est utile de diviser les journaux d'archivage en volumes dédiés. Si les journaux d'archivage partagent un groupe de volumes avec les journaux de reprise, les journaux de reprise doivent être copiés ailleurs avant la restauration de l'ensemble global de LUN afin d'éviter de perdre les transactions enregistrées finales.

Exemple de récupération complète

Supposons que les fichiers de données ont été corrompus ou détruits et qu'une restauration complète est requise. La procédure à suivre est la suivante :

```
[oracle@host1 ~]$ sqlplus / as sysdba
Connected to an idle instance.
SQL> startup mount;
ORACLE instance started.
Total System Global Area 1610612736 bytes
Fixed Size                  2924928 bytes
Variable Size            1040191104 bytes
Database Buffers          553648128 bytes
Redo Buffers               13848576 bytes
Database mounted.
SQL> recover automatic;
Media recovery complete.
SQL> alter database open;
Database altered.
SQL>
```

Exemple de restauration instantanée

Toute la procédure de restauration est une commande unique : `recover automatic`.

Si une restauration à un point dans le temps est requise, l'horodatage des snapshots doit être connu et peut être identifié comme suit :

```
Cluster01::> snapshot show -vserver vserver1 -volume NTAP_oradata -fields
create-time
vserver   volume        snapshot   create-time
--------  ------------  ---------  ------------------------
vserver1  NTAP_oradata  my-backup  Thu Mar 09 10:10:06 2017
```

L'heure de création de l'instantané est répertoriée comme 9 mars et 10:10:06. Pour être sûr, une minute est ajoutée à l'heure du snapshot :

```
[oracle@host1 ~]$ sqlplus / as sysdba
Connected to an idle instance.
SQL> startup mount;
ORACLE instance started.
Total System Global Area 1610612736 bytes
Fixed Size                  2924928 bytes
Variable Size            1040191104 bytes
Database Buffers          553648128 bytes
Redo Buffers               13848576 bytes
Database mounted.
SQL> recover database until time '09-MAR-2017 10:44:15' snapshot time '09-
MAR-2017 10:11:00';
```

La restauration est maintenant lancée. Il a spécifié une heure d'instantané de 10:11:00, une minute après l'heure enregistrée pour tenir compte de la variation d'horloge possible, et un temps de récupération cible de 10:44. Ensuite, sqlplus demande les journaux d'archivage requis pour atteindre le délai de restauration souhaité de 10:44.

```
ORA-00279: change 551760 generated at 03/09/2017 05:06:07 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_31_930813377.dbf
ORA-00280: change 551760 for thread 1 is in sequence #31
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 552566 generated at 03/09/2017 05:08:09 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_32_930813377.dbf
ORA-00280: change 552566 for thread 1 is in sequence #32
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 553045 generated at 03/09/2017 05:10:12 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_33_930813377.dbf
ORA-00280: change 553045 for thread 1 is in sequence #33
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 753229 generated at 03/09/2017 05:15:58 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_34_930813377.dbf
ORA-00280: change 753229 for thread 1 is in sequence #34
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
Log applied.
Media recovery complete.
SQL> alter database open resetlogs;
Database altered.
SQL>
```

Restauration complète d'une base de données à l'aide de snapshots à l'aide de `recover automatic` la commande ne nécessite pas de licence spécifique, mais une restauration à un point dans le temps via `snapshot time` Requiert la licence Oracle Advanced compression.

Outils d'automatisation et de gestion de la base de données

Dans un environnement de base de données Oracle, la principale valeur de ONTAP provient des principales technologies ONTAP, telles que les copies Snapshot instantanées, la réplication simple SnapMirror et la création efficace de volumes FlexClone.

Dans certains cas, une configuration simple de ces fonctionnalités principales directement sur ONTAP répond aux exigences, mais les besoins plus complexes requièrent une couche d'orchestration.

SnapCenter

SnapCenter est le produit phare de la protection des données NetApp. À un niveau très bas, il est similaire aux produits SnapManager en termes d'exécution des sauvegardes de base de données, mais il a été conçu dès le départ pour proposer une gestion de la protection des données centralisée sur les systèmes de stockage NetApp.

SnapCenter inclut les fonctions de base telles que les sauvegardes et restaurations basées sur des snapshots, SnapMirror et la réplication SnapVault, ainsi que d'autres fonctionnalités nécessaires pour fonctionner à grande échelle pour les grandes entreprises. Ces fonctionnalités avancées incluent un contrôle d'accès basé sur des rôles (RBAC) étendu, des API RESTful pour l'intégration de produits d'orchestration tiers, une gestion centralisée et sans interruption des plug-ins SnapCenter sur des hôtes de base de données et une interface utilisateur conçue pour les environnements à l'échelle du cloud.

REPOS

ONTAP contient également un jeu d'API RESTful riche. Les fournisseurs tiers peuvent ainsi créer une application de protection des données et de gestion grâce à une intégration étroite avec ONTAP. De plus, l'API RESTful est facile à utiliser par les clients qui souhaitent créer leurs propres workflows et utilitaires d'automatisation.

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.

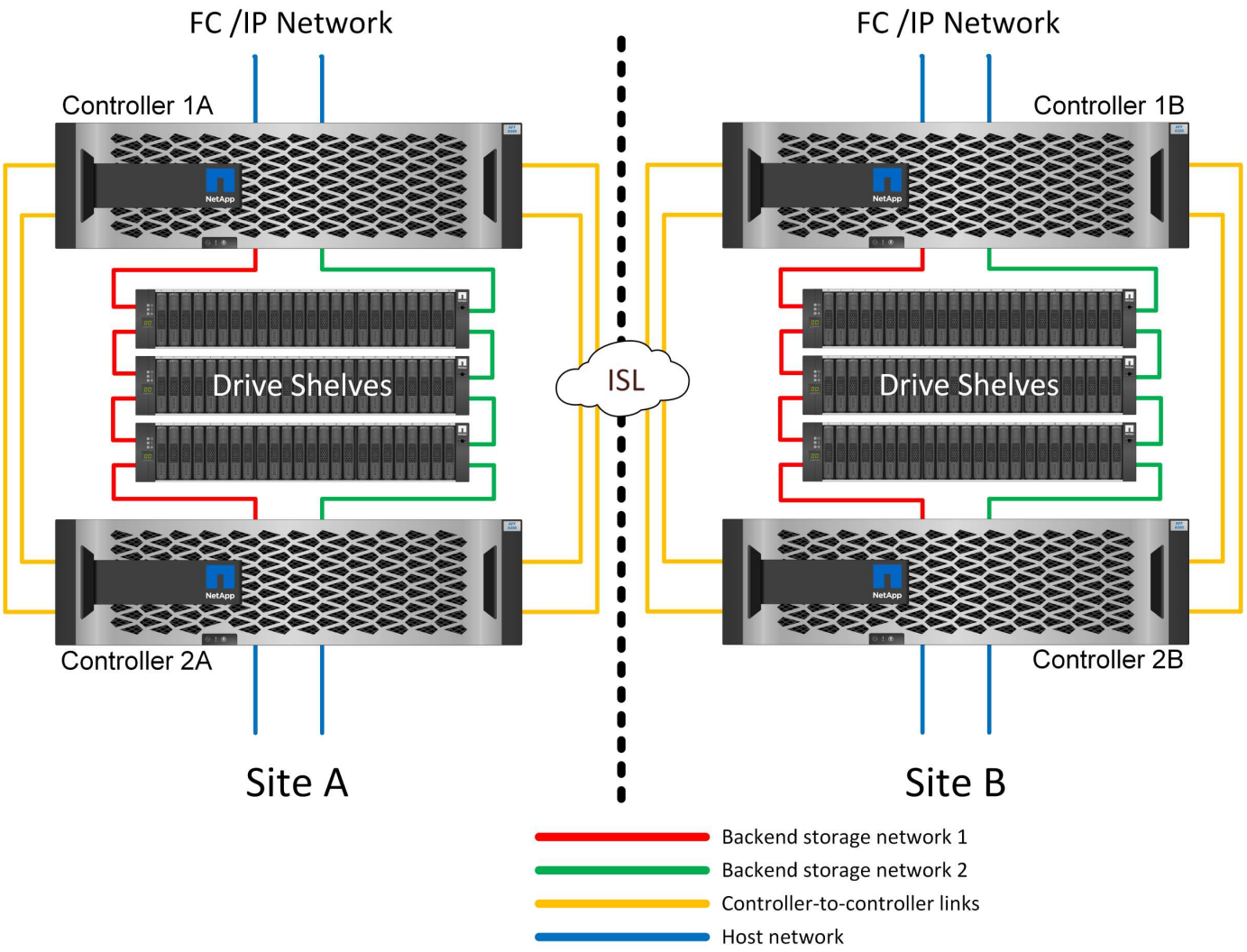
IP MetroCluster

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.

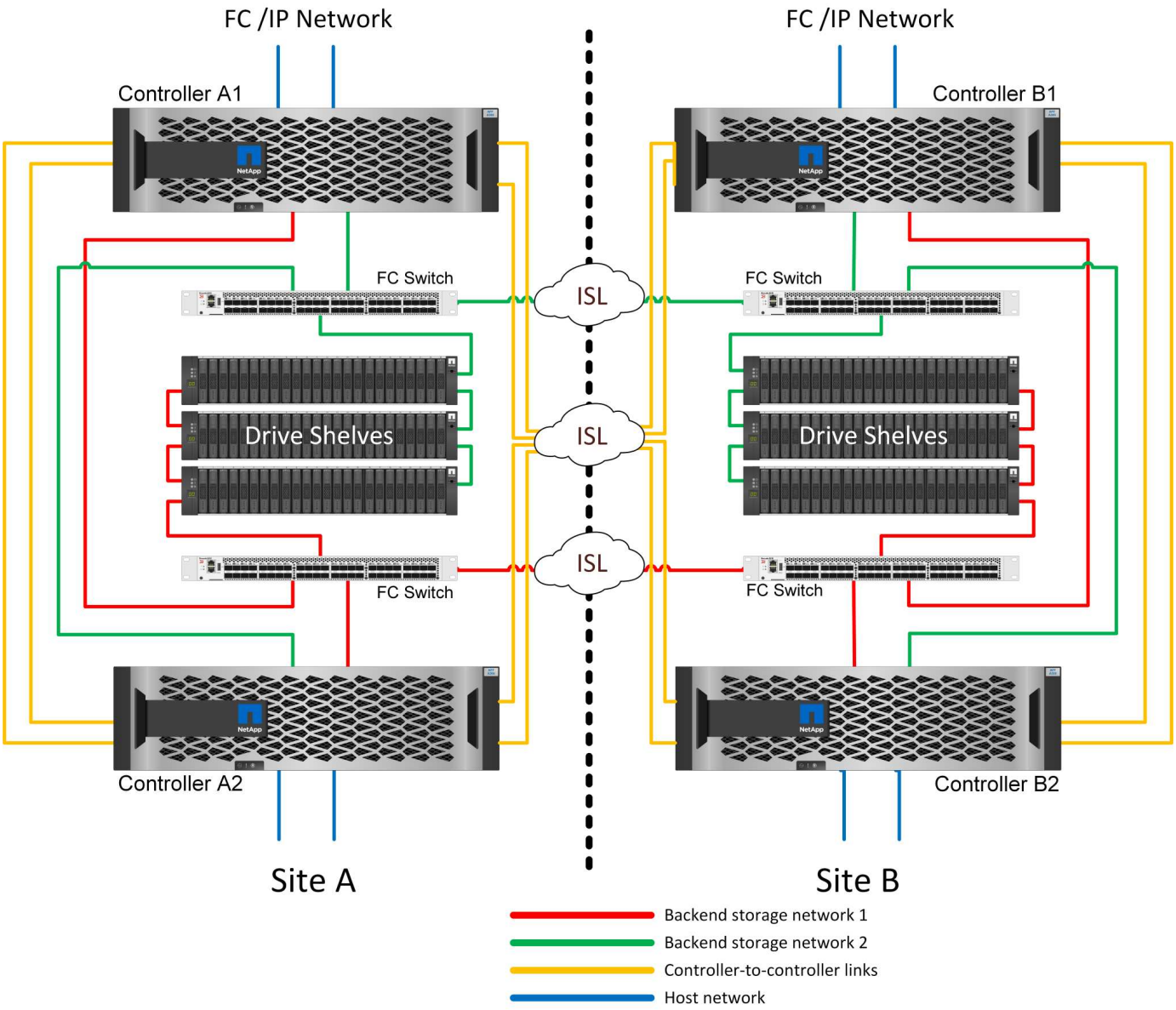
L'architecture est également plus simple, car les contrôleurs disposent des seules connexions entre les sites. Dans les MetroCluster FC, un contrôleur écrit directement sur les disques du site opposé et requiert ainsi des connexions SAN, des commutateurs et des ponts supplémentaires. En revanche, un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



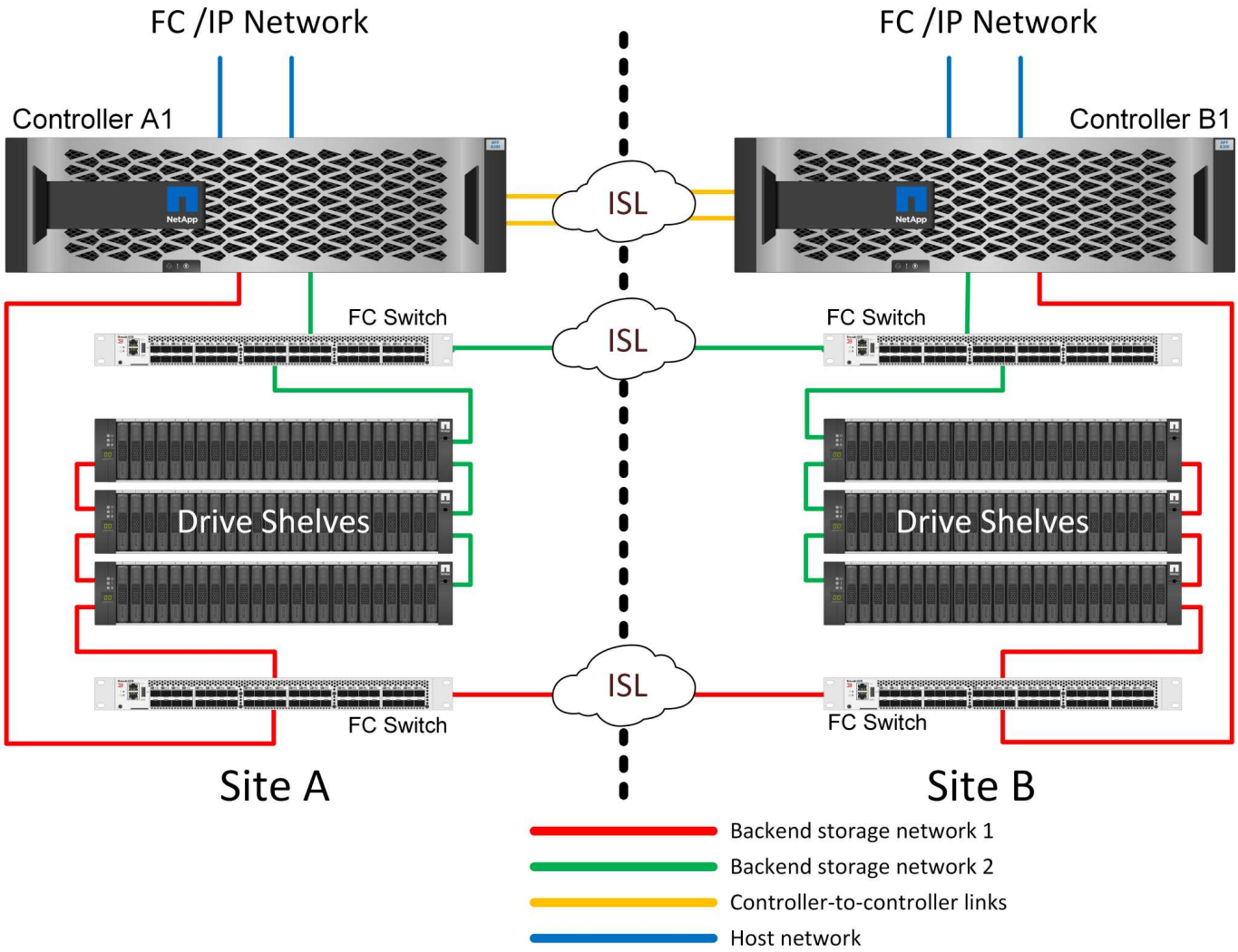
Certaines infrastructures multisites ne sont pas conçues pour les opérations en mode actif-actif. Elles sont plutôt utilisées comme site principal et site de reprise après incident. Dans ce cas, il est généralement préférable d'utiliser une option MetroCluster à paire HA pour les raisons suivantes :

- 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

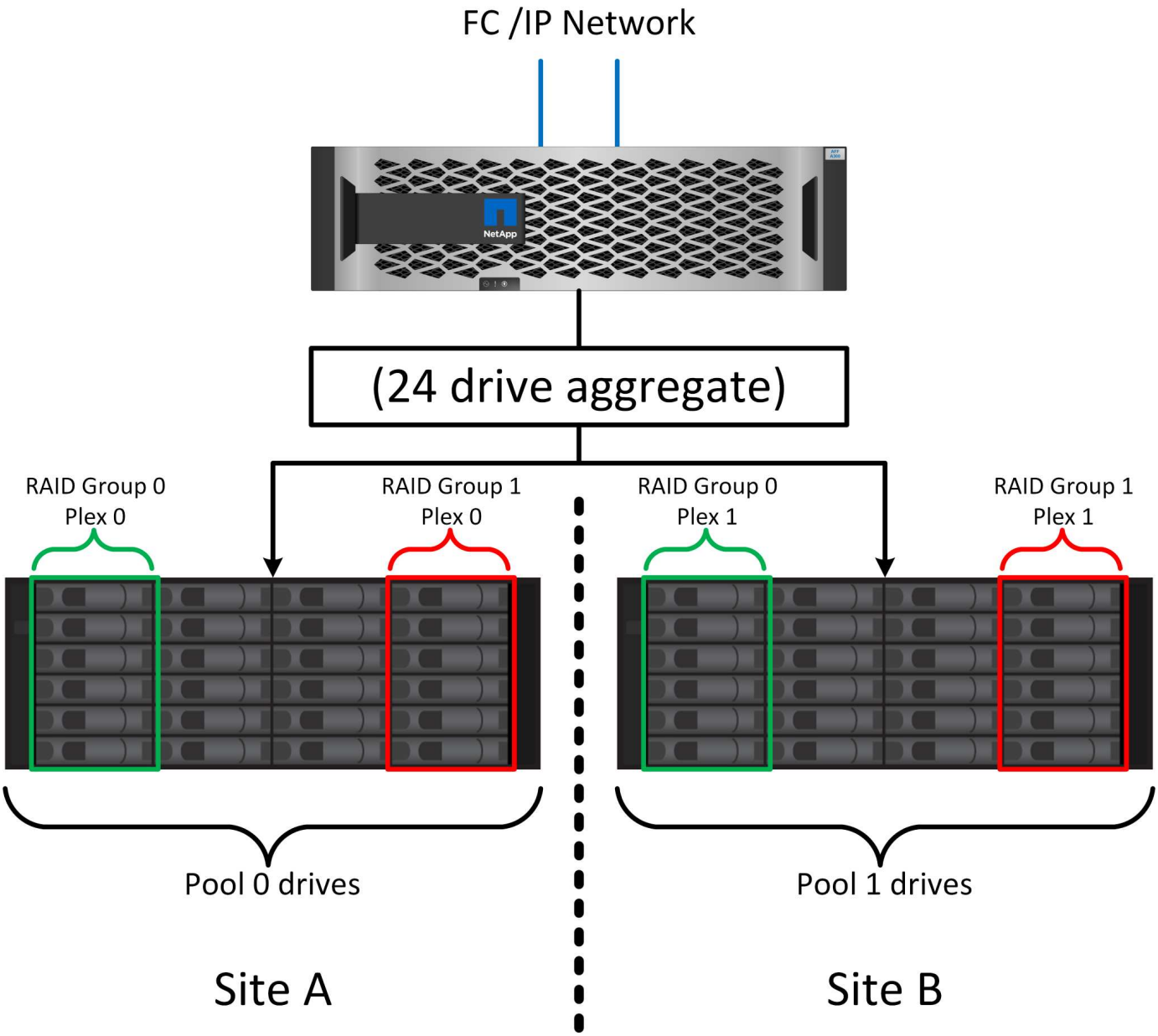
MetroCluster étend la protection des données NVRAM de plusieurs manières :

- 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

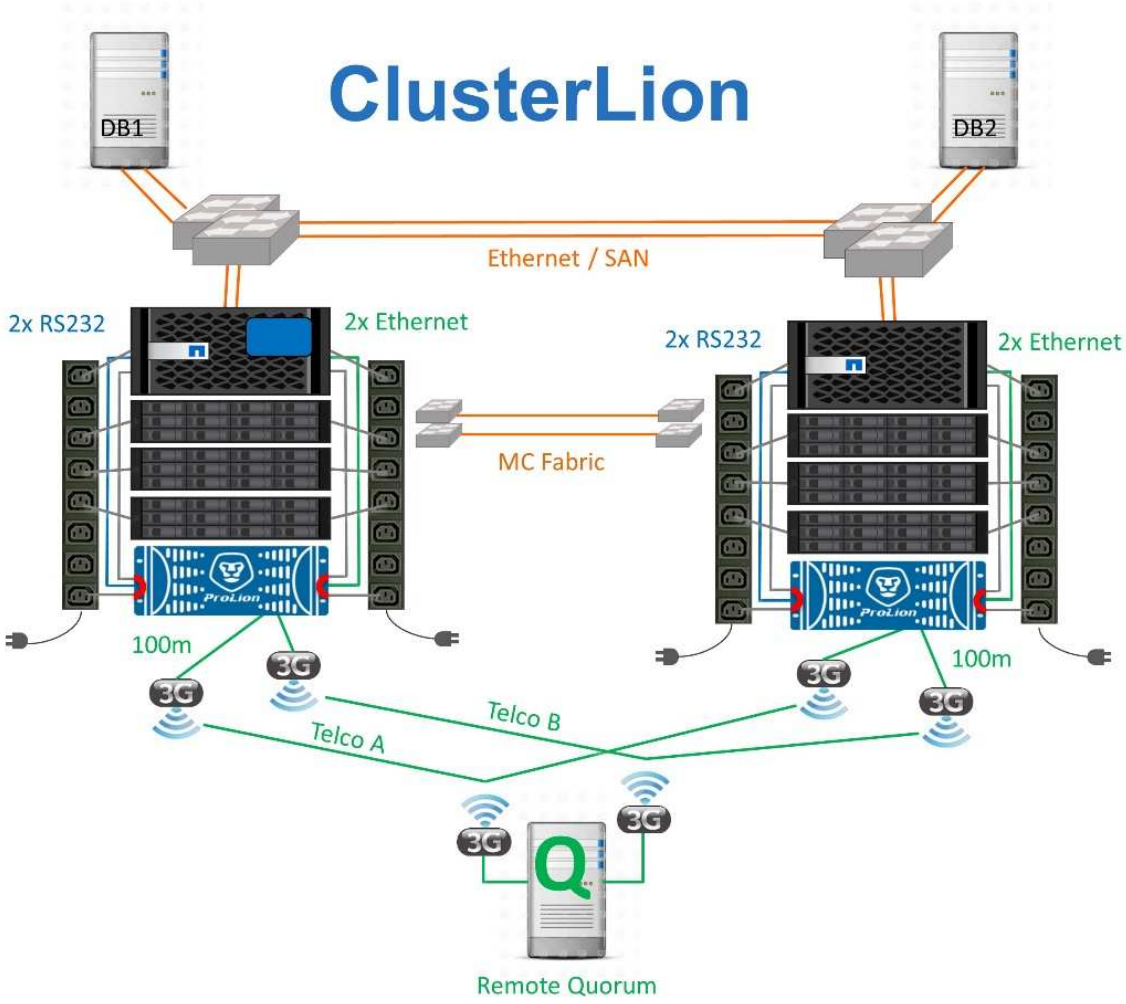
Le médiateur ONTAP est utilisé avec MetroCluster IP et certaines autres solutions ONTAP. Il fonctionne comme un service disjoncteur d'attache classique, tout comme le logiciel disjoncteur d'attache MetroCluster mentionné ci-dessus, mais comprend également une fonctionnalité essentielle, qui effectue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

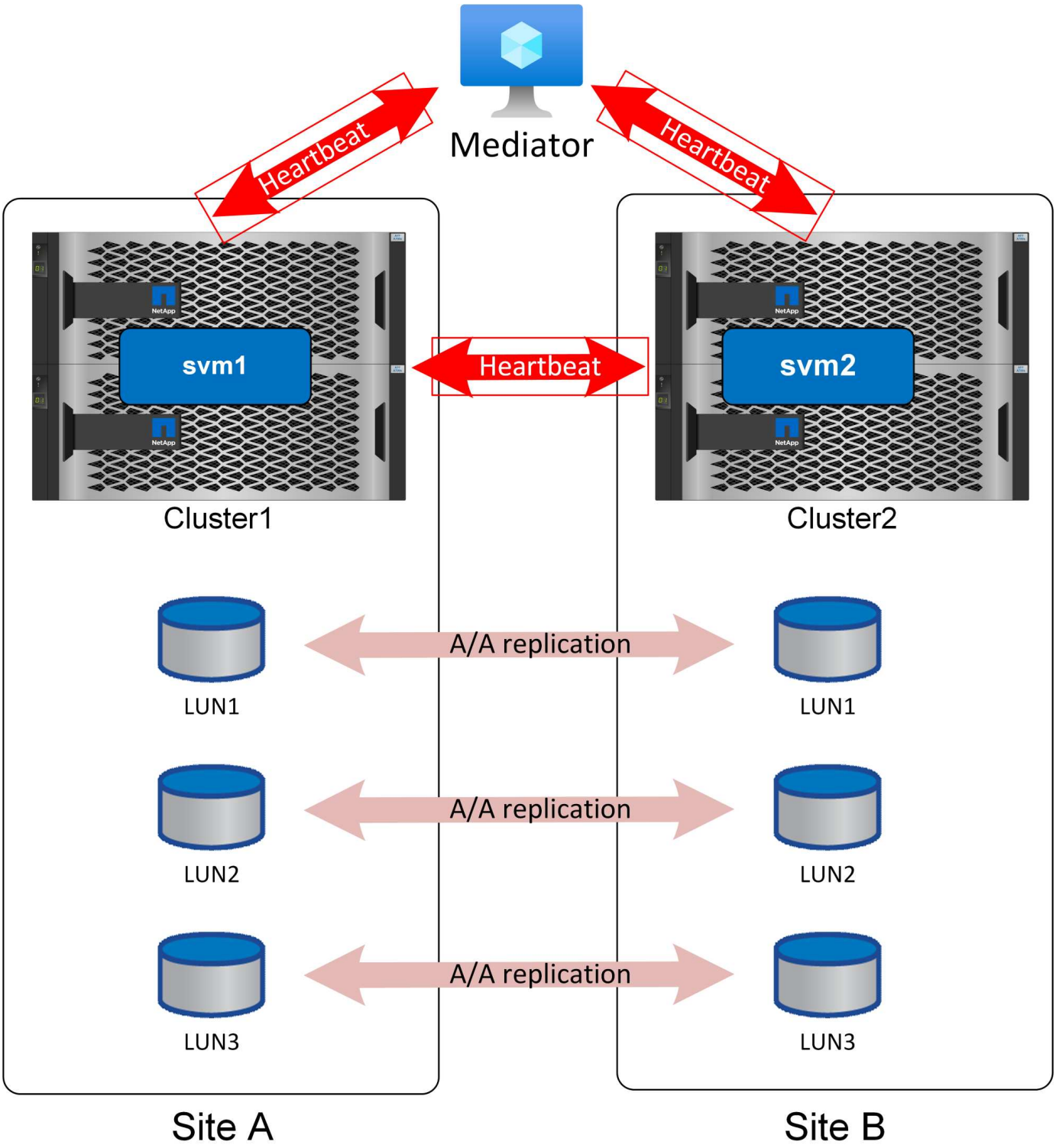
ONTAP Médiateur

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.

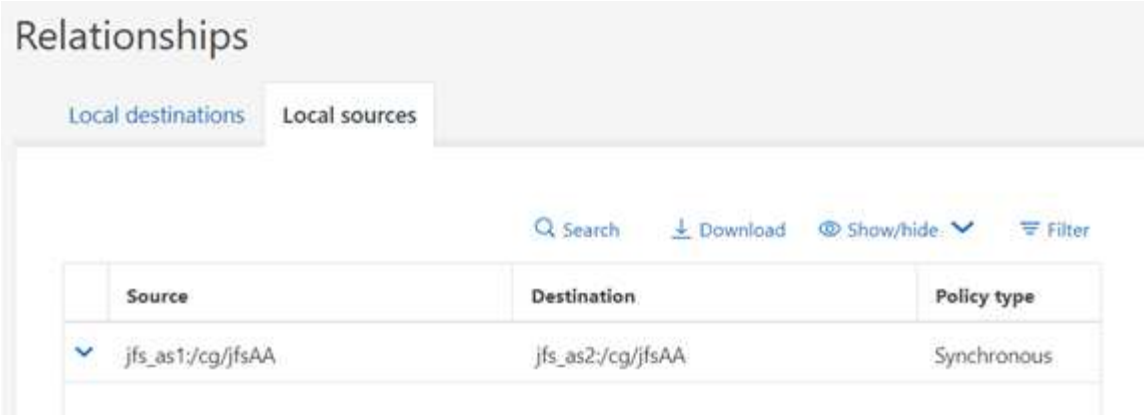
Site préféré de la synchronisation active SnapMirror

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.

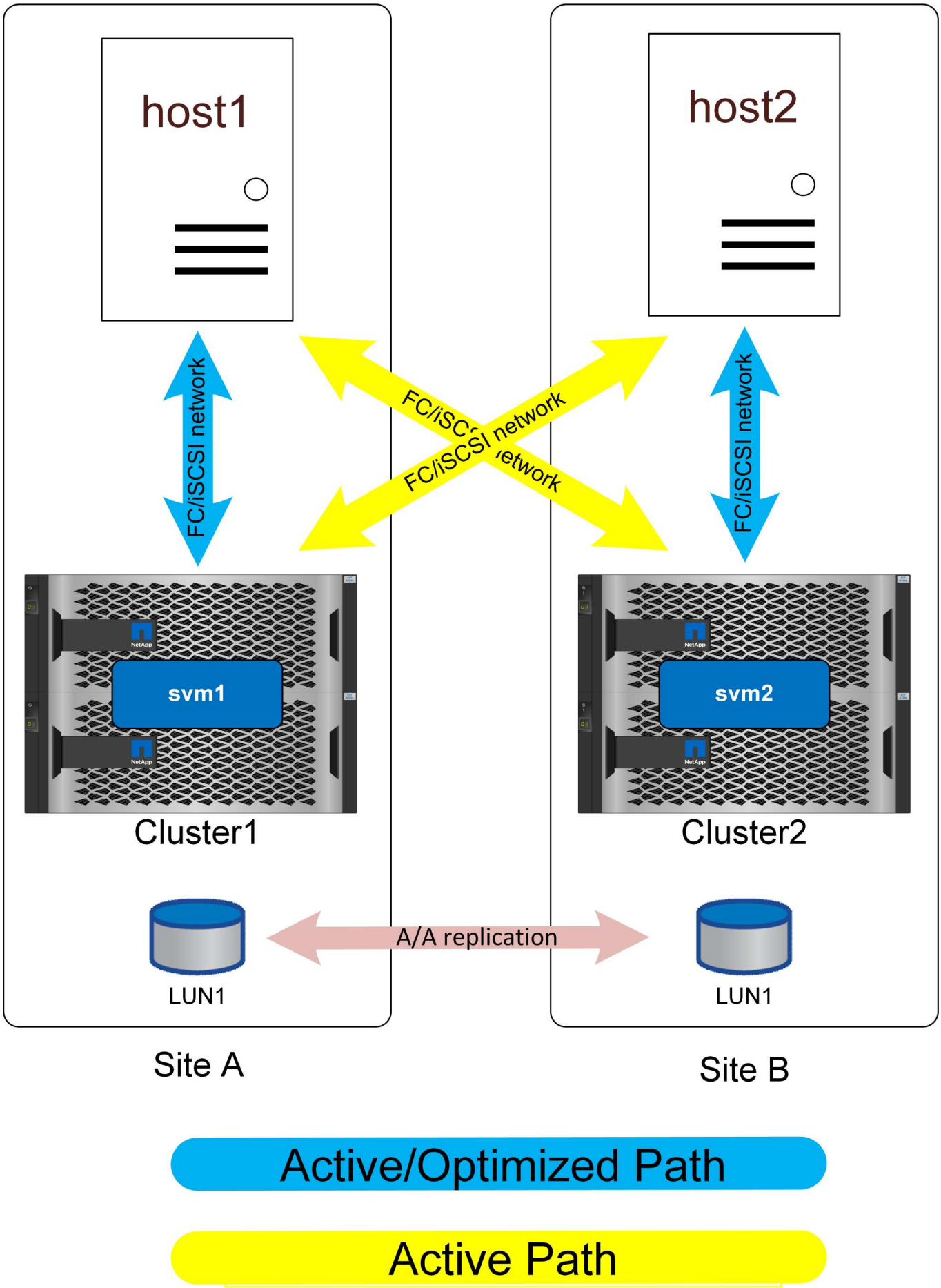
La deuxième étape facultative consiste à configurer la proximité de l'hôte. Cela ne contrôle pas l'accès, il contrôle *Priority*.

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

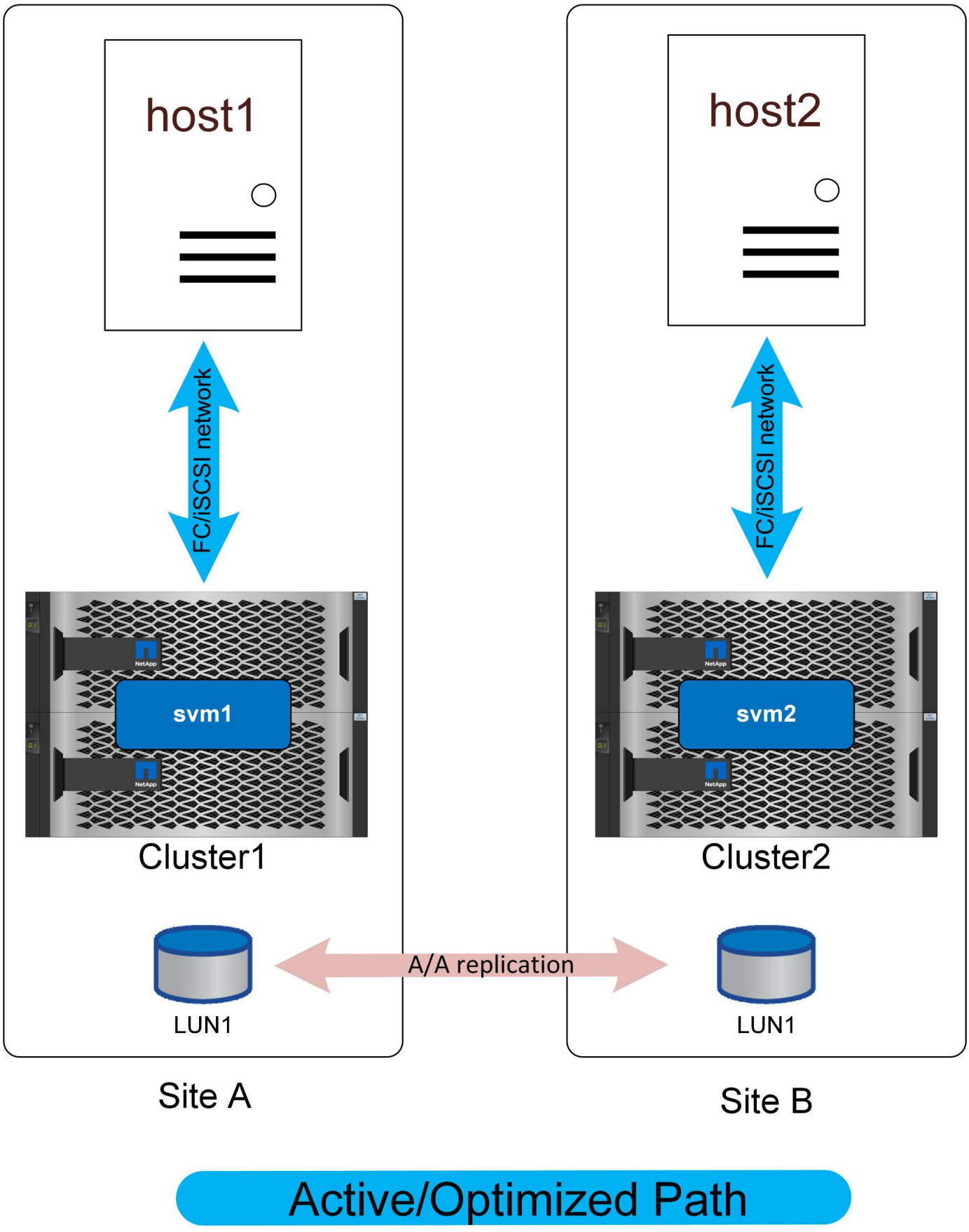
En fonctionnement normal, toutes les E/S sont des E/S locales. Les opérations de lecture et d'écriture sont gérées à partir de la baie de stockage locale. Bien entendu, les E/S en écriture devront également être répliquées par le contrôleur local sur le système distant avant d'être acquittées, mais toutes 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.



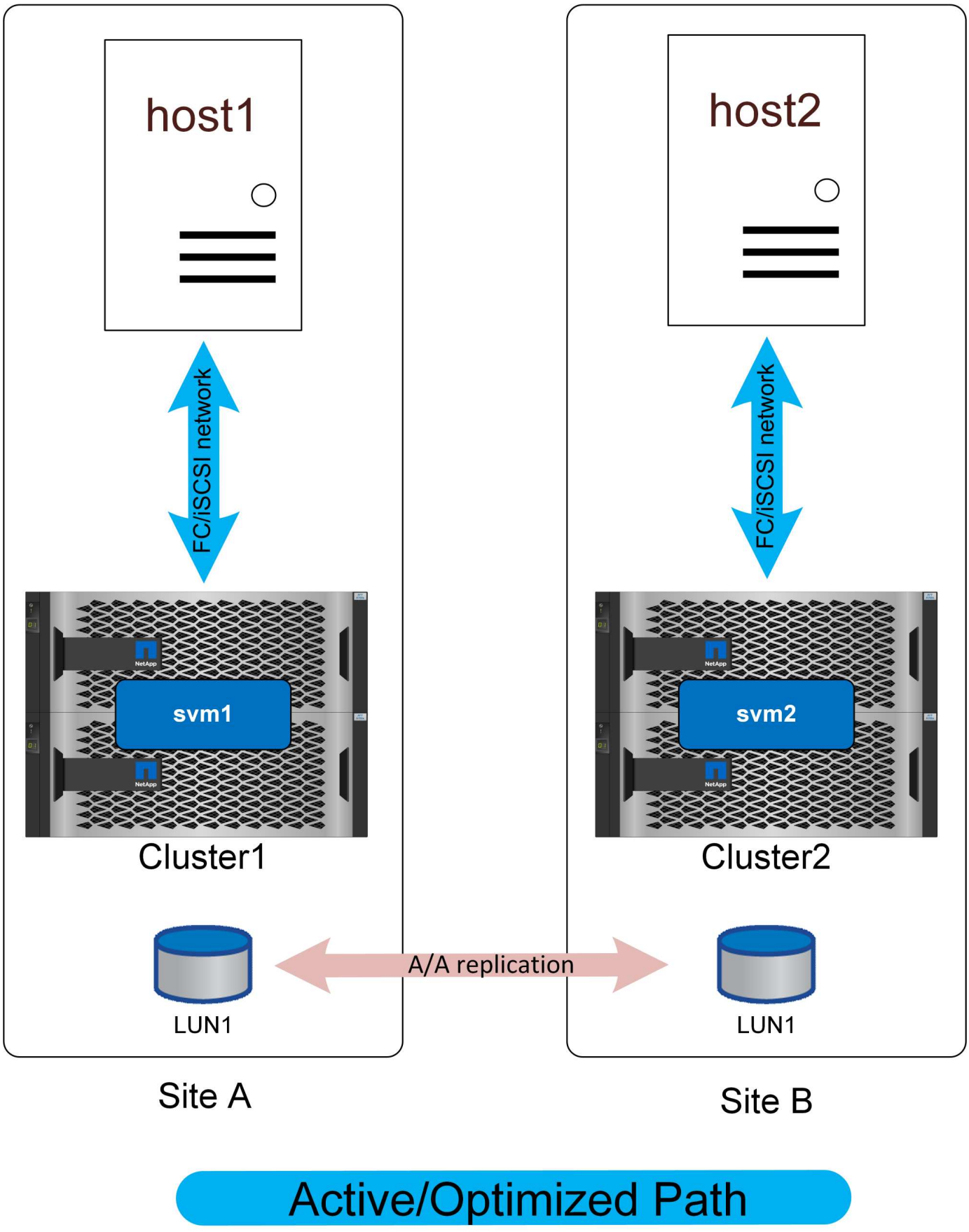
Une configuration ASA avec un accès non uniforme fonctionnera en grande partie comme avec AFF. Avec un accès uniforme, l'E/S traverserait le WAN. Cela peut être souhaitable ou non.

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.

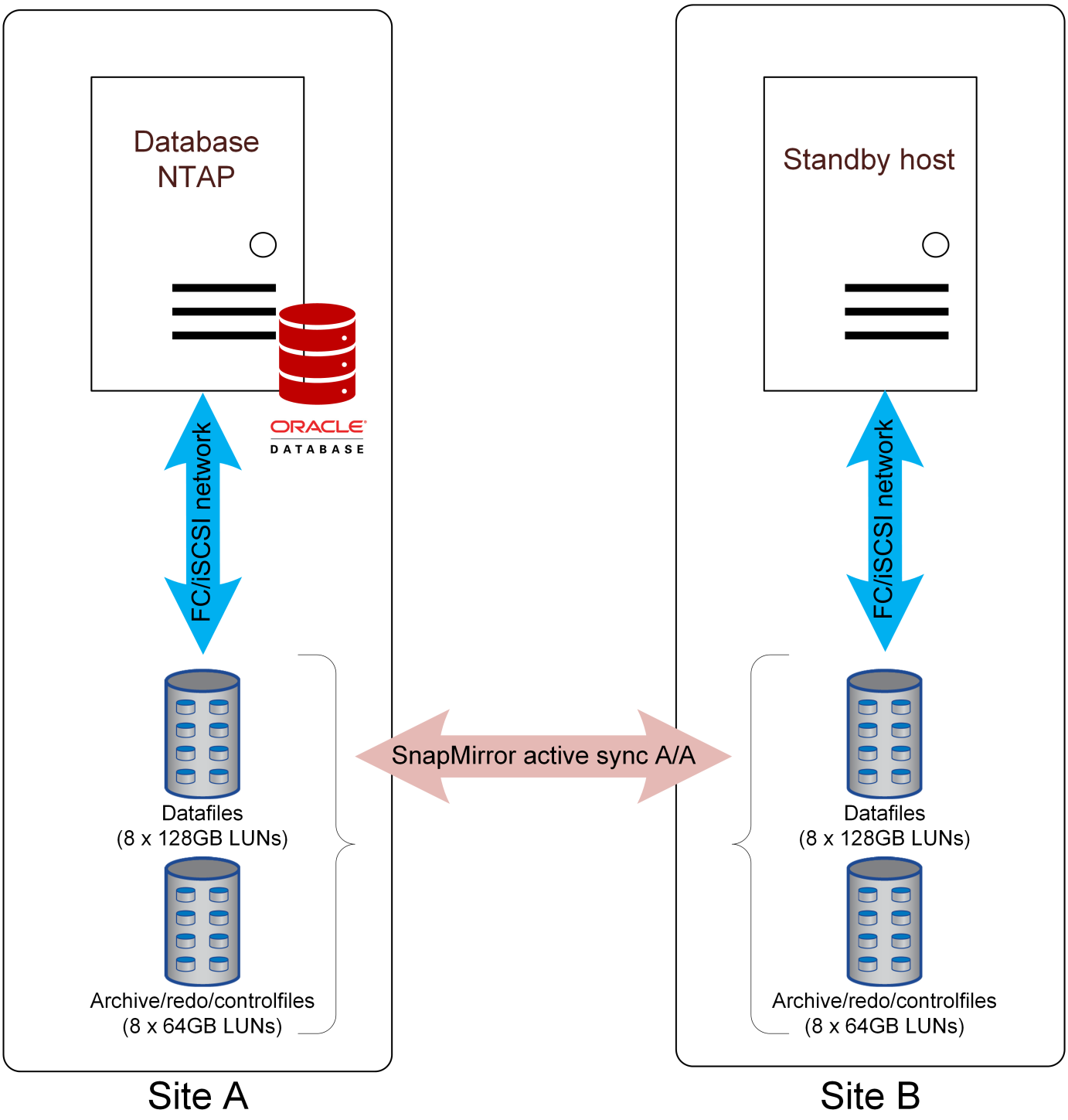
La meilleure architecture dépend des besoins de l’entreprise. Par exemple, si l’objectif est de bénéficier d’une protection RPO=0 contre la perte de données, mais que l’objectif RTO est assoupli, l’utilisation de bases de données Oracle Single instance et la réplication des LUN avec SM-AS peuvent suffire 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

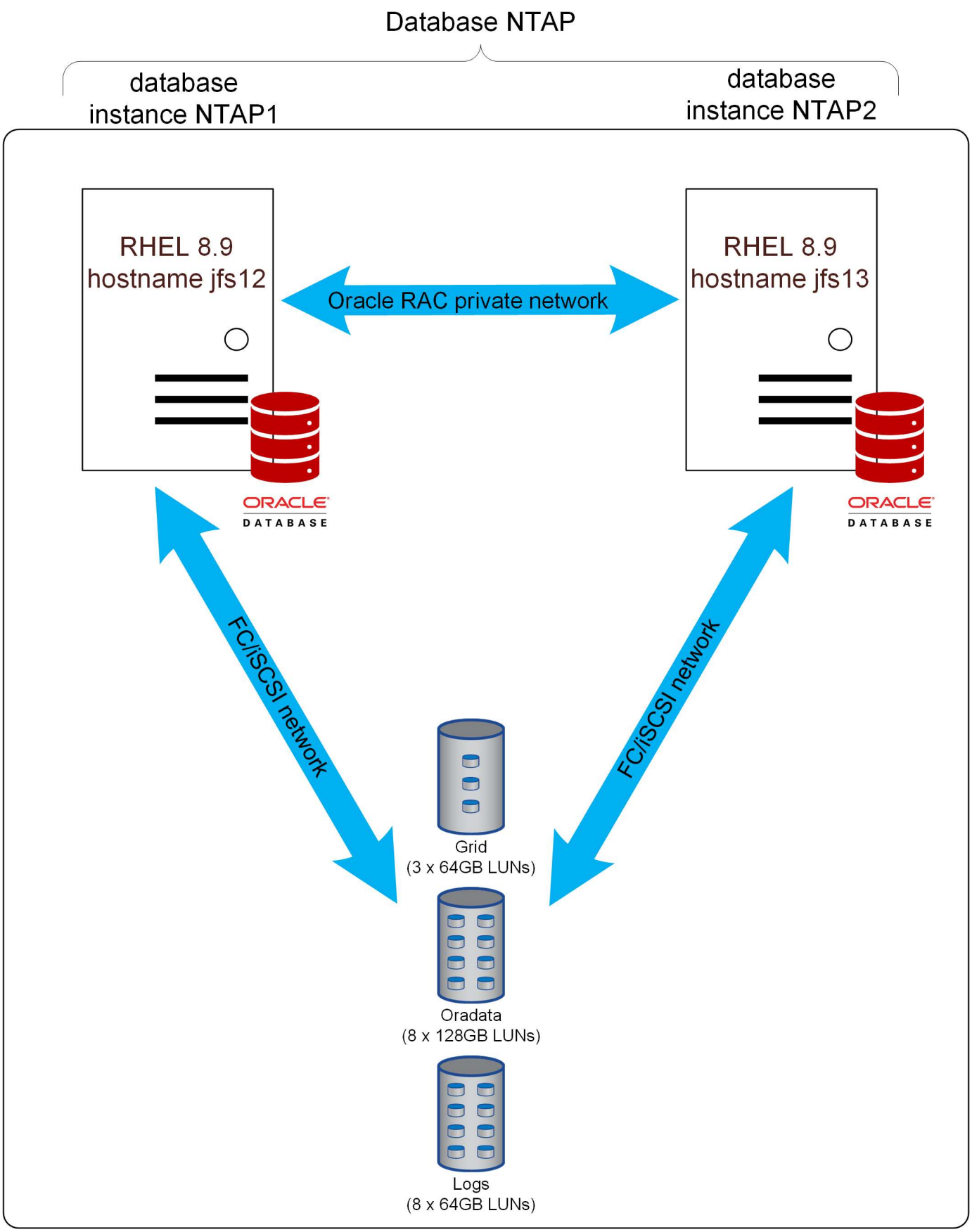
De nombreux clients optimisent leur RTO en étendant un cluster Oracle RAC sur plusieurs sites, offrant une configuration entièrement active/active. La conception globale devient plus complexe car elle doit inclure la gestion du quorum d'Oracle RAC.

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

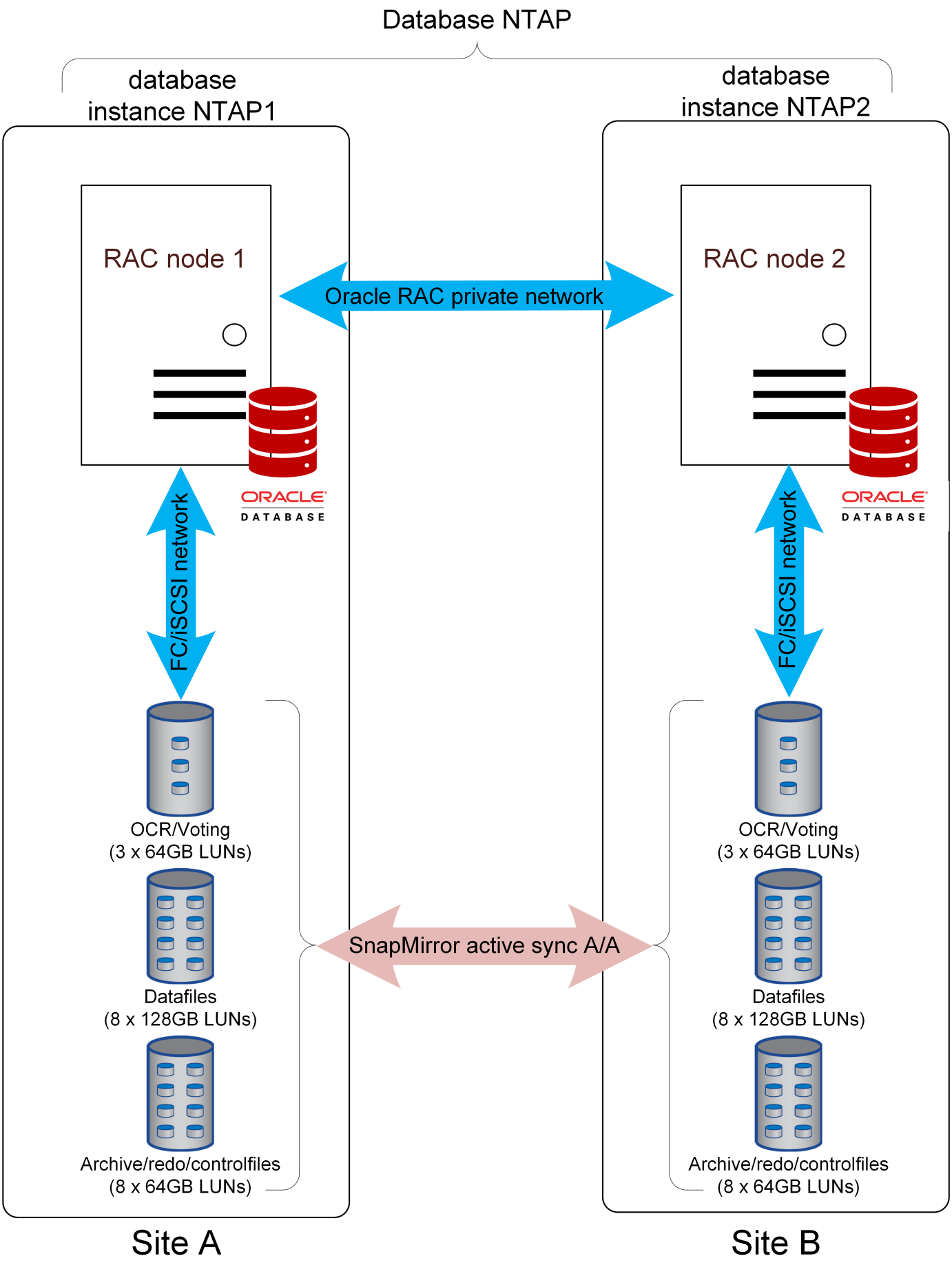
L'élément le plus important à prendre en compte avec Oracle RAC sur SnapMirror Active Sync est de savoir s'il faut utiliser un accès uniforme ou non.

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

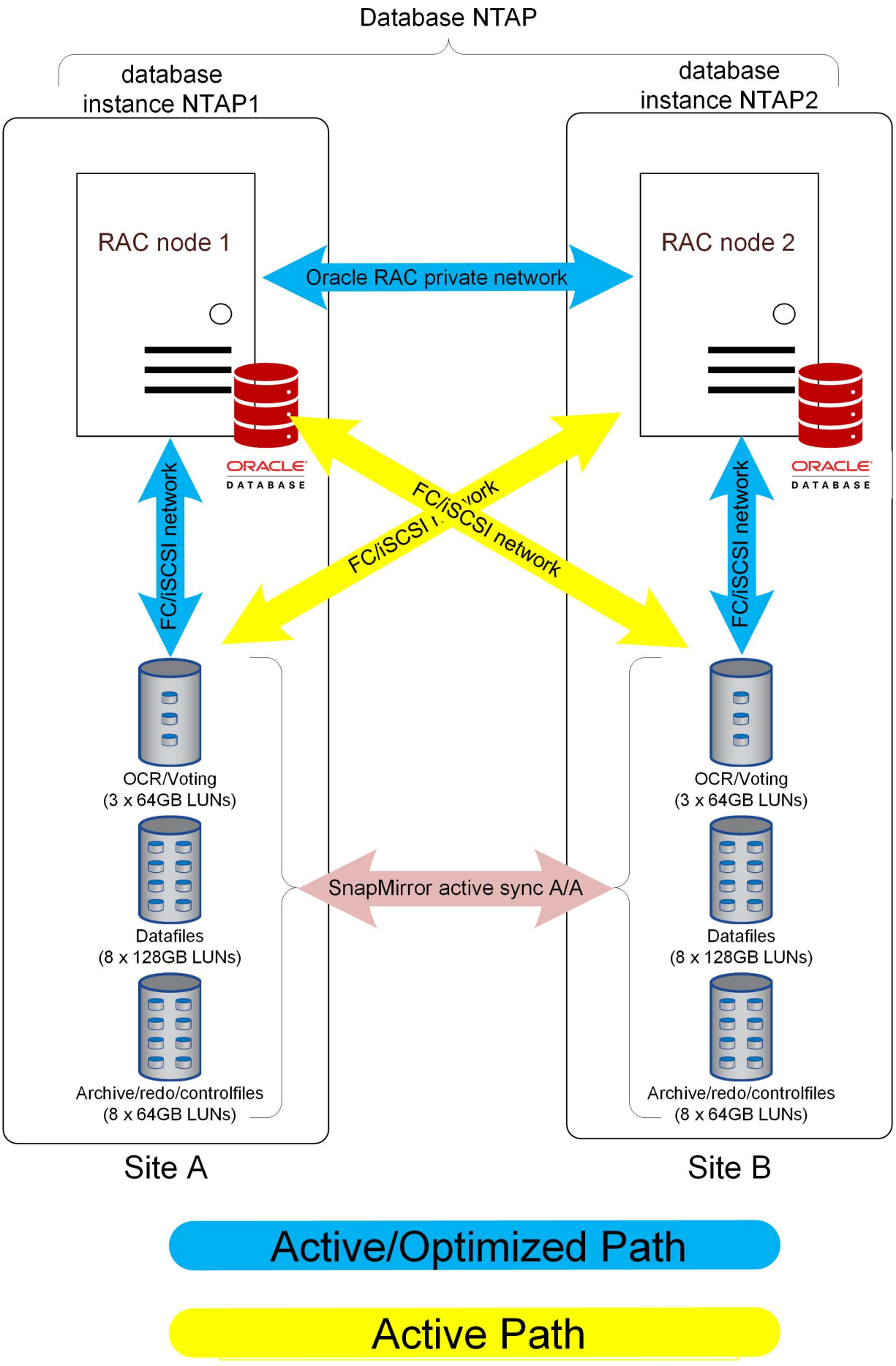
L’accès uniforme requiert l’extension du SAN sur les sites. Le principal avantage est que la perte d’un système de stockage n’entraîne pas la perte d’une instance de base de données. Au lieu de cela, cela entraînerait une modification des chemins d’accès multiples dans lesquels les chemins sont actuellement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

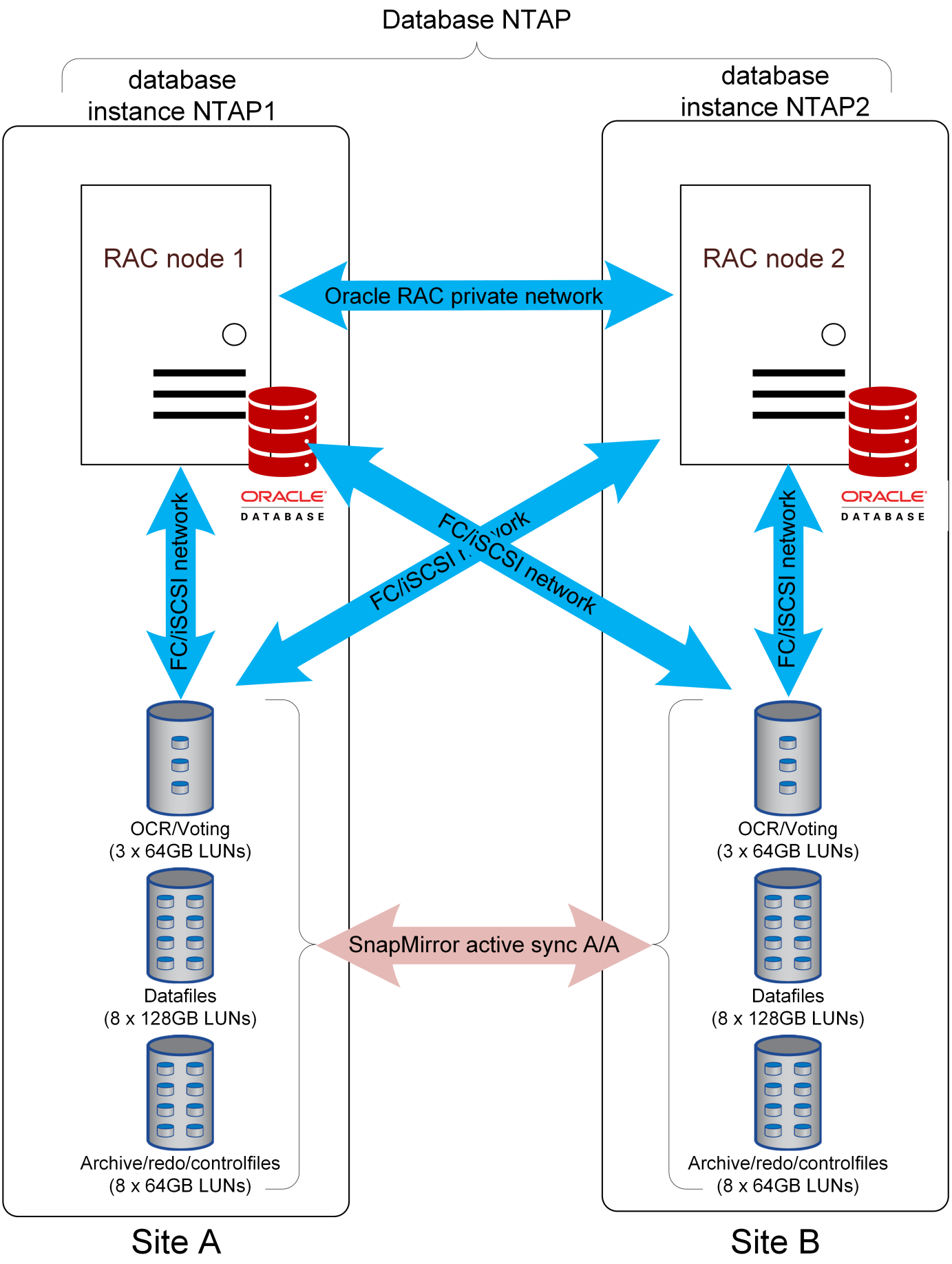
En cas de latence importante entre les sites, les systèmes AFF peuvent être configurés avec des paramètres de proximité des hôtes. Cela permet à chaque système de stockage d’identifier les hôtes locaux et distants, et d’attribuer les priorités de chemin en conséquence.


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

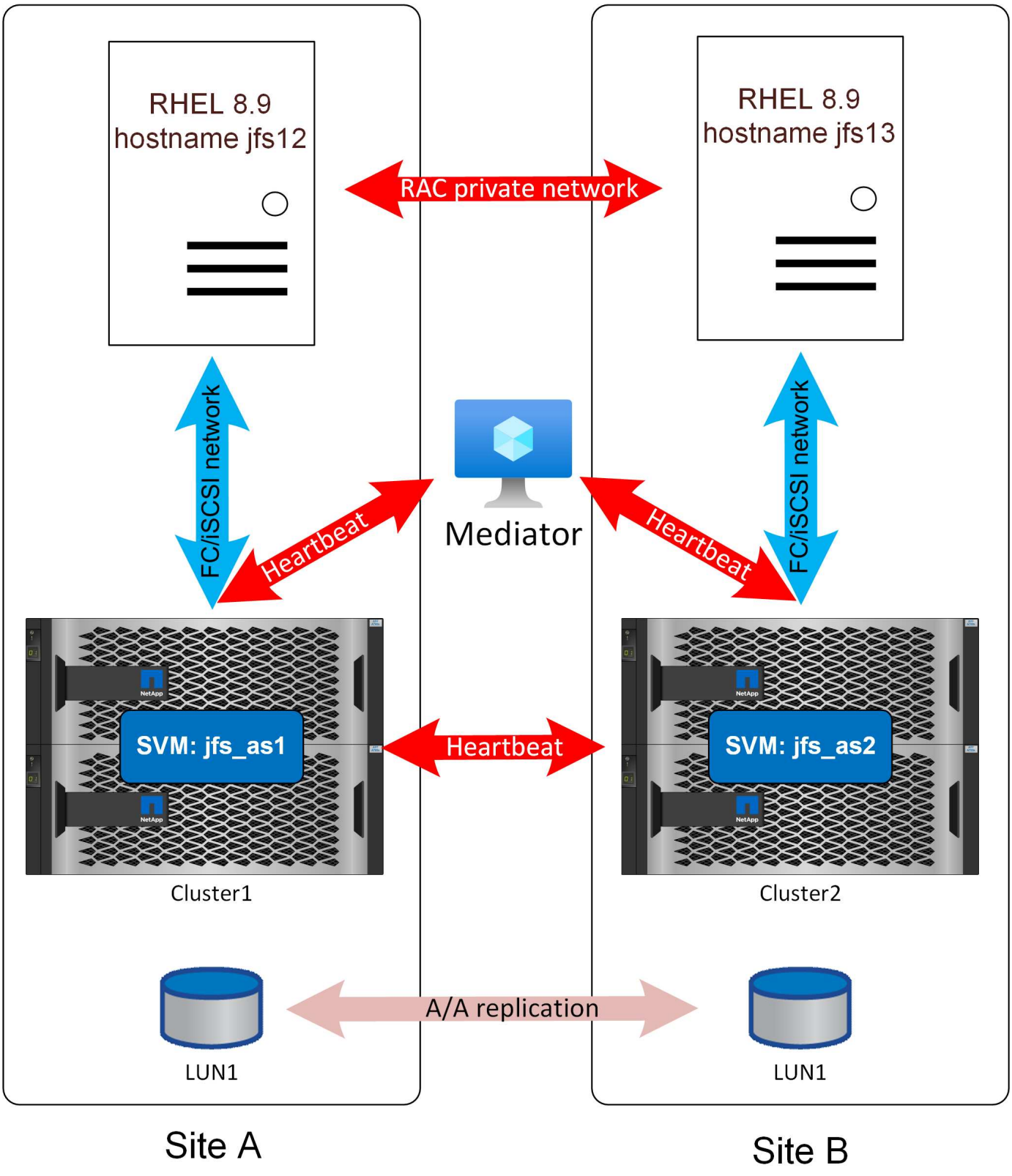
La modification du site préféré nécessite une opération simple. L'E/S s'interrompt pendant une ou deux secondes car l'autorité sur le comportement de réplication change entre les clusters, mais l'E/S n'est pas affectée.

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.

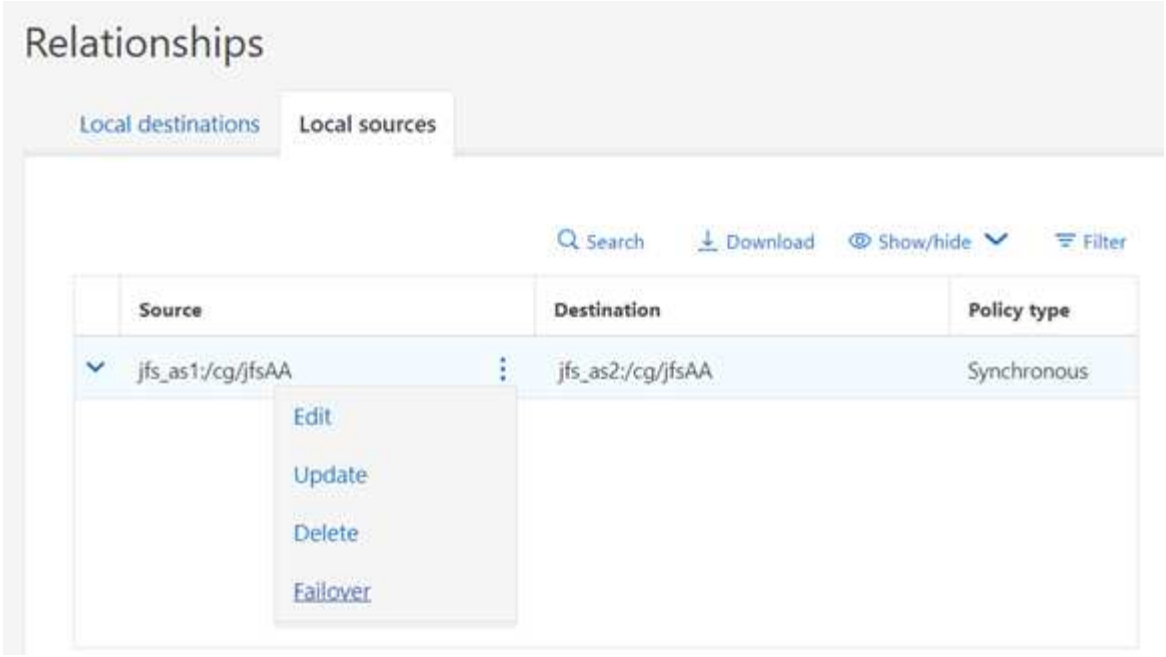
Basculement manuel

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
```

Migration de la base de données Oracle

Présentation

L'exploitation des capacités d'une nouvelle plateforme de stockage implique une seule nécessité : les données doivent être placées sur le nouveau système de stockage. ONTAP simplifie le processus de migration, notamment les migrations et les mises à niveau de ONTAP vers ONTAP, les importations de LUN étrangères et les procédures d'utilisation directe du système d'exploitation hôte ou du logiciel de base de données Oracle.

Cette documentation remplace le rapport technique *TR-4534 : migration des bases de données Oracle vers des systèmes de stockage NetApp*

Dans le cas d'un nouveau projet de base de données, cela ne pose pas de problème car les environnements de base de données et d'application sont construits en place. Cependant, la migration pose des défis

particuliers en ce qui concerne les interruptions d'activité, le temps nécessaire à la réalisation de la migration, les compétences requises et la réduction des risques.

Scripts

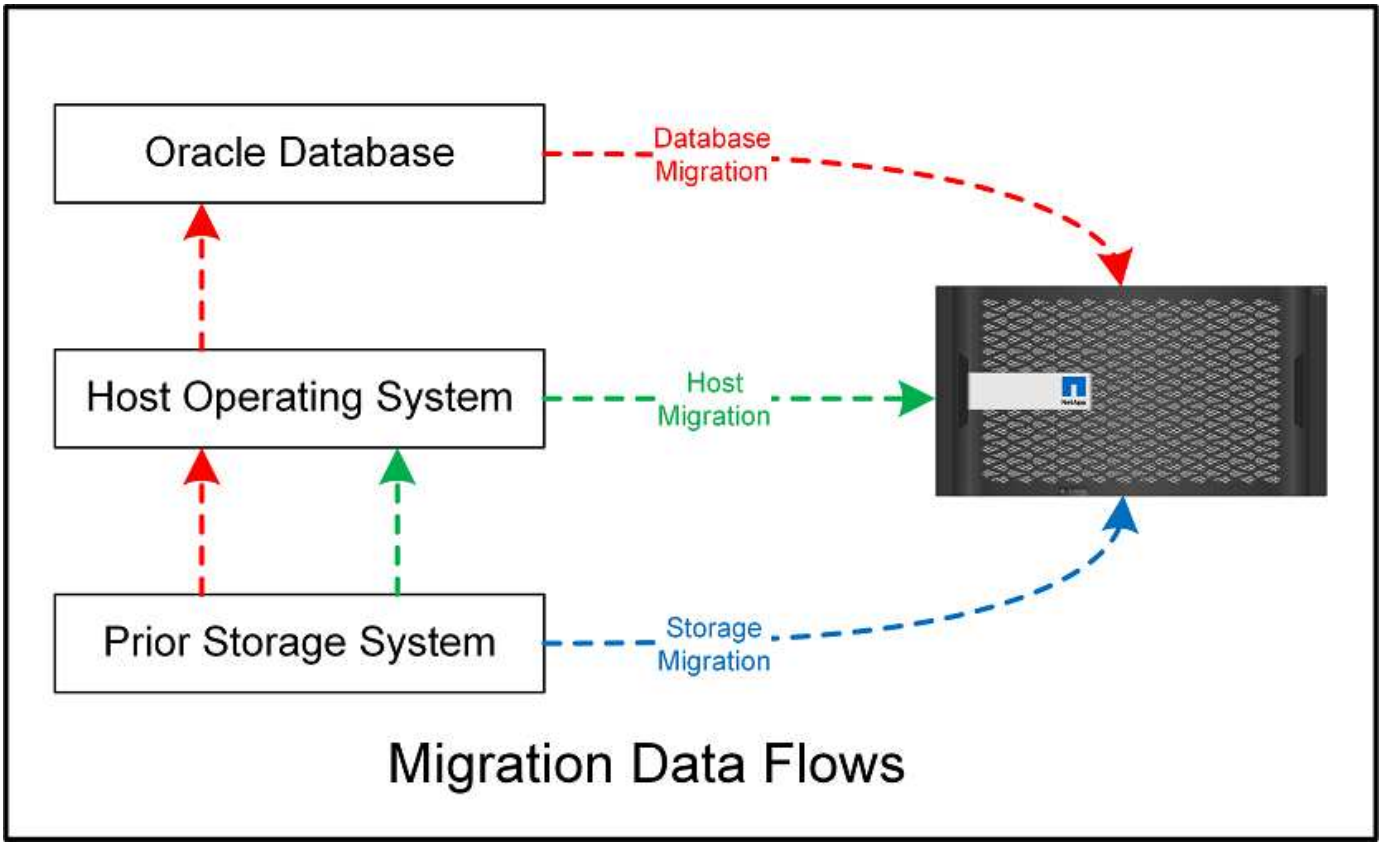
Des exemples de scripts sont fournis dans cette documentation. Ces scripts fournissent des exemples de méthodes d'automatisation de divers aspects de la migration afin de réduire le risque d'erreurs des utilisateurs. Les scripts réduisent les demandes globales de l'équipe INFORMATIQUE responsable de la migration et accélèrent le processus global. Ces scripts sont issus de projets de migration réalisés par les services professionnels de NetApp et les partenaires NetApp. Des exemples de leur utilisation sont présentés dans cette documentation.

Planification des migrations

La migration des données Oracle peut se faire à l'un des trois niveaux suivants : la base de données, l'hôte ou la baie de stockage.

Les différences résident dans la capacité du composant de la solution globale à déplacer les données : la base de données, le système d'exploitation hôte ou le système de stockage.

La figure ci-dessous présente un exemple des niveaux de migration et du flux de données. Dans le cas d'une migration au niveau de la base de données, les données sont déplacées du système de stockage d'origine vers le nouvel environnement via les couches hôte et base de données. La migration au niveau de l'hôte est similaire, mais les données ne passent pas par la couche applicative et sont écrites au nouvel emplacement à l'aide de processus hôtes. Enfin, avec la migration au niveau du stockage, une baie telle qu'un système NetApp FAS est responsable du déplacement des données.



Une migration au niveau de la base de données fait généralement référence à l'utilisation de l'envoi de journaux Oracle via une base de données de secours pour effectuer une migration au niveau de la couche

Oracle. Les migrations au niveau de l'hôte s'effectuent à l'aide de la fonctionnalité native de la configuration du système d'exploitation hôte. Cette configuration inclut des opérations de copie de fichiers à l'aide de commandes telles que cp, tar et Oracle Recovery Manager (RMAN) ou à l'aide d'un gestionnaire de volumes logiques (LVM) pour déplacer les octets sous-jacents d'un système de fichiers. Oracle Automatic Storage Management (ASM) est classé comme une fonctionnalité de niveau hôte car elle s'exécute en dessous du niveau de l'application de base de données. ASM remplace le gestionnaire de volumes logiques habituel sur un hôte. Enfin, les données peuvent être migrées au niveau de la baie de stockage, ce qui signifie en dessous du niveau du système d'exploitation.

Planification

La meilleure option de migration dépend de plusieurs facteurs, notamment de l'étendue de l'environnement à migrer, de la nécessité d'éviter les temps d'indisponibilité et des efforts globaux requis pour effectuer la migration. Les bases de données volumineuses nécessitent évidemment plus de temps et d'efforts pour la migration, mais la complexité de cette migration est minimale. Les petites bases de données peuvent être migrées rapidement. Toutefois, si des milliers d'entre elles doivent être migrées, l'ampleur des efforts peut engendrer des complications. Enfin, plus la base de données est volumineuse, plus elle est susceptible d'être stratégique, ce qui entraîne la nécessité de minimiser les temps d'indisponibilité tout en préservant un chemin « back-out ».

Voici quelques-uns des éléments à prendre en compte lors de la planification d'une stratégie de migration.

Taille des données

La taille des bases de données à migrer a de toute évidence un impact sur la planification de la migration, bien que la taille n'ait pas nécessairement un impact sur le délai de mise en service. Lorsqu'une grande quantité de données doit être migrée, la principale considération est la bande passante. Les opérations de copie s'effectuent généralement via des E/S séquentielles efficaces En guise d'estimation prudente, on suppose une utilisation de 50 % de la bande passante réseau disponible pour les opérations de copie. Par exemple, un port FC de 8 Go peut en théorie transférer environ 800 Mbit/s. Si l'on suppose une utilisation de 50 %, une base de données peut être copiée à un taux d'environ 400 Mbit/s. Ainsi, une base de données de 10 To peut être copiée en sept heures environ à ce rythme.

La migration sur de longues distances nécessite généralement une approche plus créative, comme le processus d'expédition des journaux expliqué dans "Déplacement du fichier de données en ligne". Les réseaux IP longue distance disposent rarement d'une bande passante proche des vitesses LAN ou SAN. Dans un cas, NetApp a participé à la migration à distance d'une base de données de 220 To avec des taux de génération de journaux d'archivage très élevés. L'approche choisie pour le transfert de données a été l'expédition quotidienne de bandes, parce que cette méthode offrait la bande passante maximale possible.

Nombre de bases de données

Dans de nombreux cas, le problème du déplacement d'une grande quantité de données n'est pas la taille des données, mais plutôt la complexité de la configuration qui prend en charge la base de données. Savoir qu'il faut migrer 50 To de bases de données n'est pas suffisant. Il peut s'agir d'une seule base de données stratégique de 50 To, d'un ensemble de 4 000 bases de données héritées ou d'un mélange de données de production et de données hors production. Dans certains cas, une grande partie des données est constituée de clones d'une base de données source. Il n'est pas nécessaire de migrer ces clones car ils peuvent être recréés facilement, notamment lorsque la nouvelle architecture est conçue pour exploiter les volumes NetApp FlexClone.

Pour la planification de la migration, vous devez connaître le nombre de bases de données concernées et leur priorité. À mesure que le nombre de bases de données augmente, l'option de migration privilégiée tend à être plus faible et plus faible dans la pile. Par exemple, la copie d'une seule base de données peut s'effectuer facilement avec RMAN et en cas de courte panne. Il s'agit de la réplication au niveau de l'hôte.

S'il existe 50 bases de données, il peut être plus facile d'éviter de configurer une nouvelle structure de système de fichiers pour recevoir une copie RMAN et de déplacer les données à la place. Ce processus peut être effectué en tirant parti de la migration LVM basée sur l'hôte pour déplacer les données des anciennes LUN vers les nouvelles LUN. L'équipe chargée de l'administration de la base de données (DBA) est alors détransférée vers l'équipe chargée du système d'exploitation pour que les données soient migrées de manière transparente par rapport à la base de données. La configuration du système de fichiers n'est pas modifiée.

Enfin, si 500 bases de données réparties sur 200 serveurs doivent être migrées, des options basées sur le stockage, telles que la fonctionnalité ONTAP Foreign LUN Import (FLI), peuvent être utilisées pour effectuer une migration directe des LUN.

Exigences en matière d'architecture

En général, l'organisation d'un fichier de base de données doit être modifiée pour exploiter les fonctionnalités de la nouvelle baie de stockage. Toutefois, ce n'est pas toujours le cas. Par exemple, les fonctionnalités des baies 100 % Flash EF-Series se concentrent sur les performances SAN et la fiabilité SAN. Dans la plupart des cas, les bases de données peuvent être migrées vers une baie EF-Series sans tenir compte particulière de la disposition des données. Les seules exigences sont un nombre élevé d'IOPS, une faible latence et une fiabilité robuste. Bien que certaines pratiques d'excellence soient liées à des facteurs tels que la configuration RAID ou les pools de disques dynamiques, les projets EF-Series nécessitent rarement des modifications importantes de l'architecture de stockage globale pour exploiter ces fonctionnalités.

En revanche, la migration vers ONTAP nécessite généralement une plus grande considération de la disposition de la base de données pour s'assurer que la configuration finale offre une valeur maximale. À elle seule, ONTAP offre de nombreuses fonctionnalités pour un environnement de base de données, même sans effort d'architecture spécifique. Plus important encore, il permet de migrer vers un nouveau matériel sans interruption lorsque le matériel actuel arrive en fin de vie. De manière générale, une migration vers ONTAP est la dernière migration que vous auriez à effectuer. Ensuite, le matériel est mis à niveau et les données sont migrées sans interruption vers de nouveaux supports.

Avec une certaine planification, davantage d'avantages sont disponibles. Les considérations les plus importantes concernent l'utilisation des snapshots. Les copies Snapshot sont la base des sauvegardes, des restaurations et des opérations de clonage quasi-instantanées. Comme exemple de la puissance des snapshots, l'utilisation la plus répandue concerne une base de données unique de 996 To qui s'exécute sur environ 250 LUN sur 6 contrôleurs. Cette base de données peut être sauvegardée en 2 minutes, restaurée en 2 minutes et clonée en 15 minutes. Les autres avantages sont la capacité à déplacer les données au sein du cluster en réponse aux modifications des charges de travail et les contrôles de qualité de service (QoS) appliqués pour fournir de bonnes performances cohérentes dans un environnement à plusieurs bases de données.

Les technologies comme les contrôles de qualité de service, la relocalisation des données, la copie Snapshot et le clonage fonctionnent dans presque toutes les configurations. Cependant, certains pensent généralement être nécessaires pour maximiser les avantages. Dans certains cas, les dispositions du stockage de la base de données peuvent nécessiter des modifications de conception afin d'optimiser l'investissement dans la nouvelle baie de stockage. De telles modifications de conception peuvent avoir un impact sur la stratégie de migration, car les migrations basées sur les hôtes ou sur le stockage répliquent la disposition des données d'origine. Des étapes supplémentaires peuvent être nécessaires pour mener à bien la migration et assurer une disposition des données optimisée pour ONTAP. Les procédures indiquées à la "Présentation des procédures de migration Oracle" vous pouvez par la suite présenter certaines méthodes qui vous permettent non seulement de migrer une base de données, mais aussi de la migrer vers la configuration finale optimale en un minimum d'efforts.

Délai de mise en service

Vous devez déterminer la durée maximale autorisée de l'interruption de service pendant la mise en service. C'est une erreur courante de supposer que l'ensemble du processus de migration provoque des perturbations. De nombreuses tâches peuvent être effectuées avant le début d'une interruption de service, et de nombreuses options permettent d'effectuer la migration sans interruption ni panne. Même si une interruption est inévitable, vous devez toujours définir le temps d'interruption de service maximal autorisé, car la durée de la mise en service varie d'une procédure à l'autre.

Par exemple, la copie d'une base de données de 10 To prend généralement environ sept heures. Si l'entreprise a besoin d'une interruption de service de sept heures, la copie de fichiers est une option simple et sûre pour la migration. Si cinq heures sont inacceptables, un simple processus d'envoi de journaux (voir "Envoi de journaux Oracle") peut être configuré en déployant un minimum d'efforts afin de réduire le délai de mise en service à environ 15 minutes. Pendant ce temps, un administrateur de base de données peut terminer le processus. Si 15 ce n'est pas le cas, le processus de mise en service final peut être automatisé par script afin de réduire le délai de mise en service à quelques minutes seulement. Vous pouvez toujours accélérer une migration, mais cette opération a un coût en temps et en efforts. Les délais de mise en service doivent être déterminés en fonction des objectifs acceptables pour l'entreprise.

Chemin de retour arrière

Aucune migration n'est totalement sans risque. Même si la technologie fonctionne parfaitement, il y a toujours une possibilité d'erreur de l'utilisateur. Le risque associé au chemin de migration choisi doit être pris en compte parallèlement aux conséquences d'un échec de la migration. Par exemple, la fonctionnalité de migration transparente du stockage en ligne d'Oracle ASM est l'une de ses principales fonctionnalités, et cette méthode est l'une des plus fiables connues. Cependant, les données sont copiées de manière irréversible avec cette méthode. Dans le cas peu probable où un problème se produit avec ASM, il n'y a pas de chemin de sortie simple. La seule option consiste à restaurer l'environnement d'origine ou à utiliser ASM pour restaurer la migration vers les LUN d'origine. Le risque peut être réduit, mais pas éliminé, en effectuant une sauvegarde de type Snapshot sur le système de stockage d'origine, à condition que le système soit capable d'effectuer une telle opération.

Répétition

Certaines procédures de migration doivent être entièrement vérifiées avant leur exécution. La nécessité d'une migration et d'une répétition du processus de mise en service est courante dans les bases de données stratégiques pour lesquelles la migration doit réussir et où les temps d'indisponibilité doivent être minimisés. En outre, les tests d'acceptation par l'utilisateur sont fréquemment inclus dans le travail de post-migration et le système global ne peut être remis en production qu'une fois ces tests terminés.

S'il est nécessaire de répéter, plusieurs fonctionnalités ONTAP peuvent faciliter le processus. En particulier, les snapshots peuvent réinitialiser un environnement de test et créer rapidement plusieurs copies compactes d'un environnement de base de données.

Procédures

Présentation

De nombreuses procédures sont disponibles pour la migration d'une base de données Oracle. Le bon dépend des besoins de votre entreprise.

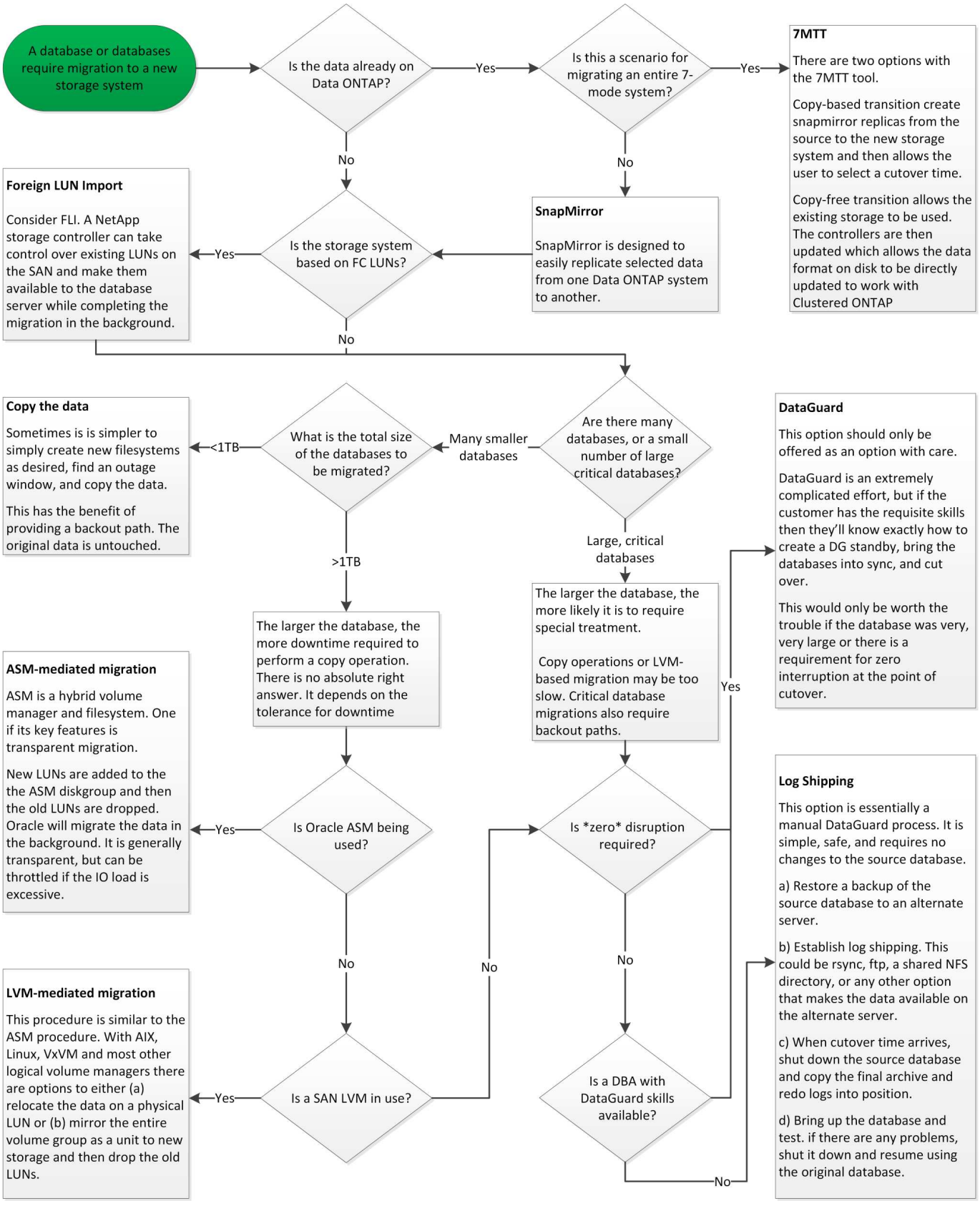
Dans de nombreux cas, les administrateurs système et les administrateurs de bases de données utilisent leurs propres méthodes de déplacement des données de volume physique, de mise en miroir et de déréplication, ou d'utilisation d'Oracle RMAN pour la copie des données.

Ces procédures sont fournies principalement à titre de conseils pour le personnel INFORMATIQUE qui connaît moins bien certaines des options disponibles. En outre, ces procédures illustrent les tâches, les exigences en termes de temps et les besoins en compétences de chaque approche de migration. Ainsi, d'autres parties, telles que NetApp et les services professionnels partenaires ou la direction INFORMATIQUE, peuvent mieux apprécier les exigences de chaque procédure.

Il n'existe pas de meilleure pratique unique pour créer une stratégie de migration. Pour créer un plan, il faut d'abord comprendre les options de disponibilité, puis sélectionner la méthode la mieux adaptée aux besoins de l'entreprise. La figure ci-dessous illustre les considérations de base et les conclusions types des clients, mais elle n'est pas universellement applicable à toutes les situations.

Par exemple, une étape soulève le problème de la taille totale de la base de données. L'étape suivante dépend si la base de données est supérieure ou inférieure à 1 To. Les étapes recommandées sont précisément des recommandations basées sur les pratiques standard des clients. La plupart des clients n'utiliseraient pas DataGuard pour copier une petite base de données, mais d'autres pourraient le faire. La plupart des clients ne tenteraient pas de copier une base de données de 50 To en raison du temps nécessaire, mais certaines peuvent disposer d'une fenêtre de maintenance suffisamment longue pour permettre une telle opération.

L'organigramme ci-dessous présente les différents éléments à prendre en compte pour déterminer le chemin de migration le plus approprié. Vous pouvez cliquer avec le bouton droit de la souris sur l'image et l'ouvrir dans un nouvel onglet pour améliorer la lisibilité.



.

Déplacement du fichier de données en ligne

Oracle 12cR1 et versions supérieures incluent la possibilité de déplacer un fichier de données pendant que la base de données reste en ligne. Il fonctionne en outre entre différents types de systèmes de fichiers. Par

exemple, un fichier de données peut être déplacé d'un système de fichiers xfs vers ASM. Cette méthode n'est généralement pas utilisée à grande échelle en raison du nombre d'opérations de déplacement de fichiers de données individuelles qui seraient requises. Toutefois, il est important de tenir compte de cette méthode avec des bases de données plus petites et moins de fichiers de données.

En outre, le simple déplacement d'un fichier de données est une bonne option pour migrer des parties de bases de données existantes. Par exemple, les fichiers de données moins actifs peuvent être transférés vers un stockage plus économique, tel qu'un volume FabricPool qui peut stocker les blocs inactifs dans le magasin d'objets.

Migration au niveau de la base de données

La migration au niveau de la base de données signifie que la base de données peut déplacer des données. Plus précisément, cela signifie l'envoi de journaux. Des technologies telles que RMAN et ASM sont des produits Oracle, mais pour la migration, elles fonctionnent au niveau de l'hôte où elles copient les fichiers et gèrent les volumes.

Envoi de journaux

La base de la migration au niveau de la base de données est le journal d'archivage Oracle, qui contient un journal des modifications apportées à la base de données. La plupart du temps, un journal d'archivage fait partie d'une stratégie de sauvegarde et de restauration. Le processus de restauration commence par la restauration d'une base de données, puis la relecture d'un ou plusieurs journaux d'archivage pour ramener la base de données à l'état souhaité. Cette même technologie de base peut être utilisée pour effectuer une migration avec une interruption des opérations nulle ou minime. Plus important encore, cette technologie permet la migration tout en conservant la base de données d'origine intacte, ce qui permet de conserver un chemin de retour.

Le processus de migration commence par la restauration d'une sauvegarde de base de données sur un serveur secondaire. Vous pouvez le faire de différentes manières, mais la plupart des clients utilisent leur application de sauvegarde normale pour restaurer les fichiers de données. Une fois les fichiers de données restaurés, les utilisateurs établissent une méthode d'envoi des journaux. L'objectif est de créer un flux constant de journaux d'archivage générés par la base de données primaire et de les relire sur la base de données restaurée afin de les conserver dans un état similaire. Lorsque le délai de mise en service arrive, la base de données source est complètement arrêtée et les journaux d'archivage finaux, et dans certains cas les journaux de reprise, sont copiés et relus. Il est essentiel que les journaux de reprise soient également pris en compte, car ils peuvent contenir certaines des transactions finales validées.

Une fois ces journaux transférés et relus, les deux bases de données sont cohérentes l'une avec l'autre. À ce stade, la plupart des clients effectuent des tests de base. Si des erreurs sont commises pendant le processus de migration, la relecture du journal doit signaler les erreurs et échouer. Il est toujours conseillé d'effectuer des tests rapides basés sur des requêtes connues ou des activités applicatives pour vérifier que la configuration est optimale. Il est également courant de créer une table de test finale avant d'arrêter la base de données d'origine pour vérifier qu'elle est présente dans la base de données migrée. Cette étape permet de s'assurer qu'aucune erreur n'a été effectuée lors de la synchronisation finale du journal.

Une simple migration d'envoi de journaux peut être configurée hors bande par rapport à la base de données d'origine, ce qui la rend particulièrement utile pour les bases de données stratégiques. Il n'est pas nécessaire de modifier la configuration de la base de données source, car la restauration et la configuration initiale de l'environnement de migration n'affectent pas les opérations de production. Une fois l'envoi de journaux configuré, il impose des demandes d'E/S sur les serveurs de production. Cependant, l'envoi de journaux se compose de simples lectures séquentielles des journaux d'archivage, qui n'ont probablement aucun impact sur les performances des bases de données de production.

L'expédition de journaux s'est avérée particulièrement utile pour les projets de migration longue distance à

taux de changement élevé. Dans un cas, une seule base de données de 220 To a été migrée vers un nouvel emplacement situé à environ 500 kilomètres. Le taux de modification était extrêmement élevé et les restrictions de sécurité empêchaient l'utilisation d'une connexion réseau. L'expédition des journaux a été effectuée à l'aide de bandes et de coursiers. Une copie de la base de données source a d'abord été restaurée à l'aide des procédures décrites ci-dessous. Les journaux ont ensuite été expédiés chaque semaine par messagerie jusqu'au moment de la mise en service, lorsque le jeu final de bandes a été livré et que les journaux ont été appliqués à la base de données de réplica.

Oracle DataGuard

Dans certains cas, un environnement DataGuard complet est garanti. Il est incorrect d'utiliser le terme DataGuard pour faire référence à toute configuration d'envoi de journaux ou de base de données de secours. Oracle DataGuard est un framework complet de gestion de la réplication de base de données, mais il ne s'agit pas d'une technologie de réplication. Le principal avantage d'un environnement DataGuard complet dans un effort de migration est le basculement transparent d'une base de données à une autre. DataGuard permet également un basculement transparent vers la base de données d'origine en cas de problème, tel qu'un problème de performances ou de connectivité réseau avec le nouvel environnement. Un environnement DataGuard entièrement configuré nécessite la configuration non seulement de la couche de base de données, mais aussi des applications pour que les applications puissent détecter un changement dans l'emplacement de la base de données primaire. En général, il n'est pas nécessaire d'utiliser DataGuard pour effectuer une migration, mais certains clients possèdent une expertise DataGuard étendue en interne et en dépendent déjà pour le travail de migration.

Architecture

Comme évoqué précédemment, l'exploitation des fonctionnalités avancées des baies de stockage nécessite parfois de modifier l'organisation de la base de données. De plus, une modification du protocole de stockage, telle que le passage d'ASM à un système de fichiers NFS, modifie nécessairement la disposition du système de fichiers.

L'un des principaux avantages des méthodes d'envoi de journaux, y compris DataGuard, est que la destination de réplication ne doit pas correspondre à la source. Il n'y a pas de problème avec l'utilisation d'une approche d'envoi de journaux pour migrer d'ASM vers un système de fichiers standard, et inversement. La disposition précise des fichiers de données peut être modifiée à la destination pour optimiser l'utilisation de la technologie de base de données enfichable (PDB) ou pour définir des contrôles QoS de manière sélective sur certains fichiers. En d'autres termes, un processus de migration basé sur l'envoi de journaux vous permet d'optimiser facilement et en toute sécurité l'organisation du stockage de la base de données.

Ressources du serveur

La migration au niveau de la base de données est limitée par le besoin d'un second serveur. Ce second serveur peut être utilisé de deux manières :

1. Vous pouvez utiliser le second serveur comme nouveau domicile permanent pour la base de données.
2. Vous pouvez utiliser le second serveur comme serveur temporaire de transfert. Une fois la migration des données vers la nouvelle baie de stockage terminée et testée, les systèmes de fichiers LUN ou NFS sont déconnectés du serveur intermédiaire et reconnectés au serveur d'origine.

La première option est la plus simple, mais son utilisation peut ne pas être possible dans les environnements très vastes nécessitant des serveurs très puissants. La deuxième option nécessite un travail supplémentaire pour replacer les systèmes de fichiers à leur emplacement d'origine. Il peut s'agir d'une opération simple dans laquelle NFS est utilisé comme protocole de stockage car les systèmes de fichiers peuvent être démontés du serveur de transfert et remontés sur le serveur d'origine.

Les systèmes de fichiers basés sur les blocs nécessitent un travail supplémentaire pour mettre à jour le zoning FC ou les initiateurs iSCSI. Avec la plupart des gestionnaires de volumes logiques (y compris ASM), les LUN sont automatiquement détectées et mises en ligne après leur mise à disposition sur le serveur d'origine. Cependant, certaines implémentations de système de fichiers et de LVM peuvent nécessiter davantage de travail pour exporter et importer les données. La procédure précise peut varier, mais il est généralement facile d'établir une procédure simple et reproductible pour terminer la migration et réexécuter les données sur le serveur d'origine.

Bien qu'il soit possible de configurer l'envoi de journaux et de répliquer une base de données dans un environnement de serveur unique, la nouvelle instance doit avoir un SID de processus différent pour pouvoir relire les journaux. Il est possible d'afficher temporairement la base de données sous un autre ensemble d'ID de processus avec un SID différent et de la modifier ultérieurement. Toutefois, cela peut entraîner de nombreuses activités de gestion complexes et mettre l'environnement de base de données en danger d'erreur de la part des utilisateurs.

Migration au niveau de l'hôte

La migration des données au niveau de l'hôte implique l'utilisation du système d'exploitation hôte et des utilitaires associés pour terminer la migration. Ce processus inclut tout utilitaire qui copie les données, y compris Oracle RMAN et Oracle ASM.

Copie de données

La valeur d'une opération de copie simple ne doit pas être sous-estimée. Les infrastructures réseau modernes peuvent déplacer des données à un taux de gigaoctets par seconde. Les opérations de copie de fichiers reposent sur des E/S efficaces en lecture et écriture séquentielles Si une opération de copie de l'hôte est plus perturbant que l'envoi de journaux, la migration ne se limite pas au déplacement des données. Elle inclut généralement les modifications apportées au réseau, au délai de redémarrage de la base de données et aux tests de post-migration.

Le temps réel nécessaire à la copie des données peut ne pas être important. En outre, une opération de copie préserve un chemin de retour garanti, car les données d'origine ne sont pas modifiées. En cas de problème pendant le processus de migration, les systèmes de fichiers d'origine avec les données d'origine peuvent être réactivés.

Changement de plate-forme

Le changement de plate-forme fait référence à un changement de type de CPU. Lorsqu'une base de données est migrée d'une plate-forme Solaris, AIX ou HP-UX traditionnelle vers Linux x86, les données doivent être reformatées en raison de modifications de l'architecture CPU. Les processeurs SPARC, IA64 et POWER sont connus sous le nom de processeurs big endian, tandis que les architectures x86 et x86_64 sont connues sous le nom de Little endian. Par conséquent, certaines données des fichiers de données Oracle sont triées différemment selon le processeur utilisé.

Jusqu'ici, les clients ont généralement utilisé DataPump pour répliquer des données sur plusieurs plateformes. DataPump est un utilitaire qui crée un type spécial d'exportation de données logiques qui peut être importé plus rapidement dans la base de données de destination. Comme il crée une copie logique des données, DataPump laisse derrière lui les dépendances de l'endianness du processeur. DataPump est encore utilisé par certains clients pour le changement de plateforme, mais une option plus rapide est désormais disponible avec Oracle 11g : les tablespaces interplateformes transportables. Cette avance permet de convertir un espace de table en un format endian différent. Il s'agit d'une transformation physique qui offre de meilleures performances qu'une exportation DataPump, qui doit convertir les octets physiques en données logiques, puis les convertir en octets physiques.

Une discussion complète sur DataPump et les tablespaces transportables va au-delà de la documentation

NetApp portée, mais NetApp propose quelques recommandations basées sur notre expérience d'assistance aux clients lors de la migration vers une nouvelle baie de stockage dans le cadre d'une nouvelle architecture de processeur :

- Si DataPump est utilisé, le temps nécessaire à la migration doit être mesuré dans un environnement de test. Les clients sont parfois surpris du temps nécessaire à la réalisation de la migration. Cette interruption supplémentaire imprévue peut provoquer des interruptions.
- De nombreux clients pensent à tort que les tablespaces transportables multi plates-formes ne nécessitent pas de conversion de données. Lorsqu'une CPU avec un autre endian est utilisée, un RMAN `convert` l'opération doit être effectuée au préalable sur les fichiers de données. Cette opération n'est pas instantanée. Dans certains cas, le processus de conversion peut être accéléré en ayant plusieurs threads fonctionnant sur différents fichiers de données, mais le processus de conversion ne peut pas être évité.

Migration basée sur le gestionnaire de volumes logiques

Les LVM fonctionnent en déregroupant un groupe d'une ou de plusieurs LUN en petites unités généralement appelées extensions. Le pool d'extensions est ensuite utilisé comme source pour créer des volumes logiques qui sont essentiellement virtualisés. Cette couche de virtualisation apporte de la valeur de plusieurs manières :

- Les volumes logiques peuvent utiliser des extensions tirées de plusieurs LUN. Lorsqu'un système de fichiers est créé sur un volume logique, il peut exploiter les performances maximales de toutes les LUN. Il favorise également le chargement homogène de toutes les LUN du groupe de volumes, pour des performances plus prévisibles.
- Les volumes logiques peuvent être redimensionnés en ajoutant et, dans certains cas, en supprimant des extensions. Le redimensionnement d'un système de fichiers sur un volume logique s'effectue généralement sans interruption.
- Le déplacement des extensions sous-jacentes permet de migrer les volumes logiques sans interruption.

La migration à l'aide d'un LVM fonctionne de deux manières : déplacer une extension ou mettre en miroir/démirroring une extension. La migration des LVM utilise des E/S séquentielles de blocs de grande taille efficaces et pose rarement des problèmes de performances. Si ce problème survient, il existe généralement des options pour limiter le taux d'E/S. Cela augmente le temps nécessaire à la migration, tout en réduisant la charge d'E/S sur l'hôte et les systèmes de stockage.

Miroir et démiroir

Certains gestionnaires de volumes, tels que AIX LVM, permettent à l'utilisateur de spécifier le nombre de copies pour chaque extension et de contrôler les périphériques qui hébergent chaque copie. La migration s'effectue par la mise en miroir d'un volume logique existant sur les extensions sous-jacentes des nouveaux volumes, l'attente de la synchronisation des copies, puis l'abandon de l'ancienne copie. Si un chemin de retour arrière est souhaité, un instantané des données d'origine peut être créé avant le point de suppression de la copie miroir. Il est également possible d'arrêter brièvement le serveur pour masquer les LUN d'origine avant de forcer la suppression des copies miroir contenues. Cela permet de conserver une copie récupérable des données à leur emplacement d'origine.

Migration d'extension

La plupart des gestionnaires de volumes permettent la migration des extensions, et il arrive parfois que plusieurs options existent. Par exemple, certains gestionnaires de volumes permettent à un administrateur de déplacer les extensions individuelles d'un volume logique spécifique de l'ancien vers le nouveau stockage. Les gestionnaires de volumes tels que Linux LVM2 offrent le `pvmove` Qui déplace toutes les extensions du périphérique LUN spécifié vers une nouvelle LUN. Une fois l'ancien LUN évacué, il est possible de le retirer.

Le risque principal pour les opérations est la suppression des anciennes LUN inutilisées de la configuration. Une attention toute particulière doit être portée au changement de segmentation FC et au retrait des périphériques LUN obsolètes.

Gestion automatique du stockage par Oracle

Oracle ASM est un gestionnaire de volumes logiques et un système de fichiers combinés. À un niveau élevé, Oracle ASM prend un ensemble de LUN, les répartit en petites unités d'allocation et les présente comme un seul volume appelé groupe de disques ASM. ASM permet également de mettre en miroir le groupe de disques en définissant le niveau de redondance. Un volume peut être sans miroir (redondance externe), en miroir (redondance normale) ou en miroir tridirectionnel (redondance élevée). La configuration du niveau de redondance doit être effectuée avec précaution car il ne peut pas être modifié après sa création.

ASM fournit également des fonctionnalités de système de fichiers. Bien que le système de fichiers ne soit pas visible directement depuis l'hôte, la base de données Oracle peut créer, déplacer et supprimer des fichiers et des répertoires sur un groupe de disques ASM. Vous pouvez également naviguer dans la structure à l'aide de l'utilitaire asmcmd.

Comme pour les autres implémentations LVM, Oracle ASM optimise les performances d'E/S en segmentant et en équilibrant les E/S de chaque fichier sur l'ensemble des LUN disponibles. Deuxièmement, les extensions sous-jacentes peuvent être déplacées pour permettre le redimensionnement du groupe de disques ASM ainsi que la migration. Oracle ASM automatise le processus tout au long de l'opération de rééquilibrage. Les nouvelles LUN sont ajoutées à un groupe de disques ASM et les anciennes LUN sont abandonnées, ce qui déclenche le déplacement d'extension et le DROP suivant de la LUN évacuée du groupe de disques. Ce processus est l'une des méthodes de migration les plus éprouvées, et la fiabilité d'ASM pour assurer une migration transparente est probablement sa fonctionnalité la plus importante.

Comme le niveau de mise en miroir d'Oracle ASM est fixe, il ne peut pas être utilisé avec la méthode de migration miroir et démiroir.

Migration au niveau du stockage

La migration au niveau du stockage implique d'effectuer la migration au-dessous des niveaux des applications et du système d'exploitation. Auparavant, il fallait parfois utiliser des périphériques spécialisés qui copiaient les LUN au niveau du réseau, mais ces fonctionnalités sont désormais natives dans ONTAP.

SnapMirror

La migration de bases de données entre des systèmes NetApp est presque effectuée de manière universelle avec le logiciel de réplication des données NetApp SnapMirror. Ce processus implique la configuration d'une relation de miroir pour les volumes à migrer, leur permettant ainsi de se synchroniser, puis d'attendre la fenêtre de mise en service. Lorsqu'elle arrive, la base de données source est arrêtée, une dernière mise à jour miroir est effectuée et le miroir est cassé. Les volumes de réplica sont alors prêts à l'emploi, soit en montant un répertoire de système de fichiers NFS contenu, soit en découvrant les LUN contenues et en démarrant la base de données.

La relocalisation des volumes dans un seul cluster ONTAP n'est pas considérée comme une migration, mais plutôt comme une routine `volume move` fonctionnement. SnapMirror est utilisé en tant que moteur de réplication des données au sein du cluster. Ce processus est entièrement automatisé. Il n'y a pas d'étape de migration supplémentaire à effectuer lorsque les attributs du volume, tels que le mappage de LUN ou les autorisations d'exportation NFS, sont déplacés avec le volume lui-même. La relocalisation ne prend pas en charge l'hôte. Dans certains cas, il convient de mettre à jour l'accès au réseau pour s'assurer que les données nouvellement déplacées sont accessibles de la manière la plus efficace possible, mais sans interruption.

Importation de LUN étrangères (FLI)

La FLI est une fonctionnalité qui permet à un système Data ONTAP exécutant la version 8.3 ou supérieure de migrer un LUN existant à partir d'une autre baie de stockage. La procédure est simple : le système ONTAP est zoné sur la baie de stockage existante comme s'il s'agissait d'un autre hôte SAN. Data ONTAP prend alors le contrôle des LUN héritées souhaitées et migre les données sous-jacentes. De plus, le processus d'importation utilise les paramètres d'efficacité du nouveau volume lors de la migration des données. Ainsi, les données peuvent être compressées et dédupliquées en ligne pendant le processus de migration.

La première implémentation de FLI dans Data ONTAP 8.3 a permis uniquement la migration hors ligne. Ce transfert était extrêmement rapide, mais cela signifiait que les données de LUN étaient indisponibles jusqu'à la fin de la migration. La migration en ligne a été introduite dans Data ONTAP 8.3.1. Ce type de migration minimise les interruptions en permettant à ONTAP de transmettre des données LUN lors du processus de transfert. Il y a une brève interruption lors de la remise en place de l'hôte pour l'utilisation des LUN via ONTAP. Cependant, dès que ces modifications sont apportées, les données sont de nouveau accessibles et restent accessibles tout au long du processus de migration.

Les E/S de lecture sont proxées via ONTAP jusqu'à la fin de l'opération de copie, tandis que les E/S d'écriture sont écrites de manière synchrone sur les LUN étrangères et ONTAP. Les deux copies LUN sont ainsi synchronisées jusqu'à ce que l'administrateur exécute une mise en service complète qui libère le LUN étranger et ne réplique plus les écritures.

FLI est conçu pour fonctionner avec FC. Toutefois, si vous souhaitez passer à iSCSI, le LUN migré peut facilement être remappé en tant que LUN iSCSI une fois la migration terminée.

Parmi les caractéristiques de FLI figurent la détection et le réglage automatiques de l'alignement. Dans ce contexte, le terme alignement fait référence à une partition sur un périphérique LUN. Pour des performances optimales, les E/S doivent être alignées sur des blocs de 4 Ko. Si une partition est placée à un décalage qui n'est pas un multiple de 4K, les performances en pâtissent.

Il existe un deuxième aspect de l'alignement qui ne peut pas être corrigé en réglant un décalage de partition, c'est-à-dire la taille du bloc du système de fichiers. Par exemple, un système de fichiers ZFS prend généralement par défaut une taille de bloc interne de 512 octets. D'autres clients utilisant AIX ont parfois créé des systèmes de fichiers jfs2 avec une taille de bloc de 512 ou 1, 024 octets. Bien que le système de fichiers puisse être aligné sur une limite de 4 Ko, les fichiers créés dans ce système de fichiers ne le sont pas et les performances en pâtissent.

FLI ne doit pas être utilisé dans ces circonstances. Bien que les données soient accessibles après la migration, vous obtenez des systèmes de fichiers avec de graves limitations de performances. En principe, tout système de fichiers prenant en charge une charge de travail de remplacement aléatoire sur ONTAP doit utiliser une taille de bloc de 4 Ko. Cela s'applique principalement aux charges de travail telles que les fichiers de données de base de données et les déploiements VDI. La taille de bloc peut être identifiée à l'aide des commandes appropriées du système d'exploitation hôte.

Par exemple, sous AIX, la taille de bloc peut être affichée avec `lsfs -q`. Avec Linux, `xfs_info` et `tune2fs` peut être utilisé pour `xfs` et `ext3/ext4`, respectivement. Avec `zfs`, la commande est `zdb -C`.

Le paramètre qui contrôle la taille du bloc est `ashift` et la valeur par défaut est généralement 9, soit 2^9, ou 512 octets. Pour des performances optimales, le `ashift` La valeur doit être 12 (2^12=4K). Cette valeur est définie au moment de la création du zpool et ne peut pas être modifiée, ce qui signifie que les zpools de données avec un `ashift` une migration autre que 12 doit être effectuée en copiant les données vers un nouveau zpool.

Oracle ASM n'a pas de taille de bloc fondamentale. La seule exigence est que la partition sur laquelle le disque ASM est construit doit être correctement alignée.

Outil de transition 7-mode

L'outil 7-mode transition Tool (7MTT) est un utilitaire d'automatisation utilisé pour migrer de grandes configurations 7-mode vers ONTAP. La plupart des clients de bases de données trouvent d'autres méthodes plus faciles, notamment parce qu'ils migrent généralement leurs environnements de bases de données par base de données plutôt que de déplacer l'intégralité de l'empreinte du stockage. De plus, les bases de données ne font souvent partie que d'un environnement de stockage plus important. Les bases de données sont donc souvent migrées individuellement, puis le reste de l'environnement peut être déplacé avec 7MTT.

Les clients sont de petite taille, mais nombreux. Ils disposent de systèmes de stockage dédiés à des environnements de base de données complexes. Ces environnements peuvent contenir de nombreux volumes, snapshots et de nombreuses informations de configuration telles que les autorisations d'exportation, les groupes initiateurs de LUN, les autorisations utilisateur et la configuration du protocole d'accès aux répertoires légers. Dans de tels cas, les fonctionnalités d'automatisation de l'outil 7MTT simplifient considérablement la migration.

7MTT peut fonctionner dans deux modes :

- **Transition basée sur les copies (CBT).** dans le nouvel environnement, l'outil 7MTT avec CBT configure les volumes SnapMirror à partir d'un système 7- mode existant. Une fois les données synchronisées, l'outil 7MTT orchestre le processus de mise en service.
- **Transition sans copie.** 7MTT avec la transition sans copie repose sur la conversion des tiroirs disques 7-mode existants sans déplacement des données. Aucune donnée n'est copiée et les tiroirs disques existants peuvent être réutilisés. La protection des données et la configuration de l'efficacité du stockage existantes sont préservées.

La différence principale entre ces deux options est que la transition sans copie constitue une approche globale où tous les tiroirs disques rattachés à la paire HA 7-mode d'origine doivent être transférés vers le nouvel environnement. Il n'existe aucune option pour déplacer un sous-ensemble de tiroirs. L'approche basée sur les copies permet de déplacer des volumes sélectionnés. Par ailleurs, une fenêtre de mise en service peut être plus longue et la transition sans copie est liée à l'alignement des tiroirs disques et à la conversion des métadonnées. En fonction de son expérience sur le terrain, NetApp recommande de consacrer 1 heure au déplacement et à la réinstallation des tiroirs disques, et entre 15 minutes et 2 heures à la conversion des métadonnées.

Migration des fichiers de données

Vous pouvez déplacer individuellement les fichiers de données Oracle via une seule commande.

Par exemple, la commande suivante déplace le fichier de données IOPST.dbf du système de fichiers `/oradata2` vers le système de fichiers `/oradata3`.

```
SQL> alter database move datafile  '/oradata2/NTAP/IOPS002.dbf' to
'/oradata3/NTAP/IOPS002.dbf';
Database altered.
```

Le déplacement d'un fichier de données avec cette méthode peut être lent, mais il ne doit normalement pas produire suffisamment d'E/S pour interférer avec les charges de travail quotidiennes des bases de données. En revanche, la migration via le rééquilibrage d'ASM peut s'exécuter beaucoup plus rapidement, mais au détriment du ralentissement de la base de données globale pendant le déplacement des données.

Le temps nécessaire à la migration des fichiers de données peut être mesuré en créant un fichier de données de test et en le déplaçant. Le temps écoulé pour l'opération est enregistré dans les données v$session :

```
SQL> set linesize 300;
SQL> select elapsed_seconds||':'||message from v$session_longops;
ELAPSED_SECONDS||':'||MESSAGE
------------------------------------------------------------------------
---------------
351:Online data file move: data file 8: 22548578304 out of 22548578304
bytes done
SQL> select bytes / 1024 / 1024 /1024 as GB from dba_data_files where
FILE_ID = 8;
        GB
----------
        21
```

Dans cet exemple, le fichier déplacé était le fichier de données 8, dont la taille était de 21 Go et dont la migration nécessitait environ 6 minutes. Le temps nécessaire dépend évidemment des capacités du système de stockage, du réseau de stockage et de l'activité globale de la base de données au moment de la migration.

Envoi de journaux

L'objectif d'une migration à l'aide de l'envoi de journaux est de créer une copie des fichiers de données d'origine à un nouvel emplacement, puis d'établir une méthode d'expédition des modifications dans le nouvel environnement.

Une fois établie, l'envoi et la relecture des journaux peuvent être automatisés afin de maintenir la base de données de réplica largement synchronisée avec la source. Par exemple, une tâche cron peut être planifiée pour (a) copier les journaux les plus récents vers le nouvel emplacement et (b) les relire toutes les 15 minutes. L'interruption au moment de la mise en service est ainsi minimale, car la lecture des journaux d'archivage ne doit pas dépasser 15 minutes.

La procédure présentée ci-dessous est également essentiellement une opération de clonage de base de données. La logique illustrée est similaire au moteur de NetApp SnapManager pour Oracle (SMO) et du plug-in Oracle NetApp SnapCenter. Certains clients ont utilisé la procédure présentée dans des scripts ou des workflows WFA pour des opérations de clonage personnalisé. Bien que cette procédure soit plus manuelle qu'avec SMO ou SnapCenter, elle reste facilement scriptée, et les API de gestion des données de ONTAP simplifient davantage le processus.

Envoi de journaux - système de fichiers vers le système de fichiers

Cet exemple illustre la migration d'une base de données appelée WAFFLE d'un système de fichiers ordinaire vers un autre système de fichiers ordinaire situé sur un serveur différent. Il illustre également l'utilisation de SnapMirror pour effectuer une copie rapide des fichiers de données, mais cela ne fait pas partie intégrante de la procédure globale.

Créer une sauvegarde de base de données

La première étape consiste à créer une sauvegarde de base de données. Plus précisément, cette procédure nécessite un ensemble de fichiers de données pouvant être utilisés pour la relecture des journaux d'archivage.

De production

Dans cet exemple, la base de données source se trouve sur un système ONTAP. La méthode la plus simple pour créer une sauvegarde d'une base de données consiste à utiliser un instantané. La base de données est placée en mode de sauvegarde à chaud pendant quelques secondes `snapshot create` l'opération est exécutée sur le volume hébergeant les fichiers de données.

```
SQL> alter database begin backup;
Database altered.
```

```
Cluster01::*> snapshot create -vserver vserver1 -volume jfsc1_oradata
hotbackup
Cluster01::*>
```

```
SQL> alter database end backup;
Database altered.
```

Le résultat est un instantané sur le disque appelé `hotbackup` qui contient une image des fichiers de données en mode de sauvegarde à chaud. Lorsqu'elles sont combinées avec les journaux d'archivage appropriés pour assurer la cohérence des fichiers de données, les données de cet instantané peuvent servir de base à une restauration ou à un clone. Dans ce cas, il est répliqué sur le nouveau serveur.

Restaurer dans un nouvel environnement

La sauvegarde doit maintenant être restaurée dans le nouvel environnement. Cette opération peut être effectuée de plusieurs façons, notamment Oracle RMAN, la restauration à partir d'une application de sauvegarde comme NetBackup ou une simple opération de copie des fichiers de données placés en mode de sauvegarde à chaud.

Dans cet exemple, SnapMirror est utilisé pour répliquer la sauvegarde à chaud de snapshot vers un nouvel emplacement.

1. Créez un volume pour recevoir les données de snapshot. Initialiser la mise en miroir à partir de `jfsc1_oradata` à `vol_oradata`.

   ```
   Cluster01::*> volume create -vserver vserver1 -volume vol_oradata
   -aggregate data_01 -size 20g -state online -type DP -snapshot-policy
   none -policy jfsc3
   [Job 833] Job succeeded: Successful
   ```

```
Cluster01::*> snapmirror initialize -source-path vserver1:jfsc1_oradata
-destination-path vserver1:vol_oradata
Operation is queued: snapmirror initialize of destination
"vserver1:vol_oradata".
Cluster01::*> volume mount -vserver vserver1 -volume vol_oradata
-junction-path /vol_oradata
Cluster01::*>
```

2. Une fois l'état défini par SnapMirror, indiquant que la synchronisation est terminée, mettre à jour le miroir en fonction du snapshot souhaité.

```
Cluster01::*> snapmirror show -destination-path vserver1:vol_oradata
-fields state
source-path             destination-path        state
----------------------- ----------------------- ------------
vserver1:jfsc1_oradata vserver1:vol_oradata SnapMirrored
```

```
Cluster01::*> snapmirror update -destination-path vserver1:vol_oradata
-source-snapshot hotbackup
Operation is queued: snapmirror update of destination
"vserver1:vol_oradata".
```

3. La synchronisation peut être vérifiée en affichant le `newest-snapshot` champ sur le volume miroir.

```
Cluster01::*> snapmirror show -destination-path vserver1:vol_oradata
-fields newest-snapshot
source-path             destination-path        newest-snapshot
----------------------- ----------------------- ---------------
vserver1:jfsc1_oradata vserver1:vol_oradata hotbackup
```

4. Le miroir peut alors être cassé.

```
Cluster01::> snapmirror break -destination-path vserver1:vol_oradata
Operation succeeded: snapmirror break for destination
"vserver1:vol_oradata".
Cluster01::>
```

5. Montez le nouveau système de fichiers.avec les systèmes de fichiers en mode bloc, les procédures précises varient en fonction du LVM utilisé. Le zoning FC ou les connexions iSCSI doivent être configurés. Une fois la connectivité aux LUN établie, des commandes telles que Linux `pvscan` Il peut être nécessaire de déterminer quels groupes de volumes ou LUN doivent être configurés correctement pour être détectables par ASM.

Dans cet exemple, un simple système de fichiers NFS est utilisé. Ce système de fichiers peut être monté directement.

```
fas8060-nfs1:/vol_oradata        19922944   1639360   18283584   9%
/oradata
fas8060-nfs1:/vol_logs            9961472       128    9961344   1%
/logs
```

Créer un modèle de création de fichier de contrôle

Vous devez ensuite créer un modèle de fichier de contrôle. Le `backup controlfile to trace` commande crée des commandes texte pour recréer un fichier de contrôle. Dans certaines circonstances, cette fonction peut être utile pour restaurer une base de données à partir d'une sauvegarde, et elle est souvent utilisée avec des scripts qui effectuent des tâches telles que le clonage de base de données.

1. Le résultat de la commande suivante est utilisé pour recréer les fichiers de contrôle pour la base de données migrée.

```
SQL> alter database backup controlfile to trace as '/tmp/waffle.ctrl';
Database altered.
```

2. Une fois les fichiers de contrôle créés, copiez-les sur le nouveau serveur.

```
[oracle@jfsc3 tmp]$ scp oracle@jfsc1:/tmp/waffle.ctrl /tmp/
oracle@jfsc1's password:
waffle.ctrl                                           100% 5199
5.1KB/s   00:00
```

Sauvegarde du fichier de paramètres

Un fichier de paramètres est également requis dans le nouvel environnement. La méthode la plus simple consiste à créer un fichier pfile à partir du fichier spfile ou pfile actuel. Dans cet exemple, la base de données source utilise un fichier spfile.

```
SQL> create pfile='/tmp/waffle.tmp.pfile' from spfile;
File created.
```

Créer une entrée oratab

La création d'une entrée oratab est requise pour le bon fonctionnement des utilitaires tels que oraenv. Pour créer une entrée oratab, procédez comme suit.

```
WAFFLE:/orabin/product/12.1.0/dbhome_1:N
```

Préparer la structure du répertoire

Si les répertoires requis n'étaient pas déjà présents, vous devez les créer ou la procédure de démarrage de la base de données échoue. Pour préparer la structure de répertoires, remplissez les conditions minimales suivantes.

```
[oracle@jfsc3 ~]$ . oraenv
ORACLE_SID = [oracle] ? WAFFLE
The Oracle base has been set to /orabin
[oracle@jfsc3 ~]$ cd $ORACLE_BASE
[oracle@jfsc3 orabin]$ cd admin
[oracle@jfsc3 admin]$ mkdir WAFFLE
[oracle@jfsc3 admin]$ cd WAFFLE
[oracle@jfsc3 WAFFLE]$ mkdir adump dpdump pfile scripts xdb_wallet
```

Mises à jour du fichier de paramètres

1. Pour copier le fichier de paramètres sur le nouveau serveur, exécutez les commandes suivantes. L'emplacement par défaut est le `$ORACLE_HOME/dbs` répertoire. Dans ce cas, le fichier pfile peut être placé n'importe où. Il est utilisé uniquement comme étape intermédiaire dans le processus de migration.

```
[oracle@jfsc3 admin]$ scp oracle@jfsc1:/tmp/waffle.tmp.pfile
$ORACLE_HOME/dbs/waffle.tmp.pfile
oracle@jfsc1's password:
waffle.pfile                                          100%  916
0.9KB/s   00:00
```

1. Modifiez le fichier selon vos besoins. Par exemple, si l'emplacement du journal d'archive a changé, le fichier pfile doit être modifié pour refléter le nouvel emplacement. Dans cet exemple, seuls les fichiers de contrôle sont déplacés, en partie pour les distribuer entre les systèmes de fichiers journaux et de données.

```
[root@jfsc1 tmp]# cat waffle.pfile
WAFFLE.__data_transfer_cache_size=0
WAFFLE.__db_cache_size=507510784
WAFFLE.__java_pool_size=4194304
WAFFLE.__large_pool_size=20971520
WAFFLE.__oracle_base='/orabin'#ORACLE_BASE set from environment
WAFFLE.__pga_aggregate_target=268435456
WAFFLE.__sga_target=805306368
WAFFLE.__shared_io_pool_size=29360128
WAFFLE.__shared_pool_size=234881024
WAFFLE.__streams_pool_size=0
*.audit_file_dest='/orabin/admin/WAFFLE/adump'
*.audit_trail='db'
*.compatible='12.1.0.2.0'
*.control_files='/oradata//WAFFLE/control01.ctl','/oradata//WAFFLE/contr
ol02.ctl'
*.control_files='/oradata/WAFFLE/control01.ctl','/logs/WAFFLE/control02.
ctl'
*.db_block_size=8192
*.db_domain=''
*.db_name='WAFFLE'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=WAFFLEXDB)'
*.log_archive_dest_1='LOCATION=/logs/WAFFLE/arch'
*.log_archive_format='%t_%s_%r.dbf'
*.open_cursors=300
*.pga_aggregate_target=256m
*.processes=300
*.remote_login_passwordfile='EXCLUSIVE'
*.sga_target=768m
*.undo_tablespace='UNDOTBS1'
```

2. Une fois les modifications terminées, créez un fichier spfile basé sur ce fichier pfile.

```
SQL> create spfile from pfile='waffle.tmp.pfile';
File created.
```

Recréer les fichiers de contrôle

Dans une étape précédente, la sortie de `backup controlfile to trace` a été copié sur le nouveau serveur. La partie spécifique de la sortie requise est le `controlfile recreation` commande. Ces informations se trouvent dans le fichier sous la section marquée `Set #1. NORESETLOGS`. Il commence par la ligne `create controlfile reuse database` et doit inclure le mot `noresetlogs`. Il se termine par le caractère point-virgule (;).

1. Dans cet exemple de procédure, le fichier se lit comme suit.

```
CREATE CONTROLFILE REUSE DATABASE "WAFFLE" NORESETLOGS  ARCHIVELOG
    MAXLOGFILES 16
    MAXLOGMEMBERS 3
    MAXDATAFILES 100
    MAXINSTANCES 8
    MAXLOGHISTORY 292
LOGFILE
  GROUP 1 '/logs/WAFFLE/redo/redo01.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 2 '/logs/WAFFLE/redo/redo02.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 3 '/logs/WAFFLE/redo/redo03.log'  SIZE 50M BLOCKSIZE 512
-- STANDBY LOGFILE
DATAFILE
  '/oradata/WAFFLE/system01.dbf',
  '/oradata/WAFFLE/sysaux01.dbf',
  '/oradata/WAFFLE/undotbs01.dbf',
  '/oradata/WAFFLE/users01.dbf'
CHARACTER SET WE8MSWIN1252
;
```

2. Modifiez ce script comme vous le souhaitez pour refléter le nouvel emplacement des différents fichiers. Par exemple, certains fichiers de données connus pour prendre en charge des E/S élevées peuvent être redirigés vers un système de fichiers sur un niveau de stockage hautes performances. Dans d'autres cas, les modifications peuvent être uniquement pour des raisons d'administrateur, telles que l'isolation des fichiers de données d'un PDB donné dans des volumes dédiés.
3. Dans cet exemple, le `DATAFILE` la strophe reste inchangée, mais les journaux de reprise sont déplacés vers un nouvel emplacement dans `/redo` plutôt que de partager de l'espace avec les journaux d'archivage `/logs`.

```
CREATE CONTROLFILE REUSE DATABASE "WAFFLE" NORESETLOGS  ARCHIVELOG
    MAXLOGFILES 16
    MAXLOGMEMBERS 3
    MAXDATAFILES 100
    MAXINSTANCES 8
    MAXLOGHISTORY 292
LOGFILE
  GROUP 1 '/redo/redo01.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 2 '/redo/redo02.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 3 '/redo/redo03.log'  SIZE 50M BLOCKSIZE 512
-- STANDBY LOGFILE
DATAFILE
  '/oradata/WAFFLE/system01.dbf',
  '/oradata/WAFFLE/sysaux01.dbf',
  '/oradata/WAFFLE/undotbs01.dbf',
  '/oradata/WAFFLE/users01.dbf'
CHARACTER SET WE8MSWIN1252
;
```

```
SQL> startup nomount;
ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             331353200 bytes
Database Buffers          465567744 bytes
Redo Buffers                5455872 bytes
SQL> CREATE CONTROLFILE REUSE DATABASE "WAFFLE" NORESETLOGS  ARCHIVELOG
  2      MAXLOGFILES 16
  3      MAXLOGMEMBERS 3
  4      MAXDATAFILES 100
  5      MAXINSTANCES 8
  6      MAXLOGHISTORY 292
  7  LOGFILE
  8    GROUP 1 '/redo/redo01.log'  SIZE 50M BLOCKSIZE 512,
  9    GROUP 2 '/redo/redo02.log'  SIZE 50M BLOCKSIZE 512,
 10    GROUP 3 '/redo/redo03.log'  SIZE 50M BLOCKSIZE 512
 11  -- STANDBY LOGFILE
 12  DATAFILE
 13    '/oradata/WAFFLE/system01.dbf',
 14    '/oradata/WAFFLE/sysaux01.dbf',
 15    '/oradata/WAFFLE/undotbs01.dbf',
 16    '/oradata/WAFFLE/users01.dbf'
 17  CHARACTER SET WE8MSWIN1252
 18  ;
Control file created.
SQL>
```

Si des fichiers sont mal placés ou si des paramètres sont mal configurés, des erreurs sont générées et indiquent ce qui doit être corrigé. La base de données est montée, mais elle n'est pas encore ouverte et ne peut pas être ouverte car les fichiers de données utilisés sont toujours marqués comme étant en mode de sauvegarde à chaud. Les journaux d'archivage doivent d'abord être appliqués pour rendre la base de données cohérente.

Réplication initiale du journal

Au moins une opération de réponse de journal est nécessaire pour rendre les fichiers de données cohérents. De nombreuses options sont disponibles pour relire les journaux. Dans certains cas, l'emplacement du journal d'archivage d'origine sur le serveur d'origine peut être partagé via NFS et la réponse du journal peut être effectuée directement. Dans d'autres cas, les journaux d'archivage doivent être copiés.

Par exemple, un simple `scp` l'opération peut copier tous les journaux en cours du serveur source vers le serveur de migration :

```
[oracle@jfsc3 arch]$ scp jfsc1:/logs/WAFFLE/arch/* ./
oracle@jfsc1's password:
1_22_912662036.dbf                                          100%   47MB
47.0MB/s   00:01
1_23_912662036.dbf                                          100%   40MB
40.4MB/s   00:00
1_24_912662036.dbf                                          100%   45MB
45.4MB/s   00:00
1_25_912662036.dbf                                          100%   41MB
40.9MB/s   00:01
1_26_912662036.dbf                                          100%   39MB
39.4MB/s   00:00
1_27_912662036.dbf                                          100%   39MB
38.7MB/s   00:00
1_28_912662036.dbf                                          100%   40MB
40.1MB/s   00:01
1_29_912662036.dbf                                          100%   17MB
16.9MB/s   00:00
1_30_912662036.dbf                                          100%  636KB
636.0KB/s   00:00
```

Relecture initiale du journal

Une fois les fichiers à l'emplacement du journal d'archivage, ils peuvent être relus en exécutant la commande `recover database until cancel` suivi de la réponse `AUTO` pour relire automatiquement tous les journaux disponibles.

```
SQL> recover database until cancel;
ORA-00279: change 382713 generated at 05/24/2016 09:00:54 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_23_912662036.dbf
ORA-00280: change 382713 for thread 1 is in sequence #23
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
AUTO
ORA-00279: change 405712 generated at 05/24/2016 15:01:05 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_24_912662036.dbf
ORA-00280: change 405712 for thread 1 is in sequence #24
ORA-00278: log file '/logs/WAFFLE/arch/1_23_912662036.dbf' no longer
needed for
this recovery
...
ORA-00279: change 713874 generated at 05/26/2016 04:26:43 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_31_912662036.dbf
ORA-00280: change 713874 for thread 1 is in sequence #31
ORA-00278: log file '/logs/WAFFLE/arch/1_30_912662036.dbf' no longer
needed for
this recovery
ORA-00308: cannot open archived log '/logs/WAFFLE/arch/1_31_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
```

La réponse finale au journal d'archivage signale une erreur, mais c'est normal. Le journal l'indique `sqlplus` a cherché un fichier journal particulier et ne l'a pas trouvé. La raison est, très probablement, que le fichier journal n'existe pas encore.

Si la base de données source peut être arrêtée avant de copier les journaux d'archivage, cette étape ne doit être effectuée qu'une seule fois. Les journaux d'archivage sont copiés et relus. Le processus peut ensuite se poursuivre directement vers le processus de mise en service qui réplique les journaux de reprise critiques.

Réplication et relecture incrémentielles du journal

Dans la plupart des cas, la migration n'est pas effectuée immédiatement. La fin du processus de migration peut prendre plusieurs jours, voire plusieurs semaines, ce qui signifie que les journaux doivent être envoyés en continu à la base de données de réplica et relus. Par conséquent, lors de la mise en service, un nombre minimal de données doit être transféré et relu.

Cela peut être scripté de plusieurs manières, mais l'une des méthodes les plus courantes est l'utilisation de rsync, un utilitaire commun de réplication de fichiers. La façon la plus sûre d'utiliser cet utilitaire est de le configurer en tant que démon. Par exemple, le `rsyncd.conf` le fichier suivant montre comment créer une ressource appelée `waffle.arch` Accessible avec les informations d'identification d'utilisateur Oracle et mappé sur `/logs/WAFFLE/arch`. Plus important encore, la ressource est définie en lecture seule, ce qui permet de lire les données de production sans les modifier.

```
[root@jfsc1 arch]# cat /etc/rsyncd.conf
[waffle.arch]
   uid=oracle
   gid=dba
   path=/logs/WAFFLE/arch
   read only = true
[root@jfsc1 arch]# rsync --daemon
```

La commande suivante synchronise la destination du journal d'archive du nouveau serveur avec la ressource rsync `waffle.arch` sur le serveur d'origine. Le `t` argument dans `rsync - potg` permet de comparer la liste de fichiers en fonction de l'horodatage et de copier uniquement les nouveaux fichiers. Ce processus fournit une mise à jour incrémentielle du nouveau serveur. Cette commande peut également être planifiée en cron pour s'exécuter de façon régulière.

```
[oracle@jfsc3 arch]$ rsync -potg --stats --progress jfsc1::waffle.arch/*
/logs/WAFFLE/arch/
1_31_912662036.dbf
      650240 100%  124.02MB/s    0:00:00 (xfer#1, to-check=8/18)
1_32_912662036.dbf
     4873728 100%  110.67MB/s    0:00:00 (xfer#2, to-check=7/18)
1_33_912662036.dbf
     4088832 100%   50.64MB/s    0:00:00 (xfer#3, to-check=6/18)
1_34_912662036.dbf
     8196096 100%   54.66MB/s    0:00:00 (xfer#4, to-check=5/18)
1_35_912662036.dbf
    19376128 100%   57.75MB/s    0:00:00 (xfer#5, to-check=4/18)
1_36_912662036.dbf
       71680 100%  201.15kB/s    0:00:00 (xfer#6, to-check=3/18)
1_37_912662036.dbf
     1144320 100%    3.06MB/s    0:00:00 (xfer#7, to-check=2/18)
1_38_912662036.dbf
    35757568 100%   63.74MB/s    0:00:00 (xfer#8, to-check=1/18)
1_39_912662036.dbf
      984576 100%    1.63MB/s    0:00:00 (xfer#9, to-check=0/18)
Number of files: 18
Number of files transferred: 9
Total file size: 399653376 bytes
Total transferred file size: 75143168 bytes
Literal data: 75143168 bytes
Matched data: 0 bytes
File list size: 474
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 204
Total bytes received: 75153219
sent 204 bytes  received 75153219 bytes  150306846.00 bytes/sec
total size is 399653376  speedup is 5.32
```

Une fois les journaux reçus, ils doivent être relus. Les exemples précédents montrent l'utilisation de sqlplus pour une exécution manuelle `recover database until cancel`, un processus qui peut être facilement automatisé. L'exemple illustré ici utilise le script décrit dans "Relire les journaux sur la base de données". Les scripts acceptent un argument qui spécifie la base de données nécessitant une opération de relecture. Cela permet d'utiliser le même script dans un effort de migration multibase de données.

```
[oracle@jfsc3 logs]$ ./replay.logs.pl WAFFLE
ORACLE_SID = [WAFFLE] ? The Oracle base remains unchanged with value
/orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu May 26 10:47:16 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> ORA-00279: change 713874 generated at 05/26/2016 04:26:43 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_31_912662036.dbf
ORA-00280: change 713874 for thread 1 is in sequence #31
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 814256 generated at 05/26/2016 04:52:30 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_32_912662036.dbf
ORA-00280: change 814256 for thread 1 is in sequence #32
ORA-00278: log file '/logs/WAFFLE/arch/1_31_912662036.dbf' no longer
needed for
this recovery
ORA-00279: change 814780 generated at 05/26/2016 04:53:04 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_33_912662036.dbf
ORA-00280: change 814780 for thread 1 is in sequence #33
ORA-00278: log file '/logs/WAFFLE/arch/1_32_912662036.dbf' no longer
needed for
this recovery
...
ORA-00279: change 1120099 generated at 05/26/2016 09:59:21 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_40_912662036.dbf
ORA-00280: change 1120099 for thread 1 is in sequence #40
ORA-00278: log file '/logs/WAFFLE/arch/1_39_912662036.dbf' no longer
needed for
this recovery
ORA-00308: cannot open archived log '/logs/WAFFLE/arch/1_40_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
```

Mise en service

Lorsque vous êtes prêt à passer au nouvel environnement, vous devez effectuer une synchronisation finale qui inclut à la fois les journaux d'archivage et les journaux de reprise. Si l'emplacement original du journal de reprise n'est pas déjà connu, il peut être identifié comme suit :

```
SQL> select member from v$logfile;
MEMBER
--------------------------------------------------------------------------------
------
/logs/WAFFLE/redo/redo01.log
/logs/WAFFLE/redo/redo02.log
/logs/WAFFLE/redo/redo03.log
```

1. Arrêtez la base de données source.
2. Effectuez une synchronisation finale des journaux d'archivage sur le nouveau serveur avec la méthode souhaitée.
3. Les fichiers redo log source doivent être copiés sur le nouveau serveur. Dans cet exemple, les journaux de reprise ont été déplacés vers un nouveau répertoire à `/redo`.

   ```
   [oracle@jfsc3 logs]$ scp jfsc1:/logs/WAFFLE/redo/* /redo/
   oracle@jfsc1's password:
   redo01.log
   100%   50MB  50.0MB/s   00:01
   redo02.log
   100%   50MB  50.0MB/s   00:00
   redo03.log
   100%   50MB  50.0MB/s   00:00
   ```

4. À ce stade, le nouvel environnement de base de données contient tous les fichiers nécessaires pour le ramener au même état que la source. Les journaux d'archivage doivent être relus une dernière fois.

```
SQL> recover database until cancel;
ORA-00279: change 1120099 generated at 05/26/2016 09:59:21 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_40_912662036.dbf
ORA-00280: change 1120099 for thread 1 is in sequence #40
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
AUTO
ORA-00308: cannot open archived log
'/logs/WAFFLE/arch/1_40_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
ORA-00308: cannot open archived log
'/logs/WAFFLE/arch/1_40_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
```

5. Une fois l'opération terminée, les journaux de reprise doivent être relus. Si le message s'affiche `Media recovery complete` est renvoyé, le processus a réussi et les bases de données sont synchronisées et peuvent être ouvertes.

```
SQL> recover database;
Media recovery complete.
SQL> alter database open;
Database altered.
```

Envoi de journaux - ASM vers le système de fichiers

Cet exemple illustre l'utilisation d'Oracle RMAN pour migrer une base de données. Il est très similaire à l'exemple précédent de système de fichiers pour l'envoi de journaux de système de fichiers, mais les fichiers sur ASM ne sont pas visibles par l'hôte. Les seules options de migration des données situées sur les périphériques ASM sont soit le déplacement du LUN ASM, soit l'utilisation d'Oracle RMAN pour effectuer les opérations de copie.

Bien que RMAN soit obligatoire pour la copie de fichiers à partir d'Oracle ASM, l'utilisation de RMAN ne se limite pas à ASM. RMAN peut être utilisé pour migrer de tout type de stockage vers tout autre type.

Cet exemple montre le déplacement d'une base de données appelée PANCAKE depuis le stockage ASM vers un système de fichiers standard situé sur un serveur différent au niveau des chemins `/oradata` et `/logs`.

Créer une sauvegarde de base de données

La première étape consiste à créer une sauvegarde de la base de données à migrer vers un autre serveur. Comme la source utilise Oracle ASM, RMAN doit être utilisé. Une simple sauvegarde RMAN peut être effectuée comme suit. Cette méthode crée une sauvegarde balisée qui peut être facilement identifiée par RMAN plus tard dans la procédure.

La première commande définit le type de destination de la sauvegarde et l'emplacement à utiliser. La seconde lance la sauvegarde des fichiers de données uniquement.

```
RMAN> configure channel device type disk format '/rman/pancake/%U';
using target database control file instead of recovery catalog
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT   '/rman/pancake/%U';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT   '/rman/pancake/%U';
new RMAN configuration parameters are successfully stored
RMAN> backup database tag 'ONTAP_MIGRATION';
Starting backup at 24-MAY-16
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=251 device type=DISK
channel ORA_DISK_1: starting full datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
input datafile file number=00001 name=+ASM0/PANCAKE/system01.dbf
input datafile file number=00002 name=+ASM0/PANCAKE/sysaux01.dbf
input datafile file number=00003 name=+ASM0/PANCAKE/undotbs101.dbf
input datafile file number=00004 name=+ASM0/PANCAKE/users01.dbf
channel ORA_DISK_1: starting piece 1 at 24-MAY-16
channel ORA_DISK_1: finished piece 1 at 24-MAY-16
piece handle=/rman/pancake/1gr6c161_1_1 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:03
channel ORA_DISK_1: starting full datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current control file in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 24-MAY-16
channel ORA_DISK_1: finished piece 1 at 24-MAY-16
piece handle=/rman/pancake/1hr6c164_1_1 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 24-MAY-16
```

Fichier de contrôle de sauvegarde

Un fichier de contrôle de sauvegarde est requis plus tard dans la procédure pour `duplicate database` fonctionnement.

```
RMAN> backup current controlfile format '/rman/pancake/ctrl.bkp';
Starting backup at 24-MAY-16
using channel ORA_DISK_1
channel ORA_DISK_1: starting full datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current control file in backup set
channel ORA_DISK_1: starting piece 1 at 24-MAY-16
channel ORA_DISK_1: finished piece 1 at 24-MAY-16
piece handle=/rman/pancake/ctrl.bkp tag=TAG20160524T032651 comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 24-MAY-16
```

Sauvegarde du fichier de paramètres

Un fichier de paramètres est également requis dans le nouvel environnement. La méthode la plus simple consiste à créer un fichier pfile à partir du fichier spfile ou pfile actuel. Dans cet exemple, la base de données source utilise un fichier spfile.

```
RMAN> create pfile='/rman/pancake/pfile' from spfile;
Statement processed
```

Script de renommage de fichier ASM

Plusieurs emplacements de fichiers actuellement définis dans les fichiers de contrôle changent lorsque la base de données est déplacée. Le script suivant crée un script RMAN pour faciliter le processus. Cet exemple illustre une base de données comportant un très petit nombre de fichiers de données, mais en général, les bases de données contiennent des centaines, voire des milliers de fichiers de données.

Ce script est disponible dans "Conversion de noms de système de fichiers ASM en système de fichiers" et il fait deux choses.

Tout d'abord, il crée un paramètre pour redéfinir les emplacements du journal de reprise appelés `log_file_name_convert`. Il s'agit essentiellement d'une liste de champs alternatifs. Le premier champ est l'emplacement d'un journal de reprise en cours et le second est l'emplacement sur le nouveau serveur. Le schéma est alors répété.

La deuxième fonction consiste à fournir un modèle pour renommer le fichier de données. Le script passe en boucle dans les fichiers de données, extrait les informations relatives au nom et au numéro de fichier et les formate en tant que script RMAN. Il fait ensuite la même chose avec les fichiers temporaires. Le résultat est un script rman simple qui peut être modifié comme vous le souhaitez pour vous assurer que les fichiers sont restaurés à l'emplacement souhaité.

```
SQL> @/rman/mk.rename.scripts.sql
Parameters for log file conversion:
*.log_file_name_convert = '+ASM0/PANCAKE/redo01.log',
'/NEW_PATH/redo01.log','+ASM0/PANCAKE/redo02.log',
'/NEW_PATH/redo02.log','+ASM0/PANCAKE/redo03.log', '/NEW_PATH/redo03.log'
rman duplication script:
run
{
set newname for datafile 1 to '+ASM0/PANCAKE/system01.dbf';
set newname for datafile 2 to '+ASM0/PANCAKE/sysaux01.dbf';
set newname for datafile 3 to '+ASM0/PANCAKE/undotbs101.dbf';
set newname for datafile 4 to '+ASM0/PANCAKE/users01.dbf';
set newname for tempfile 1 to '+ASM0/PANCAKE/temp01.dbf';
duplicate target database for standby backup location INSERT_PATH_HERE;
}
PL/SQL procedure successfully completed.
```

Capturer la sortie de cet écran. Le `log_file_name_convert` le paramètre est placé dans le fichier pfile comme décrit ci-dessous. Le script de renommage et de duplication du fichier de données RMAN doit être modifié en conséquence pour placer les fichiers de données aux emplacements souhaités. Dans cet exemple, ils sont tous placés dans `/oradata/pancake`.

```
run
{
set newname for datafile 1 to '/oradata/pancake/pancake.dbf';
set newname for datafile 2 to '/oradata/pancake/sysaux.dbf';
set newname for datafile 3 to '/oradata/pancake/undotbs1.dbf';
set newname for datafile 4 to '/oradata/pancake/users.dbf';
set newname for tempfile 1 to '/oradata/pancake/temp.dbf';
duplicate target database for standby backup location '/rman/pancake';
}
```

Préparer la structure du répertoire

Les scripts sont presque prêts à être exécutés, mais d'abord la structure de répertoire doit être en place. Si les répertoires requis ne sont pas déjà présents, ils doivent être créés ou la procédure de démarrage de la base de données échoue. L'exemple ci-dessous reflète les exigences minimales.

```
[oracle@jfsc2 ~]$ mkdir /oradata/pancake
[oracle@jfsc2 ~]$ mkdir /logs/pancake
[oracle@jfsc2 ~]$ cd /orabin/admin
[oracle@jfsc2 admin]$ mkdir PANCAKE
[oracle@jfsc2 admin]$ cd PANCAKE
[oracle@jfsc2 PANCAKE]$ mkdir adump dpdump pfile scripts xdb_wallet
```

Créer une entrée oratab

La commande suivante est requise pour que des utilitaires tels que oraenv fonctionnent correctement.

```
PANCAKE:/orabin/product/12.1.0/dbhome_1:N
```

Mises à jour des paramètres

Le fichier pfile enregistré doit être mis à jour pour refléter toute modification de chemin sur le nouveau serveur. Les modifications du chemin d'accès au fichier de données sont modifiées par le script de duplication RMAN, et presque toutes les bases de données nécessitent des modifications `control_files` et `log_archive_dest` paramètres. Il peut également y avoir des emplacements de fichiers d'audit qui doivent être modifiés, ainsi que des paramètres tels que `db_create_file_dest` Peut ne pas être pertinent en dehors d'ASM. Un administrateur de base de données expérimenté doit examiner attentivement les modifications proposées avant de poursuivre.

Dans cet exemple, les changements de clé sont les emplacements des fichiers de contrôle, la destination de l'archive de journal et l'ajout du `log_file_name_convert` paramètre.

```
PANCAKE.__data_transfer_cache_size=0
PANCAKE.__db_cache_size=545259520
PANCAKE.__java_pool_size=4194304
PANCAKE.__large_pool_size=25165824
PANCAKE.__oracle_base='/orabin'#ORACLE_BASE set from environment
PANCAKE.__pga_aggregate_target=268435456
PANCAKE.__sga_target=805306368
PANCAKE.__shared_io_pool_size=29360128
PANCAKE.__shared_pool_size=192937984
PANCAKE.__streams_pool_size=0
*.audit_file_dest='/orabin/admin/PANCAKE/adump'
*.audit_trail='db'
*.compatible='12.1.0.2.0'
*.control_files='+ASM0/PANCAKE/control01.ctl','+ASM0/PANCAKE/control02.ctl
'
*.control_files='/oradata/pancake/control01.ctl','/logs/pancake/control02.
ctl'
*.db_block_size=8192
*.db_domain=''
*.db_name='PANCAKE'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=PANCAKEXDB)'
*.log_archive_dest_1='LOCATION=+ASM1'
*.log_archive_dest_1='LOCATION=/logs/pancake'
*.log_archive_format='%t_%s_%r.dbf'
'/logs/path/redo02.log'
*.log_file_name_convert = '+ASM0/PANCAKE/redo01.log',
'/logs/pancake/redo01.log', '+ASM0/PANCAKE/redo02.log',
'/logs/pancake/redo02.log', '+ASM0/PANCAKE/redo03.log',
'/logs/pancake/redo03.log'
*.open_cursors=300
*.pga_aggregate_target=256m
*.processes=300
*.remote_login_passwordfile='EXCLUSIVE'
*.sga_target=768m
*.undo_tablespace='UNDOTBS1'
```

Une fois les nouveaux paramètres confirmés, les paramètres doivent être mis en vigueur. Plusieurs options existent, mais la plupart des clients créent un fichier spfile basé sur le fichier pfile texte.

```
bash-4.1$ sqlplus / as sysdba
SQL*Plus: Release 12.1.0.2.0 Production on Fri Jan 8 11:17:40 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> create spfile from pfile='/rman/pancake/pfile';
File created.
```

Nom de démarrage

La dernière étape avant la réplication de la base de données consiste à afficher les processus de la base de données, mais pas à monter les fichiers. Dans cette étape, des problèmes avec le fichier spfile peuvent devenir évidents. Si le `startup nomount` la commande échoue en raison d'une erreur de paramètre, il est simple de s'arrêter, de corriger le modèle pfile, de le recharger en tant que fichier spfile et de réessayer.

```
SQL> startup nomount;
ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             373296240 bytes
Database Buffers          423624704 bytes
Redo Buffers                5455872 bytes
```

Dupliquez la base de données

La restauration de la sauvegarde RMAN précédente vers le nouvel emplacement prend plus de temps que les autres étapes de ce processus. La base de données doit être dupliquée sans modification de l'ID de base de données (DBID) ou réinitialisation des journaux. Cela empêche l'application des journaux, ce qui est une étape nécessaire pour synchroniser complètement les copies.

Connectez-vous à la base de données avec RMAN en tant qu'aux et exécutez la commande duplicate database en utilisant le script créé lors d'une étape précédente.

```
[oracle@jfsc2 pancake]$ rman auxiliary /
Recovery Manager: Release 12.1.0.2.0 - Production on Tue May 24 03:04:56
2016
Copyright (c) 1982, 2014, Oracle and/or its affiliates.  All rights
reserved.
connected to auxiliary database: PANCAKE (not mounted)
RMAN> run
2> {
3> set newname for datafile 1 to '/oradata/pancake/pancake.dbf';
4> set newname for datafile 2 to '/oradata/pancake/sysaux.dbf';
5> set newname for datafile 3 to '/oradata/pancake/undotbs1.dbf';
6> set newname for datafile 4 to '/oradata/pancake/users.dbf';
7> set newname for tempfile 1 to '/oradata/pancake/temp.dbf';
```

```
8> duplicate target database for standby backup location '/rman/pancake';
9> }
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
Starting Duplicate Db at 24-MAY-16
contents of Memory Script:
{
   restore clone standby controlfile from  '/rman/pancake/ctrl.bkp';
}
executing Memory Script
Starting restore at 24-MAY-16
allocated channel: ORA_AUX_DISK_1
channel ORA_AUX_DISK_1: SID=243 device type=DISK
channel ORA_AUX_DISK_1: restoring control file
channel ORA_AUX_DISK_1: restore complete, elapsed time: 00:00:01
output file name=/oradata/pancake/control01.ctl
output file name=/logs/pancake/control02.ctl
Finished restore at 24-MAY-16
contents of Memory Script:
{
   sql clone 'alter database mount standby database';
}
executing Memory Script
sql statement: alter database mount standby database
released channel: ORA_AUX_DISK_1
allocated channel: ORA_AUX_DISK_1
channel ORA_AUX_DISK_1: SID=243 device type=DISK
contents of Memory Script:
{
   set newname for tempfile  1 to
 "/oradata/pancake/temp.dbf";
   switch clone tempfile all;
   set newname for datafile  1 to
 "/oradata/pancake/pancake.dbf";
   set newname for datafile  2 to
 "/oradata/pancake/sysaux.dbf";
   set newname for datafile  3 to
 "/oradata/pancake/undotbs1.dbf";
   set newname for datafile  4 to
 "/oradata/pancake/users.dbf";
   restore
   clone database
   ;
```

```
}
executing Memory Script
executing command: SET NEWNAME
renamed tempfile 1 to /oradata/pancake/temp.dbf in control file
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
Starting restore at 24-MAY-16
using channel ORA_AUX_DISK_1
channel ORA_AUX_DISK_1: starting datafile backup set restore
channel ORA_AUX_DISK_1: specifying datafile(s) to restore from backup set
channel ORA_AUX_DISK_1: restoring datafile 00001 to
/oradata/pancake/pancake.dbf
channel ORA_AUX_DISK_1: restoring datafile 00002 to
/oradata/pancake/sysaux.dbf
channel ORA_AUX_DISK_1: restoring datafile 00003 to
/oradata/pancake/undotbs1.dbf
channel ORA_AUX_DISK_1: restoring datafile 00004 to
/oradata/pancake/users.dbf
channel ORA_AUX_DISK_1: reading from backup piece
/rman/pancake/1gr6c161_1_1
channel ORA_AUX_DISK_1: piece handle=/rman/pancake/1gr6c161_1_1
tag=ONTAP_MIGRATION
channel ORA_AUX_DISK_1: restored backup piece 1
channel ORA_AUX_DISK_1: restore complete, elapsed time: 00:00:07
Finished restore at 24-MAY-16
contents of Memory Script:
{
   switch clone datafile all;
}
executing Memory Script
datafile 1 switched to datafile copy
input datafile copy RECID=5 STAMP=912655725 file
name=/oradata/pancake/pancake.dbf
datafile 2 switched to datafile copy
input datafile copy RECID=6 STAMP=912655725 file
name=/oradata/pancake/sysaux.dbf
datafile 3 switched to datafile copy
input datafile copy RECID=7 STAMP=912655725 file
name=/oradata/pancake/undotbs1.dbf
datafile 4 switched to datafile copy
input datafile copy RECID=8 STAMP=912655725 file
name=/oradata/pancake/users.dbf
Finished Duplicate Db at 24-MAY-16
```

Réplication initiale du journal

Vous devez maintenant envoyer les modifications de la base de données source vers un nouvel emplacement. Cela peut nécessiter une combinaison d'étapes. La méthode la plus simple serait que RMAN sur la base de données source écrive des journaux d'archive sur une connexion réseau partagée. Si aucun emplacement partagé n'est disponible, une autre méthode consiste à utiliser RMAN pour écrire dans un système de fichiers local, puis à utiliser rcp ou rsync pour copier les fichiers.

Dans cet exemple, le `/rman` Directory est un partage NFS disponible pour la base de données d'origine et migrée.

L'une des questions importantes est la `disk format` clause. Le format de disque de la sauvegarde est `%h_%e_%a.dbf`, Ce qui signifie que vous devez utiliser le format du numéro de thread, du numéro de séquence et de l'ID d'activation de la base de données. Bien que les lettres soient différentes, cela correspond à `log_archive_format='%t_%s_%r.dbf` dans le fichier pfile. Ce paramètre spécifie également les journaux d'archivage au format de numéro de thread, de numéro de séquence et d'ID d'activation. Le résultat final est que les sauvegardes du fichier journal sur la source utilisent une convention de dénomination attendue par la base de données. Cela permet de réaliser des opérations telles que `recover database` beaucoup plus simple parce que sqlplus anticipe correctement les noms des journaux d'archive à lire.

```
RMAN> configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/arch/%h_%e_%a.dbf';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters are successfully stored
released channel: ORA_DISK_1
RMAN> backup as copy archivelog from time 'sysdate-2';
Starting backup at 24-MAY-16
current log archived
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=373 device type=DISK
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=54 RECID=70 STAMP=912658508
output file name=/rman/pancake/logship/1_54_912576125.dbf RECID=123
STAMP=912659482
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=41 RECID=29 STAMP=912654101
output file name=/rman/pancake/logship/1_41_912576125.dbf RECID=124
STAMP=912659483
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
...
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=45 RECID=33 STAMP=912654688
output file name=/rman/pancake/logship/1_45_912576125.dbf RECID=152
STAMP=912659514
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=47 RECID=36 STAMP=912654809
output file name=/rman/pancake/logship/1_47_912576125.dbf RECID=153
STAMP=912659515
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
Finished backup at 24-MAY-16
```

Relecture initiale du journal

Une fois les fichiers à l'emplacement du journal d'archivage, ils peuvent être relus en exécutant la commande `recover database until cancel` suivi de la réponse `AUTO` pour relire automatiquement tous les journaux disponibles. Le fichier de paramètres dirige actuellement les journaux d'archivage vers `/logs/archive`, Mais cela ne correspond pas à l'emplacement où RMAN a été utilisé pour enregistrer les journaux. L'emplacement peut être redirigé temporairement comme suit avant de récupérer la base de données.

```
SQL> alter system set log_archive_dest_1='LOCATION=/rman/pancake/logship'
scope=memory;
System altered.
SQL> recover standby database until cancel;
ORA-00279: change 560224 generated at 05/24/2016 03:25:53 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_49_912576125.dbf
ORA-00280: change 560224 for thread 1 is in sequence #49
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
AUTO
ORA-00279: change 560353 generated at 05/24/2016 03:29:17 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_50_912576125.dbf
ORA-00280: change 560353 for thread 1 is in sequence #50
ORA-00278: log file '/rman/pancake/logship/1_49_912576125.dbf' no longer
needed
for this recovery
...
ORA-00279: change 560591 generated at 05/24/2016 03:33:56 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_54_912576125.dbf
ORA-00280: change 560591 for thread 1 is in sequence #54
ORA-00278: log file '/rman/pancake/logship/1_53_912576125.dbf' no longer
needed
for this recovery
ORA-00308: cannot open archived log
'/rman/pancake/logship/1_54_912576125.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
```

La réponse finale au journal d'archivage signale une erreur, mais c'est normal. L'erreur indique que sqlplus recherchait un fichier journal particulier et qu'il ne l'a pas trouvé. La raison est la plus probable que le fichier journal n'existe pas encore.

Si la base de données source peut être arrêtée avant de copier les journaux d'archivage, cette étape ne doit être effectuée qu'une seule fois. Les journaux d'archivage sont copiés et relus. Le processus peut ensuite se poursuivre directement vers le processus de mise en service qui réplique les journaux de reprise critiques.

Réplication et relecture incrémentielles du journal

Dans la plupart des cas, la migration n'est pas effectuée immédiatement. La fin du processus de migration peut prendre plusieurs jours, voire plusieurs semaines, ce qui signifie que les journaux doivent être envoyés en continu à la base de données de réplica et relus. Ainsi, le transfert et la lecture de données minimales doivent être assurés à l'arrivée de la mise en service.

Ce processus peut facilement être scripté. Par exemple, la commande suivante peut être planifiée sur la base de données d'origine pour s'assurer que l'emplacement utilisé pour l'envoi des journaux est mis à jour en

permanence.

```
[oracle@jfsc1 pancake]$ cat copylogs.rman
configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
backup as copy archivelog from time 'sysdate-2';
```

```
[oracle@jfsc1 pancake]$ rman target / cmdfile=copylogs.rman
Recovery Manager: Release 12.1.0.2.0 - Production on Tue May 24 04:36:19
2016
Copyright (c) 1982, 2014, Oracle and/or its affiliates.  All rights
reserved.
connected to target database: PANCAKE (DBID=3574534589)
RMAN> configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
2> backup as copy archivelog from time 'sysdate-2';
3>
4>
using target database control file instead of recovery catalog
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters are successfully stored
Starting backup at 24-MAY-16
current log archived
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=369 device type=DISK
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=54 RECID=123 STAMP=912659482
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:22
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_54_912576125.dbf
continuing other job steps, job failed will not be re-run
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=41 RECID=124 STAMP=912659483
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:23
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_41_912576125.dbf
continuing other job steps, job failed will not be re-run
...
```

```
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=45 RECID=152 STAMP=912659514
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:55
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_45_912576125.dbf
continuing other job steps, job failed will not be re-run
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=47 RECID=153 STAMP=912659515
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:57
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_47_912576125.dbf
Recovery Manager complete.
```

Une fois les journaux reçus, ils doivent être relus. Des exemples précédents ont montré l'utilisation de sqlplus pour une exécution manuelle `recover database until cancel`, qui peut être facilement automatisé. L'exemple illustré ici utilise le script décrit dans "Relire les journaux sur la base de données de secours". Le script accepte un argument qui spécifie la base de données nécessitant une opération de relecture. Ce processus permet d'utiliser le même script dans un effort de migration multibase de données.

```
[root@jfsc2 pancake]# ./replaylogs.pl PANCAKE
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Tue May 24 04:47:10 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> ORA-00279: change 560591 generated at 05/24/2016 03:33:56 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_54_912576125.dbf
ORA-00280: change 560591 for thread 1 is in sequence #54
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 562219 generated at 05/24/2016 04:15:08 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_55_912576125.dbf
ORA-00280: change 562219 for thread 1 is in sequence #55
ORA-00278: log file '/rman/pancake/logship/1_54_912576125.dbf' no longer
needed for this recovery
ORA-00279: change 562370 generated at 05/24/2016 04:19:18 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_56_912576125.dbf
ORA-00280: change 562370 for thread 1 is in sequence #56
ORA-00278: log file '/rman/pancake/logship/1_55_912576125.dbf' no longer
needed for this recovery
...
ORA-00279: change 563137 generated at 05/24/2016 04:36:20 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_65_912576125.dbf
ORA-00280: change 563137 for thread 1 is in sequence #65
ORA-00278: log file '/rman/pancake/logship/1_64_912576125.dbf' no longer
needed for this recovery
ORA-00308: cannot open archived log
'/rman/pancake/logship/1_65_912576125.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
```

Mise en service

Lorsque vous êtes prêt à passer au nouvel environnement, vous devez effectuer une synchronisation finale. Lorsque vous travaillez avec des systèmes de fichiers réguliers, il est facile de s'assurer que la base de données migrée est synchronisée à 100 % par rapport à l'original car les journaux de reprise d'origine sont copiés et relus. Il n'y a pas de bonne façon de le faire avec ASM. Seuls les journaux d'archivage peuvent être facilement recopiés. Pour s'assurer qu'aucune donnée n'est perdue, l'arrêt final de la base de données d'origine doit être effectué avec précaution.

1. Tout d'abord, la base de données doit être mise en veille, en veillant à ce qu'aucune modification ne soit apportée. Cette mise en veille peut inclure la désactivation des opérations planifiées, l'arrêt des auditeurs et/ou l'arrêt des applications.
2. Une fois cette étape effectuée, la plupart des administrateurs de bases de données créent une table fictive qui sert de marqueur de l'arrêt.
3. Forcer l'archivage des journaux pour s'assurer que la création de la table fictive est enregistrée dans les journaux d'archivage. Pour ce faire, exécutez les commandes suivantes :

```
SQL> create table cutovercheck as select * from dba_users;
Table created.
SQL> alter system archive log current;
System altered.
SQL> shutdown immediate;
Database closed.
Database dismounted.
ORACLE instance shut down.
```

4. Pour copier le dernier des journaux d'archivage, exécutez les commandes suivantes. La base de données doit être disponible mais pas ouverte.

```
SQL> startup mount;
ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             331353200 bytes
Database Buffers          465567744 bytes
Redo Buffers                5455872 bytes
Database mounted.
```

5. Pour copier les journaux d'archivage, exécutez les commandes suivantes :

```
RMAN> configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
2> backup as copy archivelog from time 'sysdate-2';
3>
4>
using target database control file instead of recovery catalog
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters are successfully stored
Starting backup at 24-MAY-16
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=8 device type=DISK
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=54 RECID=123 STAMP=912659482
RMAN-03009: failure of backup command on ORA_DISK_1 channel at
05/24/2016 04:58:24
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_54_912576125.dbf
continuing other job steps, job failed will not be re-run
...
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=45 RECID=152 STAMP=912659514
RMAN-03009: failure of backup command on ORA_DISK_1 channel at
05/24/2016 04:58:58
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_45_912576125.dbf
continuing other job steps, job failed will not be re-run
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=47 RECID=153 STAMP=912659515
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03009: failure of backup command on ORA_DISK_1 channel at
05/24/2016 04:59:00
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_47_912576125.dbf
```

6. Enfin, rejouez les journaux d'archive restants sur le nouveau serveur.

```
[root@jfsc2 pancake]# ./replaylogs.pl PANCAKE
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Tue May 24 05:00:53 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> ORA-00279: change 563137 generated at 05/24/2016 04:36:20 needed
for thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_65_912576125.dbf
ORA-00280: change 563137 for thread 1 is in sequence #65
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 563629 generated at 05/24/2016 04:55:20 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_66_912576125.dbf
ORA-00280: change 563629 for thread 1 is in sequence #66
ORA-00278: log file '/rman/pancake/logship/1_65_912576125.dbf' no longer
needed
for this recovery
ORA-00308: cannot open archived log
'/rman/pancake/logship/1_66_912576125.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
```

7. À ce stade, répliquez toutes les données. La base de données est prête à être convertie à partir d'une base de données de secours vers une base de données opérationnelle active, puis ouverte.

```
SQL> alter database activate standby database;
Database altered.
SQL> alter database open;
Database altered.
```

8. Confirmer la présence de la table factice, puis la déposer.

```
SQL> desc cutovercheck
 Name                                      Null?    Type
 ----------------------------------------- --------
----------------------------
 USERNAME                                  NOT NULL VARCHAR2(128)
 USER_ID                                   NOT NULL NUMBER
 PASSWORD                                           VARCHAR2(4000)
 ACCOUNT_STATUS                            NOT NULL VARCHAR2(32)
 LOCK_DATE                                          DATE
 EXPIRY_DATE                                        DATE
 DEFAULT_TABLESPACE                        NOT NULL VARCHAR2(30)
 TEMPORARY_TABLESPACE                      NOT NULL VARCHAR2(30)
 CREATED                                   NOT NULL DATE
 PROFILE                                   NOT NULL VARCHAR2(128)
 INITIAL_RSRC_CONSUMER_GROUP                        VARCHAR2(128)
 EXTERNAL_NAME                                      VARCHAR2(4000)
 PASSWORD_VERSIONS                                  VARCHAR2(12)
 EDITIONS_ENABLED                                   VARCHAR2(1)
 AUTHENTICATION_TYPE                                VARCHAR2(8)
 PROXY_ONLY_CONNECT                                 VARCHAR2(1)
 COMMON                                             VARCHAR2(3)
 LAST_LOGIN                                         TIMESTAMP(9) WITH
TIME ZONE
 ORACLE_MAINTAINED                                  VARCHAR2(1)
SQL> drop table cutovercheck;
Table dropped.
```

Migration des journaux de reprise sans interruption

Il arrive qu'une base de données soit correctement organisée de manière globale, à l'exception des journaux de reprise. Cela peut se produire pour de nombreuses raisons, dont la plus courante est liée aux snapshots. Des produits tels que SnapManager pour Oracle, SnapCenter et la structure de gestion du stockage NetApp Snap Creator permettent une restauration quasi instantanée d'une base de données, mais uniquement si vous restaurez l'état des volumes de fichiers de données. Si les journaux de reprise partagent l'espace avec les fichiers de données, la restauration ne peut pas être effectuée en toute sécurité, car elle entraînerait la destruction des journaux de reprise, ce qui entraînerait probablement une perte des données. Les journaux de reprise doivent donc être déplacés.

Cette procédure est simple et peut être effectuée sans interruption.

Configuration actuelle du journal de reprise

1. Identifiez le nombre de groupes de fichiers redo log et leurs numéros de groupe respectifs.

```
SQL> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
--------------------------------------------------------------------------------
--------
1 /redo0/NTAP/redo01a.log
1 /redo1/NTAP/redo01b.log
2 /redo0/NTAP/redo02a.log
2 /redo1/NTAP/redo02b.log
3 /redo0/NTAP/redo03a.log
3 /redo1/NTAP/redo03b.log
rows selected.
```

2. Indiquez la taille des journaux de reprise.

```
SQL> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
--------------------------------------------------------------------------------
--------
1 524288000
2 524288000
3 524288000
```

Créer de nouveaux journaux

1. Pour chaque journal de reprise, créez un nouveau groupe avec la taille et le nombre de membres correspondants.

```
SQL> alter database add logfile ('/newredo0/redo01a.log',
'/newredo1/redo01b.log') size 500M;
Database altered.
SQL> alter database add logfile ('/newredo0/redo02a.log',
'/newredo1/redo02b.log') size 500M;
Database altered.
SQL> alter database add logfile ('/newredo0/redo03a.log',
'/newredo1/redo03b.log') size 500M;
Database altered.
SQL>
```

2. Vérifiez la nouvelle configuration.

```
SQL> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
--------------------------------------------------------------------------------
--------
1 /redo0/NTAP/redo01a.log
1 /redo1/NTAP/redo01b.log
2 /redo0/NTAP/redo02a.log
2 /redo1/NTAP/redo02b.log
3 /redo0/NTAP/redo03a.log
3 /redo1/NTAP/redo03b.log
4 /newredo0/redo01a.log
4 /newredo1/redo01b.log
5 /newredo0/redo02a.log
5 /newredo1/redo02b.log
6 /newredo0/redo03a.log
6 /newredo1/redo03b.log
12 rows selected.
```

Supprimez les anciens journaux

1. Supprimez les anciens journaux (groupes 1, 2 et 3).

```
SQL> alter database drop logfile group 1;
Database altered.
SQL> alter database drop logfile group 2;
Database altered.
SQL> alter database drop logfile group 3;
Database altered.
```

2. Si vous rencontrez une erreur qui vous empêche de supprimer un journal actif, forcez un commutateur au journal suivant pour libérer le verrouillage et forcer un point de contrôle global. Reportez-vous à l'exemple suivant de ce processus. La tentative de suppression du groupe de fichiers journaux 2, qui se trouvait sur l'ancien emplacement, a été refusée parce qu'il y avait encore des données actives dans ce fichier journal.

```
SQL> alter database drop logfile group 2;
alter database drop logfile group 2
*
ERROR at line 1:
ORA-01623: log 2 is current log for instance NTAP (thread 1) - cannot
drop
ORA-00312: online log 2 thread 1: '/redo0/NTAP/redo02a.log'
ORA-00312: online log 2 thread 1: '/redo1/NTAP/redo02b.log'
```

3. Un archivage de journaux suivi d'un point de contrôle vous permet de supprimer le fichier journal.

```
SQL> alter system archive log current;
System altered.
SQL> alter system checkpoint;
System altered.
SQL> alter database drop logfile group 2;
Database altered.
```

4. Supprimez ensuite les journaux du système de fichiers. Vous devez effectuer ce processus avec une extrême prudence.

Copie de données hôte

À l'instar de la migration au niveau des bases de données, la migration au niveau de la couche hôte offre une approche indépendante du fournisseur de stockage.

En d'autres termes, parfois "juste copier les fichiers" est la meilleure option.

Bien que cette approche peu technologique puisse sembler trop basique, elle offre des avantages significatifs, car aucun logiciel spécial n'est requis et les données d'origine ne sont pas modifiées en toute sécurité pendant le processus. La principale limitation est le fait qu'une migration de données de copie de fichier est un processus perturbateur, car la base de données doit être arrêtée avant le début de l'opération de copie. Il n'y a pas de bonne façon de synchroniser les modifications dans un fichier, de sorte que les fichiers doivent être complètement suspendus avant le début de la copie.

Si l'arrêt requis par une opération de copie n'est pas souhaitable, la meilleure option basée sur l'hôte suivante consiste à exploiter un gestionnaire de volumes logiques (LVM). De nombreuses options LVM existent, y compris Oracle ASM, toutes avec des capacités similaires, mais avec certaines limitations qui doivent être prises en compte. Dans la plupart des cas, la migration peut s'effectuer sans interruption ni perturbation.

Copie du système de fichiers vers le système de fichiers

L'utilité d'une simple opération de copie ne doit pas être sous-estimée. Cette opération requiert un temps d'indisponibilité lors de la copie, mais le processus est extrêmement fiable et ne requiert aucune expertise particulière en matière de systèmes d'exploitation, de bases de données ou de systèmes de stockage. De plus, elle est très sûre car elle n'affecte pas les données d'origine. Généralement, un administrateur système modifie les systèmes de fichiers source pour qu'ils soient montés en lecture seule, puis redémarre un serveur pour garantir que rien ne risque d'endommager les données actuelles. Le processus de copie peut être scripté pour s'assurer qu'il s'exécute aussi rapidement que possible sans risque d'erreur de l'utilisateur. Comme le type d'E/S est un simple transfert séquentiel de données, il est très peu gourmand en bande passante.

L'exemple suivant illustre une option pour une migration sûre et rapide.

De production

L'environnement à migrer est le suivant :

- Systèmes de fichiers actuels

```
ontap-nfs1:/host1_oradata       52428800  16196928  36231872  31%
/oradata
ontap-nfs1:/host1_logs          49807360    548032  49259328   2% /logs
```

- Nouveaux systèmes de fichiers

```
ontap-nfs1:/host1_logs_new      49807360       128  49807232   1%
/new/logs
ontap-nfs1:/host1_oradata_new   49807360       128  49807232   1%
/new/oradata
```

Présentation

Il suffit à l'administrateur de bases de données de fermer la base de données et de copier les fichiers pour migrer la base de données. Toutefois, ce processus peut être facilement scripté si de nombreuses bases de données doivent être migrées ou si la réduction des temps d'indisponibilité est essentielle. L'utilisation de scripts réduit également les risques d'erreur de l'utilisateur.

Les exemples de scripts présentés automatisent les opérations suivantes :

- Arrêt de la base de données
- Conversion des systèmes de fichiers existants en état de lecture seule
- Copie de toutes les données de la source vers les systèmes de fichiers cibles, ce qui préserve toutes les autorisations de fichier
- Démontage de l'ancien et du nouveau système de fichiers
- Remontage des nouveaux systèmes de fichiers aux mêmes chemins que les systèmes de fichiers précédents

Procédure

1. Arrêtez la base de données.

```
[root@host1 current]# ./dbshut.pl NTAP
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 15:58:48 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> Database closed.
Database dismounted.
ORACLE instance shut down.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP shut down
```

2. Convertissez les systèmes de fichiers en lecture seule. Ceci peut être effectué plus rapidement en utilisant un script, comme indiqué dans la "Convertir le système de fichiers en lecture seule".

```
[root@host1 current]# ./mk.fs.readonly.pl /oradata
/oradata unmounted
/oradata mounted read-only
[root@host1 current]# ./mk.fs.readonly.pl /logs
/logs unmounted
/logs mounted read-only
```

3. Vérifiez que les systèmes de fichiers sont maintenant en lecture seule.

```
ontap-nfs1:/host1_oradata on /oradata type nfs
(ro,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
ontap-nfs1:/host1_logs on /logs type nfs
(ro,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
```

4. Synchroniser le contenu du système de fichiers avec le `rsync` commande.

```
[root@host1 current]# rsync -rlpogt --stats --progress
--exclude=.snapshot /oradata/ /new/oradata/
sending incremental file list
./
NTAP/
NTAP/IOPS.dbf
```

```
 10737426432 100%  153.50MB/s    0:01:06 (xfer#1, to-check=10/13)
NTAP/iops.dbf.zip
    22823573 100%   12.09MB/s    0:00:01 (xfer#2, to-check=9/13)
...
NTAP/undotbs02.dbf
  1073750016 100%  131.60MB/s    0:00:07 (xfer#10, to-check=1/13)
NTAP/users01.dbf
     5251072 100%    3.95MB/s    0:00:01 (xfer#11, to-check=0/13)
Number of files: 13
Number of files transferred: 11
Total file size: 18570092218 bytes
Total transferred file size: 18570092218 bytes
Literal data: 18570092218 bytes
Matched data: 0 bytes
File list size: 277
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 18572359828
Total bytes received: 228
sent 18572359828 bytes  received 228 bytes  162204017.96 bytes/sec
total size is 18570092218  speedup is 1.00
[root@host1 current]# rsync -rlpogt --stats --progress
--exclude=.snapshot /logs/ /new/logs/
sending incremental file list
./
NTAP/
NTAP/1_22_897068759.dbf
    45523968 100%   95.98MB/s    0:00:00 (xfer#1, to-check=15/18)
NTAP/1_23_897068759.dbf
    40601088 100%   49.45MB/s    0:00:00 (xfer#2, to-check=14/18)
...
NTAP/redo/redo02.log
    52429312 100%   44.68MB/s    0:00:01 (xfer#12, to-check=1/18)
NTAP/redo/redo03.log
    52429312 100%   68.03MB/s    0:00:00 (xfer#13, to-check=0/18)
Number of files: 18
Number of files transferred: 13
Total file size: 527032832 bytes
Total transferred file size: 527032832 bytes
Literal data: 527032832 bytes
Matched data: 0 bytes
File list size: 413
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 527098156
Total bytes received: 278
```

```
sent 527098156 bytes  received 278 bytes  95836078.91 bytes/sec
total size is 527032832  speedup is 1.00
```

5. Démontez les anciens systèmes de fichiers et déplacez les données copiées. Ceci peut être effectué plus rapidement en utilisant un script, comme indiqué dans la "Remplacer le système de fichiers".

```
[root@host1 current]# ./swap.fs.pl /logs,/new/logs
/new/logs unmounted
/logs unmounted
Updated /logs mounted
[root@host1 current]# ./swap.fs.pl /oradata,/new/oradata
/new/oradata unmounted
/oradata unmounted
Updated /oradata mounted
```

6. Vérifiez que les nouveaux systèmes de fichiers sont en place.

```
ontap-nfs1:/host1_logs_new on /logs type nfs
(rw,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
ontap-nfs1:/host1_oradata_new on /oradata type nfs
(rw,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
```

7. Démarrez la base de données.

```
[root@host1 current]# ./dbstart.pl NTAP
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 16:10:07 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             390073456 bytes
Database Buffers          406847488 bytes
Redo Buffers                5455872 bytes
Database mounted.
Database opened.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP started
```

Mise en service entièrement automatisée

Cet exemple de script accepte les arguments du SID de la base de données suivis de paires de systèmes de fichiers délimitées par des points communs. Pour l'exemple ci-dessus, la commande est émise comme suit :

```
[root@host1 current]# ./migrate.oracle.fs.pl NTAP /logs,/new/logs
/oradata,/new/oradata
```

Lorsqu'il est exécuté, l'exemple de script tente d'exécuter la séquence suivante. Il se termine s'il rencontre une erreur dans une étape :

1. Arrêtez la base de données.
2. Convertissez les systèmes de fichiers actuels en mode lecture seule.
3. Utilisez chaque paire d'arguments de système de fichiers délimités par des virgules et synchronisez le premier système de fichiers avec le second.
4. Démonter les systèmes de fichiers précédents.
5. Mettez à jour le `/etc/fstab` classer comme suit :
   a. Créez une sauvegarde à `/etc/fstab.bak`.
   b. Commenter les entrées précédentes pour les systèmes de fichiers antérieurs et nouveaux.
   c. Créez une nouvelle entrée pour le nouveau système de fichiers qui utilise l'ancien point de montage.
6. Montez les systèmes de fichiers.
7. Démarrez la base de données.

Le texte suivant fournit un exemple d'exécution pour ce script :

```
[root@host1 current]# ./migrate.oracle.fs.pl NTAP /logs,/new/logs
/oradata,/new/oradata
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 17:05:50 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> Database closed.
Database dismounted.
ORACLE instance shut down.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP shut down
sending incremental file list
```

```
./
NTAP/
NTAP/1_22_897068759.dbf
    45523968 100%  185.40MB/s    0:00:00 (xfer#1, to-check=15/18)
NTAP/1_23_897068759.dbf
    40601088 100%   81.34MB/s    0:00:00 (xfer#2, to-check=14/18)
...
NTAP/redo/redo02.log
    52429312 100%   70.42MB/s    0:00:00 (xfer#12, to-check=1/18)
NTAP/redo/redo03.log
    52429312 100%   47.08MB/s    0:00:01 (xfer#13, to-check=0/18)
Number of files: 18
Number of files transferred: 13
Total file size: 527032832 bytes
Total transferred file size: 527032832 bytes
Literal data: 527032832 bytes
Matched data: 0 bytes
File list size: 413
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 527098156
Total bytes received: 278
sent 527098156 bytes  received 278 bytes  150599552.57 bytes/sec
total size is 527032832  speedup is 1.00
Succesfully replicated filesystem /logs to /new/logs
sending incremental file list
./
NTAP/
NTAP/IOPS.dbf
 10737426432 100%  176.55MB/s    0:00:58 (xfer#1, to-check=10/13)
NTAP/iops.dbf.zip
    22823573 100%    9.48MB/s    0:00:02 (xfer#2, to-check=9/13)
... NTAP/undotbs01.dbf
   309338112 100%   70.76MB/s    0:00:04 (xfer#9, to-check=2/13)
NTAP/undotbs02.dbf
  1073750016 100%  187.65MB/s    0:00:05 (xfer#10, to-check=1/13)
NTAP/users01.dbf
     5251072 100%    5.09MB/s    0:00:00 (xfer#11, to-check=0/13)
Number of files: 13
Number of files transferred: 11
Total file size: 18570092218 bytes
Total transferred file size: 18570092218 bytes
Literal data: 18570092218 bytes
Matched data: 0 bytes
File list size: 277
File list generation time: 0.001 seconds
```

```
File list transfer time: 0.000 seconds
Total bytes sent: 18572359828
Total bytes received: 228
sent 18572359828 bytes  received 228 bytes  177725933.55 bytes/sec
total size is 18570092218  speedup is 1.00
Succesfully replicated filesystem /oradata to /new/oradata
swap 0 /logs /new/logs
/new/logs unmounted
/logs unmounted
Mounted updated /logs
Swapped filesystem /logs for /new/logs
swap 1 /oradata /new/oradata
/new/oradata unmounted
/oradata unmounted
Mounted updated /oradata
Swapped filesystem /oradata for /new/oradata
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 17:08:59 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             390073456 bytes
Database Buffers          406847488 bytes
Redo Buffers                5455872 bytes
Database mounted.
Database opened.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP started
[root@host1 current]#
```

Migration Oracle ASM spfile et passwd

Le fichier spfile spécifique à ASM et le fichier de mots de passe constituent une difficulté pour terminer la migration impliquant ASM. Par défaut, ces fichiers de métadonnées critiques sont créés sur le premier groupe de disques ASM défini. Si un groupe de disques ASM particulier doit être évacué et supprimé, le fichier spfile et le fichier de mot de passe qui régissent cette instance ASM doivent être déplacés.

Un autre cas d'utilisation où il peut être nécessaire de déplacer ces fichiers est le cas lors du déploiement d'un logiciel de gestion de base de données, tel que SnapManager pour Oracle ou le plug-in SnapCenter pour Oracle. L'une des fonctionnalités de ces produits consiste à restaurer rapidement une base de données en rétablissant l'état des LUN ASM qui hébergent les fichiers de données. Pour ce faire, vous devez mettre le groupe de disques ASM hors ligne avant d'effectuer une restauration. Ce n'est pas un problème tant que les fichiers de données d'une base de données donnée sont isolés dans un groupe de disques ASM dédié.

Lorsque ce groupe de disques contient également le fichier ASM spfile/passwd, la seule façon de mettre le groupe de disques hors ligne est d'arrêter l'instance ASM entière. Il s'agit d'un processus perturbateur, ce qui signifie que le fichier spfile/passwd doit être déplacé.

De production

1. SID de base de données = TOAST
2. Fichiers de données actuels sur `+DATA`
3. Fichiers journaux et fichiers de contrôle actuels sur `+LOGS`
4. Nouveaux groupes de disques ASM définis en tant que `+NEWDATA` et `+NEWLOGS`

Emplacements des fichiers spfile/passwd ASM

La migration de ces fichiers peut s'effectuer sans interruption. Cependant, pour des raisons de sécurité, NetApp recommande de fermer l'environnement de base de données afin de vous assurer que les fichiers ont été déplacés et que la configuration est correctement mise à jour. Cette procédure doit être répétée si plusieurs instances ASM sont présentes sur un serveur.

Identifier les instances ASM

Identifier les instances ASM en fonction des données enregistrées dans le `oratab` fichier. Les instances ASM sont signalées par un symbole +.

```
-bash-4.1$ cat /etc/oratab | grep '^+'
+ASM:/orabin/grid:N             # line added by Agent
```

Il existe une instance ASM appelée +ASM sur ce serveur.

Assurez-vous que toutes les bases de données sont arrêtées

Le seul processus smon visible doit être le smon de l'instance ASM utilisée. La présence d'un autre processus smon indique qu'une base de données est toujours en cours d'exécution.

```
-bash-4.1$ ps -ef | grep smon
oracle     857     1  0 18:26 ?        00:00:00 asm_smon_+ASM
```

Le seul processus smon est l'instance ASM elle-même. Cela signifie qu'aucune autre base de données n'est en cours d'exécution et que vous pouvez continuer en toute sécurité sans risque d'interruption des opérations de la base de données.

Localisez les fichiers

Identifiez l'emplacement actuel du fichier spfile et du fichier de mots de passe ASM à l'aide du `spget` et `pwget` commandes.

```
bash-4.1$ asmcmd
ASMCMD> spget
+DATA/spfile.ora
```

```
ASMCMD> pwget --asm
+DATA/orapwasm
```

Les fichiers se trouvent tous deux à la base du `+DATA` groupe de disques.

Copier des fichiers

Copiez les fichiers dans le nouveau groupe de disques ASM avec le `spcopy` et `pwcopy` commandes. Si le nouveau groupe de disques a été créé récemment et est actuellement vide, il peut être nécessaire de le monter en premier.

```
ASMCMD> mount NEWDATA
```

```
ASMCMD> spcopy +DATA/spfile.ora +NEWDATA/spfile.ora
copying +DATA/spfile.ora -> +NEWDATA/spfilea.ora
```

```
ASMCMD> pwcopy +DATA/orapwasm +NEWDATA/orapwasm
copying +DATA/orapwasm -> +NEWDATA/orapwasm
```

Les fichiers ont été copiés depuis `+DATA` à `+NEWDATA`.

Mettre à jour l'instance ASM

L'instance ASM doit maintenant être mise à jour pour refléter le changement d'emplacement. Le `spset` et `pwset` Les commandes mettent à jour les métadonnées ASM requises pour démarrer le groupe de disques ASM.

```
ASMCMD> spset +NEWDATA/spfile.ora
ASMCMD> pwset --asm +NEWDATA/orapwasm
```

Activez ASM à l'aide de fichiers mis à jour

À ce stade, l'instance ASM utilise toujours les emplacements précédents de ces fichiers. L'instance doit être redémarrée pour forcer une relecture des fichiers à partir de leurs nouveaux emplacements et pour libérer les verrous sur les fichiers précédents.

```
-bash-4.1$ sqlplus / as sysasm
SQL> shutdown immediate;
ASM diskgroups volume disabled
ASM diskgroups dismounted
ASM instance shutdown
```

```
SQL> startup
ASM instance started
Total System Global Area 1140850688 bytes
Fixed Size                  2933400 bytes
Variable Size            1112751464 bytes
ASM Cache                  25165824 bytes
ORA-15032: not all alterations performed
ORA-15017: diskgroup "NEWDATA" cannot be mounted
ORA-15013: diskgroup "NEWDATA" is already mounted
```

Supprimez les anciens fichiers spfile et les anciens fichiers de mots de passe

Si la procédure a été effectuée avec succès, les fichiers précédents ne sont plus verrouillés et peuvent maintenant être supprimés.

```
-bash-4.1$ asmcmd
ASMCMD> rm +DATA/spfile.ora
ASMCMD> rm +DATA/orapwasm
```

Copie d'Oracle ASM vers ASM

Oracle ASM est essentiellement un gestionnaire de volumes combiné léger et un système de fichiers. Comme le système de fichiers n'est pas facilement visible, RMAN doit être utilisé pour effectuer des opérations de copie. Même si un processus de migration basé sur la copie est sûr et simple, il provoque certaines perturbations. Les interruptions peuvent être minimisées, mais pas totalement éliminées.

Si vous souhaitez effectuer une migration sans interruption d'une base de données ASM, il est préférable d'exploiter la capacité d'ASM à rééquilibrer les extensions ASM vers de nouveaux LUN lors de la suppression des anciennes LUN. Cette opération est généralement sûre et non disruptive, mais elle n'offre pas de chemin « back-out ». En cas de problèmes fonctionnels ou de performances, la seule option consiste à migrer les données vers la source.

Ce risque peut être évité en copiant la base de données vers le nouvel emplacement plutôt que de déplacer les données, afin que les données d'origine ne soient pas modifiées. La base de données peut être entièrement testée à son nouvel emplacement avant la mise en service, et la base de données d'origine est disponible comme option de retour en arrière si des problèmes sont détectés.

Cette procédure est l'une des nombreuses options impliquant RMAN. Il est conçu pour permettre un processus en deux étapes dans lequel la sauvegarde initiale est créée, puis synchronisée par la suite via la relecture du journal. Ce processus est recommandé pour réduire les temps d'indisponibilité, car il permet à la base de données de rester opérationnelle et d'assurer l'accès aux données pendant la copie de base initiale.

Copier la base de données

Oracle RMAN crée une copie de niveau 0 (complète) de la base de données source actuellement située sur le groupe de disques ASM `+DATA` vers le nouvel emplacement sur `+NEWDATA`.

```
-bash-4.1$ rman target /
Recovery Manager: Release 12.1.0.2.0 - Production on Sun Dec 6 17:40:03
2015
Copyright (c) 1982, 2014, Oracle and/or its affiliates.  All rights
reserved.
connected to target database: TOAST (DBID=2084313411)
RMAN> backup as copy incremental level 0 database format '+NEWDATA' tag
'ONTAP_MIGRATION';
Starting backup at 06-DEC-15
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=302 device type=DISK
channel ORA_DISK_1: starting datafile copy
input datafile file number=00001
name=+DATA/TOAST/DATAFILE/system.262.897683141
...
input datafile file number=00004
name=+DATA/TOAST/DATAFILE/users.264.897683151
output file name=+NEWDATA/TOAST/DATAFILE/users.258.897759623
tag=ONTAP_MIGRATION RECID=5 STAMP=897759622
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting incremental level 0 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 06-DEC-15
channel ORA_DISK_1: finished piece 1 at 06-DEC-15
piece
handle=+NEWDATA/TOAST/BACKUPSET/2015_12_06/nnsnn0_ontap_migration_0.262.89
7759623 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 06-DEC-15
```

Forcer le changement de journal d'archivage

Vous devez forcer un commutateur de journal d'archivage pour vous assurer que les journaux d'archivage contiennent toutes les données nécessaires pour que la copie soit totalement cohérente. Sans cette commande, les données clés peuvent toujours être présentes dans les journaux de reprise.

```
RMAN> sql 'alter system archive log current';
sql statement: alter system archive log current
```

Arrêtez la base de données source

L'interruption commence à cette étape car la base de données est arrêtée et placée en mode lecture seule à accès limité. Pour arrêter la base de données source, exécutez les commandes suivantes :

```
RMAN> shutdown immediate;
using target database control file instead of recovery catalog
database closed
database dismounted
Oracle instance shut down
RMAN> startup mount;
connected to target database (not started)
Oracle instance started
database mounted
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                390073456 bytes
Database Buffers             406847488 bytes
Redo Buffers                   5455872 bytes
```

Sauvegarde Controlfile

Vous devez sauvegarder le fichier de contrôle si vous devez abandonner la migration et revenir à l'emplacement de stockage d'origine. Une copie du fichier de contrôle de sauvegarde n'est pas nécessaire à 100 %, mais elle facilite le processus de réinitialisation des emplacements des fichiers de base de données vers leur emplacement d'origine.

```
RMAN> backup as copy current controlfile format '/tmp/TOAST.ctrl';
Starting backup at 06-DEC-15
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=358 device type=DISK
channel ORA_DISK_1: starting datafile copy
copying current control file
output file name=/tmp/TOAST.ctrl tag=TAG20151206T174753 RECID=6
STAMP=897760073
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
Finished backup at 06-DEC-15
```

Mises à jour des paramètres

Le fichier spfile actuel contient des références aux fichiers de contrôle sur leurs emplacements actuels dans l'ancien groupe de disques ASM. Il doit être édité, ce qui est facile à faire en éditant une version intermédiaire de pfile.

```
RMAN> create pfile='/tmp/pfile' from spfile;
Statement processed
```

Mettre à jour le fichier pfile

Mettez à jour tous les paramètres faisant référence aux anciens groupes de disques ASM pour refléter les nouveaux noms de groupes de disques ASM. Enregistrez ensuite le fichier pfile mis à jour. Assurez-vous que le `db_create` des paramètres sont présents.

Dans l'exemple ci-dessous, les références à `+DATA` ils ont été remplacés par `+NEWDATA` sont surlignés en jaune. Deux paramètres clés sont le `db_create` paramètres qui créent de nouveaux fichiers à l'emplacement correct.

```
*.compatible='12.1.0.2.0'
*.control_files='+NEWLOGS/TOAST/CONTROLFILE/current.258.897683139'
*.db_block_size=8192
*. db_create_file_dest='+NEWDATA'
*. db_create_online_log_dest_1='+NEWLOGS'
*.db_domain=''
*.db_name='TOAST'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=TOASTXDB)'
*.log_archive_dest_1='LOCATION=+NEWLOGS'
*.log_archive_format='%t_%s_%r.dbf'
```

Mettre à jour le fichier init.ora

La plupart des bases de données ASM utilisent un `init.ora` fichier situé dans le `$ORACLE_HOME/dbs` Répertoire, qui est un point vers le fichier spfile sur le groupe de disques ASM. Ce fichier doit être redirigé vers un emplacement du nouveau groupe de disques ASM.

```
-bash-4.1$ cd $ORACLE_HOME/dbs
-bash-4.1$ cat initTOAST.ora
SPFILE='+DATA/TOAST/spfileTOAST.ora'
```

Modifiez ce fichier comme suit :

```
SPFILE=+NEWLOGS/TOAST/spfileTOAST.ora
```

Récréation du fichier de paramètres

Le fichier spfile est maintenant prêt à être rempli par les données du fichier pfile modifié.

```
RMAN> create spfile from pfile='/tmp/pfile';
Statement processed
```

Démarrez la base de données pour commencer à utiliser le nouveau fichier spfile

Démarrez la base de données pour vous assurer qu'elle utilise maintenant le fichier spfile nouvellement créé et que toute autre modification des paramètres système est correctement enregistrée.

```
RMAN> startup nomount;
connected to target database (not started)
Oracle instance started
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                373296240 bytes
Database Buffers             423624704 bytes
Redo Buffers                   5455872 bytes
```

Restaurer le fichier de contrôle

Le fichier de contrôle de sauvegarde créé par RMAN peut également être restauré directement par RMAN à l'emplacement spécifié dans le nouveau fichier spfile.

```
RMAN> restore controlfile from
'+DATA/TOAST/CONTROLFILE/current.258.897683139';
Starting restore at 06-DEC-15
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=417 device type=DISK
channel ORA_DISK_1: copied control file copy
output file name=+NEWLOGS/TOAST/CONTROLFILE/current.273.897761061
Finished restore at 06-DEC-15
```

Montez la base de données et vérifiez l'utilisation du nouveau fichier de contrôle.

```
RMAN> alter database mount;
using target database control file instead of recovery catalog
Statement processed
```

```
SQL> show parameter control_files;
NAME                                 TYPE        VALUE
------------------------------------ -----------
------------------------------
control_files                        string
+NEWLOGS/TOAST/CONTROLFILE/cur
                                                 rent.273.897761061
```

Relecture du journal

La base de données utilise actuellement les fichiers de données dans l'ancien emplacement. Avant de pouvoir utiliser la copie, elles doivent être synchronisées. Le temps s'est écoulé pendant le processus de copie initial et les modifications ont été enregistrées principalement dans les journaux d'archivage. Ces modifications sont répliquées comme suit :

1. Effectuez une sauvegarde incrémentielle RMAN contenant les journaux d'archivage.

```
RMAN> backup incremental level 1 format '+NEWLOGS' for recover of copy
with tag 'ONTAP_MIGRATION' database;
Starting backup at 06-DEC-15
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=62 device type=DISK
channel ORA_DISK_1: starting incremental level 1 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
input datafile file number=00001
name=+DATA/TOAST/DATAFILE/system.262.897683141
input datafile file number=00002
name=+DATA/TOAST/DATAFILE/sysaux.260.897683143
input datafile file number=00003
name=+DATA/TOAST/DATAFILE/undotbs1.257.897683145
input datafile file number=00004
name=+DATA/TOAST/DATAFILE/users.264.897683151
channel ORA_DISK_1: starting piece 1 at 06-DEC-15
channel ORA_DISK_1: finished piece 1 at 06-DEC-15
piece
handle=+NEWLOGS/TOAST/BACKUPSET/2015_12_06/nnndn1_ontap_migration_0.268.
897762693 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting incremental level 1 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current control file in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 06-DEC-15
channel ORA_DISK_1: finished piece 1 at 06-DEC-15
piece
handle=+NEWLOGS/TOAST/BACKUPSET/2015_12_06/ncsnn1_ontap_migration_0.267.
897762697 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 06-DEC-15
```

2. Relire le journal.

```
RMAN> recover copy of database with tag 'ONTAP_MIGRATION';
Starting recover at 06-DEC-15
using channel ORA_DISK_1
channel ORA_DISK_1: starting incremental datafile backup set restore
channel ORA_DISK_1: specifying datafile copies to recover
recovering datafile copy file number=00001
name=+NEWDATA/TOAST/DATAFILE/system.259.897759609
recovering datafile copy file number=00002
name=+NEWDATA/TOAST/DATAFILE/sysaux.263.897759615
recovering datafile copy file number=00003
name=+NEWDATA/TOAST/DATAFILE/undotbs1.264.897759619
recovering datafile copy file number=00004
name=+NEWDATA/TOAST/DATAFILE/users.258.897759623
channel ORA_DISK_1: reading from backup piece
+NEWLOGS/TOAST/BACKUPSET/2015_12_06/nnndn1_ontap_migration_0.268.8977626
93
channel ORA_DISK_1: piece
handle=+NEWLOGS/TOAST/BACKUPSET/2015_12_06/nnndn1_ontap_migration_0.268.
897762693 tag=ONTAP_MIGRATION
channel ORA_DISK_1: restored backup piece 1
channel ORA_DISK_1: restore complete, elapsed time: 00:00:01
Finished recover at 06-DEC-15
```

Activation

Le fichier de contrôle restauré fait toujours référence aux fichiers de données à l'emplacement d'origine et contient également les informations de chemin des fichiers de données copiés.

1. Pour modifier les fichiers de données actifs, exécutez `switch database to copy` commande.

```
RMAN> switch database to copy;
datafile 1 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/system.259.897759609"
datafile 2 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/sysaux.263.897759615"
datafile 3 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/undotbs1.264.897759619"
datafile 4 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/users.258.897759623"
```

   Les fichiers de données actifs sont désormais les fichiers de données copiés, mais des modifications peuvent encore être contenues dans les journaux de reprise finaux.

2. Pour relire tous les journaux restants, exécutez le `recover database` commande. Si le message s'affiche `media recovery complete` apparaît, le processus a réussi.

```
RMAN> recover database;
Starting recover at 06-DEC-15
using channel ORA_DISK_1
starting media recovery
media recovery complete, elapsed time: 00:00:01
Finished recover at 06-DEC-15
```

Ce processus a uniquement modifié l'emplacement des fichiers de données normaux. Les fichiers de données temporaires doivent être renommés, mais ils n'ont pas besoin d'être copiés car ils sont temporaires uniquement. La base de données est actuellement inactive, il n'y a donc pas de données actives dans les fichiers de données temporaires.

3. Pour déplacer les fichiers de données temporaires, identifiez d'abord leur emplacement.

```
RMAN> select file#||' '||name from v$tempfile;
FILE#||''||NAME
------------------------------------------------------------------------
--------
1 +DATA/TOAST/TEMPFILE/temp.263.897683145
```

4. Déplacez les fichiers de données temporaires à l'aide d'une commande RMAN qui définit le nouveau nom de chaque fichier de données. Avec Oracle Managed Files (OMF), le nom complet n'est pas nécessaire ; le groupe de disques ASM est suffisant. Lorsque la base de données est ouverte, OMF est lié à l'emplacement approprié sur le groupe de disques ASM. Pour déplacer des fichiers, exécutez les commandes suivantes :

```
run {
set newname for tempfile 1 to '+NEWDATA';
switch tempfile all;
}
```

```
RMAN> run {
2> set newname for tempfile 1 to '+NEWDATA';
3> switch tempfile all;
4> }
executing command: SET NEWNAME
renamed tempfile 1 to +NEWDATA in control file
```

Migration du journal de reprise

Le processus de migration est presque terminé, mais les journaux de reprise se trouvent toujours sur le groupe de disques ASM d'origine. Les journaux de reprise ne peuvent pas être transférés directement. Un nouvel ensemble de journaux de reprise est créé et ajouté à la configuration, suivi d'un DROP des anciens journaux.

1. Identifiez le nombre de groupes de fichiers redo log et leurs numéros de groupe respectifs.

```
RMAN> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
------------------------------------------------------------------------
--------
1 +DATA/TOAST/ONLINELOG/group_1.261.897683139
2 +DATA/TOAST/ONLINELOG/group_2.259.897683139
3 +DATA/TOAST/ONLINELOG/group_3.256.897683139
```

2. Indiquez la taille des journaux de reprise.

```
RMAN> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
------------------------------------------------------------------------
--------
1 52428800
2 52428800
3 52428800
```

3. Pour chaque journal de reprise, créez un groupe avec une configuration correspondante. Si vous n'utilisez pas OMF, vous devez spécifier le chemin complet. C'est également un exemple qui utilise le `db_create_online_log` paramètres. Comme indiqué précédemment, ce paramètre a été défini sur +NEWLOGS. Cette configuration vous permet d'utiliser les commandes suivantes pour créer de nouveaux journaux en ligne sans avoir à spécifier un emplacement de fichier ou même un groupe de disques ASM spécifique.

```
RMAN> alter database add logfile size 52428800;
Statement processed
RMAN> alter database add logfile size 52428800;
Statement processed
RMAN> alter database add logfile size 52428800;
Statement processed
```

4. Ouvrez la base de données.

```
SQL> alter database open;
Database altered.
```

5. Supprimez les anciens journaux.

```
RMAN> alter database drop logfile group 1;
Statement processed
```

6. Si vous rencontrez une erreur qui vous empêche de supprimer un journal actif, forcez un commutateur au journal suivant pour libérer le verrouillage et forcer un point de contrôle global. Un exemple est illustré ci-dessous. La tentative de suppression du groupe de fichiers journaux 3, qui se trouvait sur l'ancien emplacement, a été refusée parce qu'il y avait encore des données actives dans ce fichier journal. Un archivage de journaux après un point de contrôle vous permet de supprimer le fichier journal.

```
RMAN> alter database drop logfile group 3;
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03002: failure of sql statement command at 12/08/2015 20:23:51
ORA-01623: log 3 is current log for instance TOAST (thread 4) - cannot
drop
ORA-00312: online log 3 thread 1:
'+LOGS/TOAST/ONLINELOG/group_3.259.897563549'
RMAN> alter system switch logfile;
Statement processed
RMAN> alter system checkpoint;
Statement processed
RMAN> alter database drop logfile group 3;
Statement processed
```

7. Vérifiez l'environnement pour vous assurer que tous les paramètres basés sur l'emplacement sont mis à jour.

```
SQL> select name from v$datafile;
SQL> select member from v$logfile;
SQL> select name from v$tempfile;
SQL> show parameter spfile;
SQL> select name, value from v$parameter where value is not null;
```

8. Le script suivant explique comment simplifier ce processus :

```
[root@host1 current]# ./checkdbdata.pl TOAST
TOAST datafiles:
+NEWDATA/TOAST/DATAFILE/system.259.897759609
+NEWDATA/TOAST/DATAFILE/sysaux.263.897759615
+NEWDATA/TOAST/DATAFILE/undotbs1.264.897759619
+NEWDATA/TOAST/DATAFILE/users.258.897759623
TOAST redo logs:
+NEWLOGS/TOAST/ONLINELOG/group_4.266.897763123
+NEWLOGS/TOAST/ONLINELOG/group_5.265.897763125
+NEWLOGS/TOAST/ONLINELOG/group_6.264.897763125
TOAST temp datafiles:
+NEWDATA/TOAST/TEMPFILE/temp.260.897763165
TOAST spfile
spfile                                string
+NEWDATA/spfiletoast.ora
TOAST key parameters
control_files +NEWLOGS/TOAST/CONTROLFILE/current.273.897761061
log_archive_dest_1 LOCATION=+NEWLOGS
db_create_file_dest +NEWDATA
db_create_online_log_dest_1 +NEWLOGS
```

9. Si les groupes de disques ASM ont été complètement évacués, ils peuvent maintenant être démontés avec `asmcmd`. Cependant, dans de nombreux cas, les fichiers appartenant à d'autres bases de données ou au fichier ASM spfile/passwd peuvent toujours être présents.

```
-bash-4.1$ . oraenv
ORACLE_SID = [TOAST] ? +ASM
The Oracle base remains unchanged with value /orabin
-bash-4.1$ asmcmd
ASMCMD> umount DATA
ASMCMD>
```

Copie d'Oracle ASM vers le système de fichiers

La procédure de copie d'Oracle ASM vers un système de fichiers est très similaire à la procédure de copie d'ASM vers ASM, avec des avantages et des restrictions similaires. La différence principale est la syntaxe des différentes commandes et paramètres de configuration lors de l'utilisation d'un système de fichiers visible par opposition à un groupe de disques ASM.

Copier la base de données

Oracle RMAN permet de créer une copie de niveau 0 (complète) de la base de données source actuellement située sur le groupe de disques ASM `+DATA` vers le nouvel emplacement sur `/oradata`.

```
RMAN> backup as copy incremental level 0 database format
'/oradata/TOAST/%U' tag 'ONTAP_MIGRATION';
Starting backup at 13-MAY-16
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=377 device type=DISK
channel ORA_DISK_1: starting datafile copy
input datafile file number=00001 name=+ASM0/TOAST/system01.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSTEM_FNO-
1_01r5fhjg tag=ONTAP_MIGRATION RECID=1 STAMP=911722099
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:07
channel ORA_DISK_1: starting datafile copy
input datafile file number=00002 name=+ASM0/TOAST/sysaux01.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSAUX_FNO-
2_02r5fhjo tag=ONTAP_MIGRATION RECID=2 STAMP=911722106
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:07
channel ORA_DISK_1: starting datafile copy
input datafile file number=00003 name=+ASM0/TOAST/undotbs101.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-UNDOTBS1_FNO-
3_03r5fhjt tag=ONTAP_MIGRATION RECID=3 STAMP=911722113
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:07
channel ORA_DISK_1: starting datafile copy
copying current control file
output file name=/oradata/TOAST/cf_D-TOAST_id-2098173325_04r5fhk5
tag=ONTAP_MIGRATION RECID=4 STAMP=911722118
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting datafile copy
input datafile file number=00004 name=+ASM0/TOAST/users01.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-USERS_FNO-
4_05r5fhk6 tag=ONTAP_MIGRATION RECID=5 STAMP=911722118
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting incremental level 0 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 13-MAY-16
channel ORA_DISK_1: finished piece 1 at 13-MAY-16
piece handle=/oradata/TOAST/06r5fhk7_1_1 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 13-MAY-16
```

Forcer le changement de journal d’archivage

Forcer le commutateur de journal d’archivage est nécessaire pour s’assurer que les journaux d’archivage contiennent toutes les données requises pour rendre la copie entièrement cohérente. Sans cette commande, les données clés peuvent toujours être présentes dans les journaux de reprise. Pour forcer un commutateur de journal d’archivage, exécutez la commande suivante :

```
RMAN> sql 'alter system archive log current';
sql statement: alter system archive log current
```

Arrêtez la base de données source

L'interruption commence à cette étape car la base de données est arrêtée et placée en mode lecture seule à accès limité. Pour arrêter la base de données source, exécutez les commandes suivantes :

```
RMAN> shutdown immediate;
using target database control file instead of recovery catalog
database closed
database dismounted
Oracle instance shut down
RMAN> startup mount;
connected to target database (not started)
Oracle instance started
database mounted
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                331353200 bytes
Database Buffers             465567744 bytes
Redo Buffers                   5455872 bytes
```

Sauvegarde Controlfile

Sauvegarder les fichiers de contrôle si vous devez abandonner la migration et revenir à l'emplacement de stockage d'origine. Une copie du fichier de contrôle de sauvegarde n'est pas nécessaire à 100 %, mais elle facilite le processus de réinitialisation des emplacements des fichiers de base de données vers leur emplacement d'origine.

```
RMAN> backup as copy current controlfile format '/tmp/TOAST.ctrl';
Starting backup at 08-DEC-15
using channel ORA_DISK_1
channel ORA_DISK_1: starting datafile copy
copying current control file
output file name=/tmp/TOAST.ctrl tag=TAG20151208T194540 RECID=30
STAMP=897939940
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
Finished backup at 08-DEC-15
```

Mises à jour des paramètres

```
RMAN> create pfile='/tmp/pfile' from spfile;
Statement processed
```

Mettre à jour le fichier pfile

Tous les paramètres faisant référence aux anciens groupes de disques ASM doivent être mis à jour et, dans certains cas, supprimés lorsqu'ils ne sont plus pertinents. Mettez-les à jour pour refléter les nouveaux chemins du système de fichiers et enregistrez le fichier pfile mis à jour. Assurez-vous que le chemin cible complet est répertorié. Pour mettre à jour ces paramètres, exécutez les commandes suivantes :

```
*.audit_file_dest='/orabin/admin/TOAST/adump'
*.audit_trail='db'
*.compatible='12.1.0.2.0'
*.control_files='/logs/TOAST/arch/control01.ctl','/logs/TOAST/redo/control
02.ctl'
*.db_block_size=8192
*.db_domain=''
*.db_name='TOAST'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=TOASTXDB)'
*.log_archive_dest_1='LOCATION=/logs/TOAST/arch'
*.log_archive_format='%t_%s_%r.dbf'
*.open_cursors=300
*.pga_aggregate_target=256m
*.processes=300
*.remote_login_passwordfile='EXCLUSIVE'
*.sga_target=768m
*.undo_tablespace='UNDOTBS1'
```

Désactivez le fichier init.ora d'origine

Ce fichier se trouve dans le `$ORACLE_HOME/dbs` Et se trouve généralement dans un fichier pfile qui sert de pointeur vers le fichier spfile sur le groupe de disques ASM. Pour vous assurer que le fichier spfile d'origine n'est plus utilisé, renommez-le. Ne le supprimez pas, cependant, car ce fichier est nécessaire si la migration doit être abandonnée.

```
[oracle@jfsc1 ~]$ cd $ORACLE_HOME/dbs
[oracle@jfsc1 dbs]$ cat initTOAST.ora
SPFILE='+ASM0/TOAST/spfileTOAST.ora'
[oracle@jfsc1 dbs]$ mv initTOAST.ora initTOAST.ora.prev
[oracle@jfsc1 dbs]$
```

Récréation du fichier de paramètres

Il s'agit de la dernière étape de la relocalisation de fichier spfile. Le fichier spfile d'origine n'est plus utilisé et la base de données est actuellement démarrée (mais pas montée) à l'aide du fichier intermédiaire. Le contenu de ce fichier peut être écrit dans le nouvel emplacement spfile comme suit :

```
RMAN> create spfile from pfile='/tmp/pfile';
Statement processed
```

Démarrez la base de données pour commencer à utiliser le nouveau fichier spfile

Vous devez démarrer la base de données pour libérer les verrous sur le fichier intermédiaire et démarrer la base de données en utilisant uniquement le nouveau fichier spfile. Le démarrage de la base de données prouve également que le nouvel emplacement spfile est correct et que ses données sont valides.

```
RMAN> shutdown immediate;
Oracle instance shut down
RMAN> startup nomount;
connected to target database (not started)
Oracle instance started
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                331353200 bytes
Database Buffers             465567744 bytes
Redo Buffers                   5455872 bytes
```

Restaurer le fichier de contrôle

Un fichier de contrôle de sauvegarde a été créé au niveau du chemin `/tmp/TOAST.ctrl` plus tôt dans la procédure. Le nouveau fichier spfile définit les emplacements des fichiers de contrôle comme `/logfs/TOAST/ctrl/ctrlfile1.ctrl` et `/logfs/TOAST/redo/ctrlfile2.ctrl`. Cependant, ces fichiers n'existent pas encore.

1. Cette commande restaure les données du fichier de contrôle dans les chemins définis dans le fichier spfile.

   ```
   RMAN> restore controlfile from '/tmp/TOAST.ctrl';
   Starting restore at 13-MAY-16
   using channel ORA_DISK_1
   channel ORA_DISK_1: copied control file copy
   output file name=/logs/TOAST/arch/control01.ctl
   output file name=/logs/TOAST/redo/control02.ctl
   Finished restore at 13-MAY-16
   ```

2. Exécutez la commande mount pour que les fichiers de contrôle soient correctement découverts et contiennent des données valides.

```
RMAN> alter database mount;
Statement processed
released channel: ORA_DISK_1
```

Pour valider le `control_files` paramètre, exécutez la commande suivante :

```
SQL> show parameter control_files;
NAME                                 TYPE        VALUE
------------------------------------ -----------
------------------------------
control_files                        string
/logs/TOAST/arch/control01.ctl
                                                 ,
/logs/TOAST/redo/control02.c
                                                 tl
```

Relecture du journal

La base de données utilise actuellement les fichiers de données dans l'ancien emplacement. Avant de pouvoir utiliser la copie, les fichiers de données doivent être synchronisés. Le temps s'est écoulé pendant le processus de copie initial et les modifications ont été enregistrées principalement dans les journaux d'archivage. Ces modifications sont répliquées dans les deux étapes suivantes.

1. Effectuez une sauvegarde incrémentielle RMAN contenant les journaux d'archivage.

```
RMAN>  backup incremental level 1 format '/logs/TOAST/arch/%U' for
recover of copy with tag 'ONTAP_MIGRATION' database;
Starting backup at 13-MAY-16
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=124 device type=DISK
channel ORA_DISK_1: starting incremental level 1 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
input datafile file number=00001 name=+ASM0/TOAST/system01.dbf
input datafile file number=00002 name=+ASM0/TOAST/sysaux01.dbf
input datafile file number=00003 name=+ASM0/TOAST/undotbs101.dbf
input datafile file number=00004 name=+ASM0/TOAST/users01.dbf
channel ORA_DISK_1: starting piece 1 at 13-MAY-16
channel ORA_DISK_1: finished piece 1 at 13-MAY-16
piece handle=/logs/TOAST/arch/09r5fj8i_1_1 tag=ONTAP_MIGRATION
comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 13-MAY-16
RMAN-06497: WARNING: control file is not current, control file
AUTOBACKUP skipped
```

2. Relire les journaux.

```
RMAN> recover copy of database with tag 'ONTAP_MIGRATION';
Starting recover at 13-MAY-16
using channel ORA_DISK_1
channel ORA_DISK_1: starting incremental datafile backup set restore
channel ORA_DISK_1: specifying datafile copies to recover
recovering datafile copy file number=00001 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-SYSTEM_FNO-1_01r5fhjg
recovering datafile copy file number=00002 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-SYSAUX_FNO-2_02r5fhjo
recovering datafile copy file number=00003 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-UNDOTBS1_FNO-3_03r5fhjt
recovering datafile copy file number=00004 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-USERS_FNO-4_05r5fhk6
channel ORA_DISK_1: reading from backup piece
/logs/TOAST/arch/09r5fj8i_1_1
channel ORA_DISK_1: piece handle=/logs/TOAST/arch/09r5fj8i_1_1
tag=ONTAP_MIGRATION
channel ORA_DISK_1: restored backup piece 1
channel ORA_DISK_1: restore complete, elapsed time: 00:00:01
Finished recover at 13-MAY-16
RMAN-06497: WARNING: control file is not current, control file
AUTOBACKUP skipped
```

Activation

Le fichier de contrôle restauré fait toujours référence aux fichiers de données à l'emplacement d'origine et contient également les informations de chemin des fichiers de données copiés.

1. Pour modifier les fichiers de données actifs, exécutez `switch database to copy` commande :

```
RMAN> switch database to copy;
datafile 1 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-SYSTEM_FNO-1_01r5fhjg"
datafile 2 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-SYSAUX_FNO-2_02r5fhjo"
datafile 3 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-UNDOTBS1_FNO-3_03r5fhjt"
datafile 4 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-USERS_FNO-4_05r5fhk6"
```

2. Bien que les fichiers de données soient parfaitement cohérents, une dernière étape est nécessaire pour relire les modifications restantes enregistrées dans les journaux de reprise en ligne. Utilisez le `recover database` pour relire ces modifications et rendre la copie identique à 100 % à l'original. Toutefois, la copie n'est pas encore ouverte.

```
RMAN> recover database;
Starting recover at 13-MAY-16
using channel ORA_DISK_1
starting media recovery
archived log for thread 1 with sequence 28 is already on disk as file
+ASM0/TOAST/redo01.log
archived log file name=+ASM0/TOAST/redo01.log thread=1 sequence=28
media recovery complete, elapsed time: 00:00:00
Finished recover at 13-MAY-16
```

Déplacer les fichiers de données temporaires

1. Identifiez l'emplacement des fichiers de données temporaires toujours en cours d'utilisation sur le groupe de disques d'origine.

```
RMAN> select file#||' '||name from v$tempfile;
FILE#||''||NAME
-------------------------------------------------------------------------
--------
1 +ASM0/TOAST/temp01.dbf
```

2. Pour déplacer les fichiers de données, exécutez les commandes suivantes. S'il existe de nombreux fichiers tempfiles, utilisez un éditeur de texte pour créer la commande RMAN, puis coupez-la et collez-la.

```
RMAN> run {
2> set newname for tempfile 1 to '/oradata/TOAST/temp01.dbf';
3> switch tempfile all;
4> }
executing command: SET NEWNAME
renamed tempfile 1 to /oradata/TOAST/temp01.dbf in control file
```

Migration du journal de reprise

Le processus de migration est presque terminé, mais les journaux de reprise se trouvent toujours sur le groupe de disques ASM d'origine. Les journaux de reprise ne peuvent pas être transférés directement. Un nouvel ensemble de journaux de reprise est créé et ajouté à la configuration, suivant un DROP des anciens journaux.

1. Identifiez le nombre de groupes de fichiers redo log et leurs numéros de groupe respectifs.

```
RMAN> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
--------------------------------------------------------------------------------
--------
1 +ASM0/TOAST/redo01.log
2 +ASM0/TOAST/redo02.log
3 +ASM0/TOAST/redo03.log
```

2. Indiquez la taille des journaux de reprise.

```
RMAN> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
--------------------------------------------------------------------------------
--------
1 52428800
2 52428800
3 52428800
```

3. Pour chaque fichier redo log, créez un groupe en utilisant la même taille que le groupe de fichiers redo log actuel à l'aide du nouvel emplacement du système de fichiers.

```
RMAN> alter database add logfile '/logs/TOAST/redo/log00.rdo' size
52428800;
Statement processed
RMAN> alter database add logfile '/logs/TOAST/redo/log01.rdo' size
52428800;
Statement processed
RMAN> alter database add logfile '/logs/TOAST/redo/log02.rdo' size
52428800;
Statement processed
```

4. Supprimez les anciens groupes de fichiers journaux qui se trouvent toujours sur le stockage précédent.

```
RMAN> alter database drop logfile group 4;
Statement processed
RMAN> alter database drop logfile group 5;
Statement processed
RMAN> alter database drop logfile group 6;
Statement processed
```

5. Si une erreur bloque la suppression d'un journal actif, forcez un commutateur au journal suivant pour libérer le verrouillage et forcer un point de contrôle global. Un exemple est illustré ci-dessous. La tentative de suppression du groupe de fichiers journaux 3, qui se trouvait sur l'ancien emplacement, a été refusée

parce qu'il y avait encore des données actives dans ce fichier journal. L'archivage des journaux suivi d'un point de contrôle permet la suppression des fichiers journaux.

```
RMAN> alter database drop logfile group 4;
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03002: failure of sql statement command at 12/08/2015 20:23:51
ORA-01623: log 4 is current log for instance TOAST (thread 4) - cannot
drop
ORA-00312: online log 4 thread 1:
'+NEWLOGS/TOAST/ONLINELOG/group_4.266.897763123'
RMAN> alter system switch logfile;
Statement processed
RMAN> alter system checkpoint;
Statement processed
RMAN> alter database drop logfile group 4;
Statement processed
```

6. Vérifiez l'environnement pour vous assurer que tous les paramètres basés sur l'emplacement sont mis à jour.

```
SQL> select name from v$datafile;
SQL> select member from v$logfile;
SQL> select name from v$tempfile;
SQL> show parameter spfile;
SQL> select name, value from v$parameter where value is not null;
```

7. Le script suivant explique comment faciliter ce processus.

```
[root@jfsc1 current]# ./checkdbdata.pl TOAST
TOAST datafiles:
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSTEM_FNO-1_01r5fhjg
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSAUX_FNO-2_02r5fhjo
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-UNDOTBS1_FNO-3_03r5fhjt
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-USERS_FNO-4_05r5fhk6
TOAST redo logs:
/logs/TOAST/redo/log00.rdo
/logs/TOAST/redo/log01.rdo
/logs/TOAST/redo/log02.rdo
TOAST temp datafiles:
/oradata/TOAST/temp01.dbf
TOAST spfile
spfile                                string
/orabin/product/12.1.0/dbhome_
                                                    1/dbs/spfileTOAST.ora
TOAST key parameters
control_files /logs/TOAST/arch/control01.ctl,
/logs/TOAST/redo/control02.ctl
log_archive_dest_1 LOCATION=/logs/TOAST/arch
```

8. Si les groupes de disques ASM ont été complètement évacués, ils peuvent maintenant être démontés avec `asmcmd`. Dans de nombreux cas, les fichiers appartenant à d'autres bases de données ou au fichier ASM spfile/passwd peuvent toujours être présents.

```
-bash-4.1$ . oraenv
ORACLE_SID = [TOAST] ? +ASM
The Oracle base remains unchanged with value /orabin
-bash-4.1$ asmcmd
ASMCMD> umount DATA
ASMCMD>
```

Procédure de nettoyage du fichier de données

Le processus de migration peut donner lieu à des fichiers de données avec une syntaxe longue ou chiffrée, selon la façon dont Oracle RMAN a été utilisé. Dans l'exemple illustré ici, la sauvegarde a été effectuée avec le format de fichier de `/oradata/TOAST/%U`. `%U` Indique que RMAN doit créer un nom unique par défaut pour chaque fichier de données. Le résultat est similaire à ce qui est affiché dans le texte suivant. Les noms traditionnels des fichiers de données sont incorporés dans les noms. Pour ce faire, utilisez l'approche par script illustrée à la "Nettoyage de migration ASM".

```
[root@jfsc1 current]# ./fixuniquenames.pl TOAST
#sqlplus Commands
shutdown immediate;
startup mount;
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSTEM_FNO-1_01r5fhjg
/oradata/TOAST/system.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSAUX_FNO-2_02r5fhjo
/oradata/TOAST/sysaux.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-UNDOTBS1_FNO-
3_03r5fhjt /oradata/TOAST/undotbs1.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-USERS_FNO-4_05r5fhk6
/oradata/TOAST/users.dbf
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
SYSTEM_FNO-1_01r5fhjg' to '/oradata/TOAST/system.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
SYSAUX_FNO-2_02r5fhjo' to '/oradata/TOAST/sysaux.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
UNDOTBS1_FNO-3_03r5fhjt' to '/oradata/TOAST/undotbs1.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
USERS_FNO-4_05r5fhk6' to '/oradata/TOAST/users.dbf';
alter database open;
```

Rééquilibrage d'Oracle ASM

Comme nous l'avons vu précédemment, un groupe de disques Oracle ASM peut être migré en toute transparence vers un nouveau système de stockage en utilisant le processus de rééquilibrage. En résumé, le processus de rééquilibrage nécessite l'ajout de LUN de taille égale au groupe existant de LUN, suivi d'une opération de DROP de la LUN précédente. Oracle ASM déplace automatiquement les données sous-jacentes vers un nouveau stockage selon une disposition optimale, puis libère les anciens LUN une fois l'opération terminée.

Le processus de migration utilise des E/S séquentielles efficaces et ne provoque généralement aucune interruption des performances. En revanche, le taux de migration peut être ralenti lorsque cela est nécessaire.

Identifiez les données à migrer

```
SQL> select name||' '||group_number||' '||total_mb||' '||path||'
'||header_status from v$asm_disk;
NEWDATA_0003 1 10240 /dev/mapper/3600a098038303537762b47594c315864 MEMBER
NEWDATA_0002 1 10240 /dev/mapper/3600a098038303537762b47594c315863 MEMBER
NEWDATA_0000 1 10240 /dev/mapper/3600a098038303537762b47594c315861 MEMBER
NEWDATA_0001 1 10240 /dev/mapper/3600a098038303537762b47594c315862 MEMBER
SQL> select group_number||' '||name from v$asm_diskgroup;
1 NEWDATA
```

Créer des LUN

Créez de nouvelles LUN de la même taille et définissez l'appartenance des utilisateurs et des groupes selon les besoins. Les LUN doivent s'afficher comme `CANDIDATE` disques.

```
SQL> select name||' '||group_number||' '||total_mb||' '||path||'
'||header_status from v$asm_disk;
 0 0 /dev/mapper/3600a098038303537762b47594c31586b CANDIDATE
 0 0 /dev/mapper/3600a098038303537762b47594c315869 CANDIDATE
 0 0 /dev/mapper/3600a098038303537762b47594c315858 CANDIDATE
 0 0 /dev/mapper/3600a098038303537762b47594c31586a CANDIDATE
NEWDATA_0003 1 10240 /dev/mapper/3600a098038303537762b47594c315864 MEMBER
NEWDATA_0002 1 10240 /dev/mapper/3600a098038303537762b47594c315863 MEMBER
NEWDATA_0000 1 10240 /dev/mapper/3600a098038303537762b47594c315861 MEMBER
NEWDATA_0001 1 10240 /dev/mapper/3600a098038303537762b47594c315862 MEMBER
```

Ajouter de nouvelles LUN

Même si les opérations d'ajout et de suppression peuvent être effectuées ensemble, il est généralement plus facile d'ajouter de nouvelles LUN en deux étapes. Commencez par ajouter les nouvelles LUN au groupe de disques. Cette étape entraîne la migration de la moitié des extensions des LUN ASM actuelles vers les nouvelles LUN.

La puissance de rééquilibrage indique la vitesse à laquelle les données sont transférées. Plus le nombre est élevé, plus le parallélisme du transfert de données est élevé. La migration s'effectue au moyen d'opérations d'E/S séquentielles efficaces, peu susceptibles d'entraîner des problèmes de performances. Toutefois, si nécessaire, le pouvoir de rééquilibrage d'une migration en cours peut être ajusté avec le `alter diskgroup [name] rebalance power [level]` commande. Les migrations types utilisent une valeur de 5.

```
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c31586b' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c315869' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c315858' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c31586a' rebalance power 5;
Diskgroup altered.
```

Surveiller le fonctionnement

Une opération de rééquilibrage peut être contrôlée et gérée de plusieurs manières. Nous avons utilisé la commande suivante dans cet exemple.

```
SQL> select group_number,operation,state from v$asm_operation;
GROUP_NUMBER OPERA STAT
------------ ----- ----
           1 REBAL RUN
           1 REBAL WAIT
```

Une fois la migration terminée, aucune opération de rééquilibrage n'est signalée.

```
SQL> select group_number,operation,state from v$asm_operation;
no rows selected
```

Supprimez les anciennes LUN

La migration est maintenant terminée à mi-chemin. Il peut être souhaitable d'effectuer quelques tests de performances de base pour s'assurer que l'environnement est sain. Après confirmation, les données restantes peuvent être déplacées en déposant les anciennes LUN. Notez que cela ne provoque pas la publication immédiate des LUN. L'opération de DROP indique à Oracle ASM de déplacer d'abord les extensions, puis de libérer la LUN.

```
sqlplus / as sysasm
SQL> alter diskgroup NEWDATA drop disk NEWDATA_0000 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA drop disk NEWDATA_0001 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup newdata drop disk NEWDATA_0002 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup newdata drop disk NEWDATA_0003 rebalance power 5;
Diskgroup altered.
```

Surveiller le fonctionnement

L'opération de rééquilibrage peut être contrôlée et gérée de plusieurs manières. Nous avons utilisé la commande suivante dans cet exemple :

```
SQL> select group_number,operation,state from v$asm_operation;
GROUP_NUMBER OPERA STAT
------------ ----- ----
           1 REBAL RUN
           1 REBAL WAIT
```

Une fois la migration terminée, aucune opération de rééquilibrage n'est signalée.

```
SQL> select group_number,operation,state from v$asm_operation;
no rows selected
```

Supprimer les anciens LUN

Avant de supprimer les anciennes LUN du groupe de disques, vous devez effectuer une dernière vérification de l'état de l'en-tête. Une fois qu'une LUN est libérée d'ASM, son nom n'est plus répertorié et son état est répertorié comme `FORMER`. Cela signifie que ces LUN peuvent être supprimées du système en toute sécurité.

```
SQL> select name||' '||group_number||' '||total_mb||' '||path||'
'||header_status from v$asm_disk;
NAME||''||GROUP_NUMBER||''||TOTAL_MB||''||PATH||''||HEADER_STATUS
------------------------------------------------------------------------
------
 0 0 /dev/mapper/3600a098038303537762b47594c315863 FORMER
 0 0 /dev/mapper/3600a098038303537762b47594c315864 FORMER
 0 0 /dev/mapper/3600a098038303537762b47594c315861 FORMER
 0 0 /dev/mapper/3600a098038303537762b47594c315862 FORMER
NEWDATA_0005 1 10240 /dev/mapper/3600a098038303537762b47594c315869 MEMBER
NEWDATA_0007 1 10240 /dev/mapper/3600a098038303537762b47594c31586a MEMBER
NEWDATA_0004 1 10240 /dev/mapper/3600a098038303537762b47594c31586b MEMBER
NEWDATA_0006 1 10240 /dev/mapper/3600a098038303537762b47594c315858 MEMBER
8 rows selected.
```

Migration LVM

La procédure présentée ici présente les principes d'une migration basée sur LVM d'un groupe de volumes appelé `datavg`. Les exemples sont tirés du LVM Linux, mais les principes s'appliquent également à AIX, HP-UX et VxVM. Les commandes précises peuvent varier.

1. Identifiez les LUN actuellement dans le `datavg` groupe de volumes.

```
[root@host1 ~]# pvdisplay -C | grep datavg
  /dev/mapper/3600a098038303537762b47594c31582f datavg lvm2 a--  10.00g
10.00g
  /dev/mapper/3600a098038303537762b47594c31585a datavg lvm2 a--  10.00g
10.00g
  /dev/mapper/3600a098038303537762b47594c315859 datavg lvm2 a--  10.00g
10.00g
  /dev/mapper/3600a098038303537762b47594c31586c datavg lvm2 a--  10.00g
10.00g
```

2. Créez de nouvelles LUN de taille physique identique ou légèrement supérieure et définissez-les comme volumes physiques.

```
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315864
  Physical volume "/dev/mapper/3600a098038303537762b47594c315864"
successfully created
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315863
  Physical volume "/dev/mapper/3600a098038303537762b47594c315863"
successfully created
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315862
  Physical volume "/dev/mapper/3600a098038303537762b47594c315862"
successfully created
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315861
  Physical volume "/dev/mapper/3600a098038303537762b47594c315861"
successfully created
```

3. Ajoutez les nouveaux volumes au groupe de volumes.

```
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315864
  Volume group "datavg" successfully extended
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315863
  Volume group "datavg" successfully extended
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315862
  Volume group "datavg" successfully extended
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315861
  Volume group "datavg" successfully extended
```

4. Émettez le `pvmove` Commande permettant de déplacer les extensions de chaque LUN actuelle vers la nouvelle LUN. Le `-i [seconds]` l'argument surveille la progression de l'opération.

```
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c31582f
/dev/mapper/3600a098038303537762b47594c315864
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 14.2%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 28.4%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 42.5%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 57.1%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 72.3%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 87.3%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 100.0%
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c31585a
/dev/mapper/3600a098038303537762b47594c315863
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 14.9%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 29.9%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 44.8%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 60.1%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 75.8%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 90.9%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 100.0%
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c315859
/dev/mapper/3600a098038303537762b47594c315862
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 14.8%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 29.8%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 45.5%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 61.1%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 76.6%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 91.7%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 100.0%
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c31586c
/dev/mapper/3600a098038303537762b47594c315861
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 15.0%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 30.4%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 46.0%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 61.4%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 77.2%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 92.3%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 100.0%
```

5. Une fois ce processus terminé, supprimez les anciennes LUN du groupe de volumes à l'aide du `vgreduce` commande. En cas de réussite, la LUN peut être supprimée en toute sécurité du système.

```
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c31582f
Removed "/dev/mapper/3600a098038303537762b47594c31582f" from volume
group "datavg"
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c31585a
  Removed "/dev/mapper/3600a098038303537762b47594c31585a" from volume
group "datavg"
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c315859
  Removed "/dev/mapper/3600a098038303537762b47594c315859" from volume
group "datavg"
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c31586c
  Removed "/dev/mapper/3600a098038303537762b47594c31586c" from volume
group "datavg"
```

Importation de LUN étrangères

Planification

Les procédures de migration des ressources SAN à l'aide de FLI sont décrites dans NetApp "Documentation relative à l'importation de LUN étrangères de ONTAP" .

Du point de vue de la base de données et de l'hôte, aucune étape particulière n'est requise. Une fois les zones FC mises à jour et les LUN disponibles sur ONTAP, LVM doit pouvoir lire les métadonnées LVM des LUN. De plus, les groupes de volumes sont prêts à être utilisés sans étape de configuration supplémentaire. Dans de rares cas, les environnements peuvent inclure des fichiers de configuration codés en dur avec des références à la baie de stockage précédente. Par exemple, un système Linux inclus `/etc/multipath.conf` Les règles qui référençaient un WWN d'un périphérique donné doivent être mises à jour pour refléter les modifications introduites par FLI.

> Reportez-vous à la matrice de compatibilité NetApp pour plus d'informations sur les configurations prises en charge. Si votre environnement n'est pas inclus, contactez votre représentant NetApp pour obtenir de l'aide.

Cet exemple montre la migration des LUN ASM et LVM hébergées sur un serveur Linux. FLI est pris en charge par d'autres systèmes d'exploitation. Bien que les commandes côté hôte puissent différer, les principes sont les mêmes et les procédures ONTAP sont identiques.

Identifier les LUN LVM

La première étape de la préparation consiste à identifier les LUN à migrer. Dans l'exemple illustré ici, deux systèmes de fichiers SAN sont montés sur `/orabin` et `/backups`.

```
[root@host1 ~]# df -k
Filesystem                          1K-blocks      Used Available Use%
Mounted on
/dev/mapper/rhel-root                52403200   8811464  43591736  17% /
devtmpfs                             65882776         0  65882776   0% /dev
...
fas8060-nfs-public:/install         199229440 119368128  79861312  60%
/install
/dev/mapper/sanvg-lvorabin           20961280  12348476   8612804  59%
/orabin
/dev/mapper/sanvg-lvbackups          73364480  62947536  10416944  86%
/backups
```

Le nom du groupe de volumes peut être extrait du nom du périphérique, qui utilise le format (nom du groupe de volumes)-(nom du volume logique). Dans ce cas, le groupe de volumes est appelé `sanvg`.

Le `pvdisplay` Vous pouvez utiliser la commande suivante pour identifier les LUN qui prennent en charge ce groupe de volumes. Dans ce cas, 10 LUN constituent le `sanvg` groupe de volumes.

```
[root@host1 ~]# pvdisplay -C -o pv_name,pv_size,pv_fmt,vg_name
  PV                                                PSize   VG
  /dev/mapper/3600a0980383030445424487556574266  10.00g sanvg
  /dev/mapper/3600a0980383030445424487556574267  10.00g sanvg
  /dev/mapper/3600a0980383030445424487556574268  10.00g sanvg
  /dev/mapper/3600a0980383030445424487556574269  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426a  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426b  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426c  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426d  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426e  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426f  10.00g sanvg
  /dev/sda2                                       278.38g rhel
```

Identifier les LUN ASM

Les LUN ASM doivent également être migrés. Pour obtenir le nombre de LUN et de chemins de LUN depuis sqlplus en tant qu'utilisateur sysasm, exécutez la commande suivante :

```
SQL> select path||' '||os_mb from v$asm_disk;
PATH||''||OS_MB
--------------------------------------------------------------------------
------
/dev/oracleasm/disks/ASM0 10240
/dev/oracleasm/disks/ASM9 10240
/dev/oracleasm/disks/ASM8 10240
/dev/oracleasm/disks/ASM7 10240
/dev/oracleasm/disks/ASM6 10240
/dev/oracleasm/disks/ASM5 10240
/dev/oracleasm/disks/ASM4 10240
/dev/oracleasm/disks/ASM1 10240
/dev/oracleasm/disks/ASM3 10240
/dev/oracleasm/disks/ASM2 10240
10 rows selected.
SQL>
```

Modifications du réseau FC

L'environnement actuel contient 20 LUN à migrer. Mettez à jour le SAN actuel de sorte que ONTAP puisse accéder aux LUN actuelles. Les données n'ont pas encore été migrées, mais ONTAP doit lire les informations de configuration des LUN actuelles pour créer le nouveau home pour ces données.

Au moins un port HBA sur le système AFF/FAS doit être configuré en tant que port initiateur. En outre, les zones FC doivent être mises à jour de sorte que ONTAP puisse accéder aux LUN de la baie de stockage étrangère. Certaines baies de stockage ont configuré le masquage des LUN, ce qui limite les WWN pouvant accéder à une LUN donnée. Dans ce cas, le masquage de LUN doit également être mis à jour pour autoriser l'accès aux WWN de ONTAP.

Une fois cette étape terminée, ONTAP doit être en mesure d'afficher la baie de stockage étrangère avec le `storage array show` commande. Le champ de clé renvoyé est le préfixe utilisé pour identifier la LUN étrangère sur le système. Dans l'exemple ci-dessous, les LUN de la baie étrangère `FOREIGN_1` Apparaissent dans ONTAP en utilisant le préfixe de `FOR-1`.

Identifiez le tableau étranger

```
Cluster01::> storage array show -fields name,prefix
name          prefix
------------- ------
FOREIGN_1     FOR-1
Cluster01::>
```

Identifiez les LUN étrangères

Vous pouvez lister les LUN en transmettant le `array-name` à la `storage disk show` commande. Les données renvoyées sont référencées plusieurs fois pendant la procédure de migration.

```
Cluster01::> storage disk show -array-name FOREIGN_1 -fields disk,serial
disk     serial-number
-------- -------------
FOR-1.1  800DT$HuVWBX
FOR-1.2  800DT$HuVWBZ
FOR-1.3  800DT$HuVWBW
FOR-1.4  800DT$HuVWBY
FOR-1.5  800DT$HuVWB/
FOR-1.6  800DT$HuVWBa
FOR-1.7  800DT$HuVWBd
FOR-1.8  800DT$HuVWBb
FOR-1.9  800DT$HuVWBc
FOR-1.10 800DT$HuVWBe
FOR-1.11 800DT$HuVWBf
FOR-1.12 800DT$HuVWBg
FOR-1.13 800DT$HuVWBi
FOR-1.14 800DT$HuVWBh
FOR-1.15 800DT$HuVWBj
FOR-1.16 800DT$HuVWBk
FOR-1.17 800DT$HuVWBm
FOR-1.18 800DT$HuVWBl
FOR-1.19 800DT$HuVWBo
FOR-1.20 800DT$HuVWBn
20 entries were displayed.
Cluster01::>
```

Enregistrer des LUN de baies étrangères en tant que candidats à l'importation

Les LUN étrangères sont initialement classées comme tout type de LUN particulier. Avant de pouvoir importer des données, les LUN doivent être marquées comme étrangères et par conséquent comme candidates au processus d'importation. Cette étape est terminée en transmettant le numéro de série au `storage disk modify` comme indiqué dans l'exemple suivant. Notez que ce processus balise uniquement la LUN comme étant étrangère dans ONTAP. Aucune donnée n'est écrite sur la LUN étrangère elle-même.

```
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBW} -is
-foreign true
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBX} -is
-foreign true
...
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBn} -is
-foreign true
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBo} -is
-foreign true
Cluster01::*>
```

Création de volumes pour héberger les LUN migrés

Un volume est nécessaire pour héberger les LUN migrées. La configuration exacte du volume dépend du plan global d'exploitation des fonctionnalités ONTAP. Dans cet exemple, les LUN ASM sont placées dans un volume et les LUN LVM sont placées dans un second volume. Vous pouvez ainsi gérer les LUN en tant que groupes indépendants à des fins telles que la hiérarchisation, la création de snapshots ou la définition de contrôles de QoS.

Réglez le `snapshot-policy `to `none`. Le processus de migration peut inclure une grande partie du transfert des données. Par conséquent, si des snapshots sont créés par accident, la consommation d'espace peut augmenter de façon importante, car des données indésirables sont capturées dans les snapshots.

```
Cluster01::> volume create -volume new_asm -aggregate data_02 -size 120G
-snapshot-policy none
[Job 1152] Job succeeded: Successful
Cluster01::> volume create -volume new_lvm -aggregate data_02 -size 120G
-snapshot-policy none
[Job 1153] Job succeeded: Successful
Cluster01::>
```

Créer des LUN ONTAP

Une fois les volumes créés, les nouvelles LUN doivent être créées. Normalement, la création d'une LUN nécessite que l'utilisateur indique des informations telles que la taille de LUN, mais dans ce cas, l'argument disque étranger est transmis à la commande. Par conséquent, ONTAP réplique les données de configuration actuelle du LUN à partir du numéro de série spécifié. Il utilise également la géométrie des LUN et les données de la table de partition pour ajuster l'alignement des LUN et établir des performances optimales.

Dans cette étape, les numéros de série doivent être référencés avec le tableau étranger pour s'assurer que le LUN étranger correct est associé au nouveau LUN correct.

```
Cluster01::*> lun create -vserver vserver1 -path /vol/new_asm/LUN0 -ostype
linux -foreign-disk 800DT$HuVWBW
Created a LUN of size 10g (10737418240)
Cluster01::*> lun create -vserver vserver1 -path /vol/new_asm/LUN1 -ostype
linux -foreign-disk 800DT$HuVWBX
Created a LUN of size 10g (10737418240)
...
Created a LUN of size 10g (10737418240)
Cluster01::*> lun create -vserver vserver1 -path /vol/new_lvm/LUN8 -ostype
linux -foreign-disk 800DT$HuVWBn
Created a LUN of size 10g (10737418240)
Cluster01::*> lun create -vserver vserver1 -path /vol/new_lvm/LUN9 -ostype
linux -foreign-disk 800DT$HuVWBo
Created a LUN of size 10g (10737418240)
```

Créer des relations d'importation

Les LUN ont été créées, mais ne sont pas configurées en tant que destination de réplication. Avant de pouvoir réaliser cette étape, les LUN doivent d'abord être mises hors ligne. Cette étape supplémentaire est conçue pour protéger les données contre les erreurs de l'utilisateur. Si ONTAP permettait l'exécution d'une migration sur une LUN en ligne, une erreur typographique risquerait d'écraser les données actives. L'étape supplémentaire consistant à forcer l'utilisateur à mettre d'abord une LUN hors ligne permet de vérifier que la LUN cible correcte est utilisée comme destination de migration.

```
Cluster01::*> lun offline -vserver vserver1 -path /vol/new_asm/LUN0
Warning: This command will take LUN "/vol/new_asm/LUN0" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
Cluster01::*> lun offline -vserver vserver1 -path /vol/new_asm/LUN1
Warning: This command will take LUN "/vol/new_asm/LUN1" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
...
Warning: This command will take LUN "/vol/new_lvm/LUN8" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
Cluster01::*> lun offline -vserver vserver1 -path /vol/new_lvm/LUN9
Warning: This command will take LUN "/vol/new_lvm/LUN9" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
```

Une fois les LUN hors ligne, vous pouvez établir la relation d'importation en transmettant le numéro de série de la LUN étrangère à `lun import create` commande.

```
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_asm/LUN0
-foreign-disk 800DT$HuVWBW
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_asm/LUN1
-foreign-disk 800DT$HuVWBX
...
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_lvm/LUN8
-foreign-disk 800DT$HuVWBn
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_lvm/LUN9
-foreign-disk 800DT$HuVWBo
Cluster01::*>
```

Une fois toutes les relations d'importation établies, les LUN peuvent être remis en ligne.

```
Cluster01::*> lun online -vserver vserver1 -path /vol/new_asm/LUN0
Cluster01::*> lun online -vserver vserver1 -path /vol/new_asm/LUN1
...
Cluster01::*> lun online -vserver vserver1 -path /vol/new_lvm/LUN8
Cluster01::*> lun online -vserver vserver1 -path /vol/new_lvm/LUN9
Cluster01::*>
```

Créer le groupe initiateur

Un groupe initiateur (igroup) fait partie de l'architecture de masquage des LUN ONTAP. L'accès à une LUN nouvellement créée n'est pas accessible à moins qu'un hôte ne bénéficie au préalable d'un accès. Pour ce faire, vous devez créer un groupe initiateur qui répertorie les WWN FC ou les noms d'initiateurs iSCSI auxquels l'accès doit être accordé. Au moment de la rédaction de ce rapport, FLI était pris en charge uniquement pour les LUN FC. Cependant, la conversion en iSCSI après migration est une tâche simple, comme illustré dans la "Conversion de protocoles".

Dans cet exemple, un groupe initiateur est créé et contient deux WWN correspondant aux deux ports disponibles sur l'adaptateur HBA de l'hôte.

```
Cluster01::*> igroup create linuxhost -protocol fcp -ostype linux
-initiator 21:00:00:0e:1e:16:63:50 21:00:00:0e:1e:16:63:51
```

Mappez les nouvelles LUN sur l'hôte

Après la création du groupe initiateur, les LUN sont ensuite mappées sur le groupe initiateur défini. Ces LUN sont uniquement disponibles pour les WWN inclus dans ce groupe initiateur. NetApp suppose, à ce stade du processus de migration, que l'hôte n'a pas été segmenté vers ONTAP. Cela est important, car si l'hôte est segmenté simultanément sur la baie étrangère et le nouveau système ONTAP, il est possible de détecter sur chaque baie des LUN portant le même numéro de série. Cette situation peut entraîner des dysfonctionnements des chemins d'accès multiples ou endommager les données.

```
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN0 -igroup
linuxhost
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN1 -igroup
linuxhost
...
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN8 -igroup
linuxhost
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN9 -igroup
linuxhost
Cluster01::*>
```

Mise en service

Certaines perturbations lors de l'importation d'une LUN étrangère sont inévitables en raison de la nécessité de modifier la configuration du réseau FC. Cependant,

l'interruption ne doit pas durer beaucoup plus longtemps que le temps nécessaire pour redémarrer l'environnement de base de données et mettre à jour la segmentation FC pour basculer la connectivité FC de l'hôte de la LUN étrangère vers ONTAP.

Ce processus peut être résumé comme suit :

1. Mettez toutes les activités de LUN au repos sur les LUN étrangères.
2. Rediriger les connexions FC de l'hôte vers le nouveau système ONTAP.
3. Déclencher le processus d'importation.
4. Redécouvrez les LUN.
5. Redémarrez la base de données.

Inutile d'attendre la fin du processus de migration. Dès que la migration d'une LUN donnée commence, celle-ci est disponible sur ONTAP et peut assurer le service des données pendant que le processus de copie des données se poursuit. Toutes les lectures sont transmises au LUN étranger et toutes les écritures sont écrites de manière synchrone sur les deux baies. L'opération de copie est très rapide et la surcharge liée à la redirection du trafic FC est minimale. Par conséquent, tout impact sur les performances doit être transitoire et minimal. En cas de problème, vous pouvez retarder le redémarrage de l'environnement jusqu'à ce que le processus de migration soit terminé et que les relations d'importation aient été supprimées.

Arrêtez la base de données

Dans cet exemple, la première étape de la mise en veille de l'environnement consiste à arrêter la base de données.

```
[oracle@host1 bin]$ . oraenv
ORACLE_SID = [oracle] ? FLIDB
The Oracle base remains unchanged with value /orabin
[oracle@host1 bin]$ sqlplus / as sysdba
SQL*Plus: Release 12.1.0.2.0
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, Automatic Storage Management, OLAP, Advanced
Analytics
and Real Application Testing options
SQL> shutdown immediate;
Database closed.
Database dismounted.
ORACLE instance shut down.
SQL>
```

Fermez les services de grille

L'un des systèmes de fichiers SAN en cours de migration inclut également les services Oracle ASM. La mise en veille des LUN sous-jacentes nécessite la suspension des systèmes de fichiers, ce qui signifie l'arrêt des processus avec des fichiers ouverts sur ce système de fichiers.

```
[oracle@host1 bin]$ ./crsctl stop has -f
CRS-2791: Starting shutdown of Oracle High Availability Services-managed
resources on 'host1'
CRS-2673: Attempting to stop 'ora.evmd' on 'host1'
CRS-2673: Attempting to stop 'ora.DATA.dg' on 'host1'
CRS-2673: Attempting to stop 'ora.LISTENER.lsnr' on 'host1'
CRS-2677: Stop of 'ora.DATA.dg' on 'host1' succeeded
CRS-2673: Attempting to stop 'ora.asm' on 'host1'
CRS-2677: Stop of 'ora.LISTENER.lsnr' on 'host1' succeeded
CRS-2677: Stop of 'ora.evmd' on 'host1' succeeded
CRS-2677: Stop of 'ora.asm' on 'host1' succeeded
CRS-2673: Attempting to stop 'ora.cssd' on 'host1'
CRS-2677: Stop of 'ora.cssd' on 'host1' succeeded
CRS-2793: Shutdown of Oracle High Availability Services-managed resources
on 'host1' has completed
CRS-4133: Oracle High Availability Services has been stopped.
[oracle@host1 bin]$
```

Démonter les systèmes de fichiers

Si tous les processus sont arrêtés, l'opération de montage a réussi. Si l'autorisation est refusée, il doit y avoir un processus avec un verrou sur le système de fichiers. Le `fuser` permet d'identifier ces processus.

```
[root@host1 ~]# umount /orabin
[root@host1 ~]# umount /backups
```

Désactiver les groupes de volumes

Une fois tous les systèmes de fichiers d'un groupe de volumes donné démontés, le groupe de volumes peut être désactivé.

```
[root@host1 ~]# vgchange --activate n sanvg
  0 logical volume(s) in volume group "sanvg" now active
[root@host1 ~]#
```

Modifications du réseau FC

Les zones FC peuvent maintenant être mises à jour pour supprimer tout accès de l'hôte à la baie étrangère et établir l'accès à ONTAP.

Démarrer le processus d'importation

Pour démarrer les processus d'importation de LUN, exécutez `lun import start` commande.

```
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_asm/LUN0
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_asm/LUN1
...
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_lvm/LUN8
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_lvm/LUN9
Cluster01::lun import*>
```

Surveiller la progression de l'importation

L'opération d'importation peut être surveillée avec `lun import show` commande. Comme indiqué ci-dessous, l'importation des 20 LUN est en cours, ce qui signifie que les données sont désormais accessibles via ONTAP, même si la copie des données progresse.

```
Cluster01::lun import*> lun import show -fields path,percent-complete
vserver   foreign-disk path                 percent-complete
--------- ------------ ------------------ ----------------
vserver1  800DT$HuVWB/ /vol/new_asm/LUN4 5
vserver1  800DT$HuVWBW /vol/new_asm/LUN0 5
vserver1  800DT$HuVWBX /vol/new_asm/LUN1 6
vserver1  800DT$HuVWBY /vol/new_asm/LUN2 6
vserver1  800DT$HuVWBZ /vol/new_asm/LUN3 5
vserver1  800DT$HuVWBa /vol/new_asm/LUN5 4
vserver1  800DT$HuVWBb /vol/new_asm/LUN6 4
vserver1  800DT$HuVWBc /vol/new_asm/LUN7 4
vserver1  800DT$HuVWBd /vol/new_asm/LUN8 4
vserver1  800DT$HuVWBe /vol/new_asm/LUN9 4
vserver1  800DT$HuVWBf /vol/new_lvm/LUN0 5
vserver1  800DT$HuVWBg /vol/new_lvm/LUN1 4
vserver1  800DT$HuVWBh /vol/new_lvm/LUN2 4
vserver1  800DT$HuVWBi /vol/new_lvm/LUN3 3
vserver1  800DT$HuVWBj /vol/new_lvm/LUN4 3
vserver1  800DT$HuVWBk /vol/new_lvm/LUN5 3
vserver1  800DT$HuVWBl /vol/new_lvm/LUN6 4
vserver1  800DT$HuVWBm /vol/new_lvm/LUN7 3
vserver1  800DT$HuVWBn /vol/new_lvm/LUN8 2
vserver1  800DT$HuVWBo /vol/new_lvm/LUN9 2
20 entries were displayed.
```

Si vous avez besoin d'un processus hors ligne, retardez la redécouverte ou le redémarrage des services jusqu'à ce que la `lun import show` commande indique que la migration a abouti. Vous pouvez ensuite terminer le processus de migration comme décrit à la section "Importation de LUN étrangères—fin".

Si vous avez besoin d'une migration en ligne, redécouvrez les LUN de leur nouveau domicile et accédez aux services.

Recherchez les modifications de périphérique SCSI

Dans la plupart des cas, l'option la plus simple pour redécouvrir de nouvelles LUN consiste à redémarrer l'hôte. Cela supprime automatiquement les anciens périphériques obsolètes, détecte correctement toutes les nouvelles LUN et construit les périphériques associés, tels que les périphériques multivoies. L'exemple ci-dessous montre un processus entièrement en ligne à des fins de démonstration.

Attention : avant de redémarrer un hôte, assurez-vous que toutes les entrées dans `/etc/fstab` Les ressources SAN migrées de cette référence sont commentées. Si ce n'est pas le cas et si des problèmes surviennent lors de l'accès aux LUN, le système d'exploitation risque de ne pas démarrer. Cette situation n'endommage pas les données. Cependant, il peut être très peu commode de démarrer en mode de secours ou un mode similaire et de corriger le `/etc/fstab` Afin que le système d'exploitation puisse être démarré pour permettre le dépannage.

Les LUN de la version de Linux utilisée dans cet exemple peuvent être renumérisées avec `rescan-scsi-bus.sh` commande. Si la commande réussit, chaque chemin de LUN doit apparaître dans le résultat de la commande. Le résultat de cette commande peut être difficile à interpréter, mais si la configuration de zoning et d'igroup était correcte, de nombreuses LUN doivent apparaître et inclure un `NETAPP` chaîne du fournisseur.

```
[root@host1 /]# rescan-scsi-bus.sh
Scanning SCSI subsystem for new devices
Scanning host 0 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
 Scanning for device 0 2 0 0 ...
OLD: Host: scsi0 Channel: 02 Id: 00 Lun: 00
      Vendor: LSI      Model: RAID SAS 6G 0/1  Rev: 2.13
      Type:   Direct-Access                    ANSI SCSI revision: 05
Scanning host 1 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
 Scanning for device 1 0 0 0 ...
OLD: Host: scsi1 Channel: 00 Id: 00 Lun: 00
      Vendor: Optiarc  Model: DVD RW AD-7760H  Rev: 1.41
      Type:   CD-ROM                           ANSI SCSI revision: 05
Scanning host 2 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 3 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 4 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 5 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 6 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 7 for  all SCSI target IDs, all LUNs
 Scanning for device 7 0 0 10 ...
OLD: Host: scsi7 Channel: 00 Id: 00 Lun: 10
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
 Scanning for device 7 0 0 11 ...
OLD: Host: scsi7 Channel: 00 Id: 00 Lun: 11
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
 Scanning for device 7 0 0 12 ...
...
OLD: Host: scsi9 Channel: 00 Id: 01 Lun: 18
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
 Scanning for device 9 0 1 19 ...
OLD: Host: scsi9 Channel: 00 Id: 01 Lun: 19
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
0 new or changed device(s) found.
0 remapped or resized device(s) found.
0 device(s) removed.
```

Vérifiez la présence de périphériques multivoies

Le processus de découverte des LUN déclenche également la recréation des périphériques multivoies, mais il est connu que le pilote de chemins d'accès multiples Linux présente des problèmes occasionnels. La sortie de `multipath -ll` doit être vérifié pour vérifier que la sortie semble correcte. Par exemple, le résultat ci-dessous affiche les périphériques à chemins d'accès multiples associés à un `NETAPP` chaîne du fournisseur. Chaque périphérique a quatre chemins, dont deux avec une priorité de 50 et deux avec une priorité de 10.

Bien que le résultat exact puisse varier selon les versions de Linux, ce résultat semble normal.

Reportez-vous à la documentation des utilitaires hôtes pour connaître la version de Linux que vous utilisez pour vérifier que l' `/etc/multipath.conf` les paramètres sont corrects.

```
[root@host1 /]# multipath -ll
3600a098038303558735d493762504b36 dm-5 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:4  sdat 66:208 active ready running
| `- 9:0:1:4  sdbn 68:16  active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:4  sdf  8:80   active ready running
  `- 9:0:0:4  sdz  65:144 active ready running
3600a098038303558735d493762504b2d dm-10 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:8  sdax 67:16  active ready running
| `- 9:0:1:8  sdbr 68:80  active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:8  sdj  8:144  active ready running
  `- 9:0:0:8  sdad 65:208 active ready running
...
3600a098038303558735d493762504b37 dm-8 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:5  sdau 66:224 active ready running
| `- 9:0:1:5  sdbo 68:32  active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:5  sdg  8:96   active ready running
  `- 9:0:0:5  sdaa 65:160 active ready running
3600a098038303558735d493762504b4b dm-22 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:19 sdbi 67:192 active ready running
| `- 9:0:1:19 sdcc 69:0   active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:19 sdu  65:64  active ready running
  `- 9:0:0:19 sdao 66:128 active ready running
```

Réactiver le groupe de volumes LVM

Si les LUN LVM ont été correctement découvertes, le système `vgchange --activate y` la commande doit réussir. C'est un bon exemple de la valeur d'un gestionnaire de volumes logiques. Une modification du WWN d'une LUN ou même d'un numéro de série n'est pas importante, car les métadonnées du groupe de volumes sont écrites sur la LUN elle-même.

Le système d'exploitation a analysé les LUN et découvert une petite quantité de données écrites sur la LUN qui l'identifie comme un volume physique appartenant au système `sanvg volumegroup`. Il a ensuite construit tous les périphériques requis. Il suffit de réactiver le groupe de volumes.

```
[root@host1 /]# vgchange --activate y sanvg
  Found duplicate PV fpCzdLTuKfy2xDZjai1NliJh3TjLUBiT: using
/dev/mapper/3600a098038303558735d493762504b46 not /dev/sdp
  Using duplicate PV /dev/mapper/3600a098038303558735d493762504b46 from
subsystem DM, ignoring /dev/sdp
  2 logical volume(s) in volume group "sanvg" now active
```

Remonter les systèmes de fichiers

Une fois le groupe de volumes réactivé, les systèmes de fichiers peuvent être montés avec toutes les données d'origine intactes. Comme nous l'avons vu précédemment, les systèmes de fichiers sont pleinement opérationnels, même si la réplication des données est toujours active dans le groupe en arrière-plan.

```
[root@host1 /]# mount /orabin
[root@host1 /]# mount /backups
[root@host1 /]# df -k
Filesystem                          1K-blocks      Used Available Use%
Mounted on
/dev/mapper/rhel-root                52403200   8837100  43566100  17% /
devtmpfs                             65882776         0  65882776   0% /dev
tmpfs                                 6291456        84   6291372   1%
/dev/shm
tmpfs                                65898668      9884  65888784   1% /run
tmpfs                                65898668         0  65898668   0%
/sys/fs/cgroup
/dev/sda1                              505580    224828    280752  45% /boot
fas8060-nfs-public:/install          199229440 119368256  79861184  60%
/install
fas8040-nfs-routable:/snapomatic      9961472     30528   9930944   1%
/snapomatic
tmpfs                                13179736        16  13179720   1%
/run/user/42
tmpfs                                13179736         0  13179736   0%
/run/user/0
/dev/mapper/sanvg-lvorabin           20961280  12357456   8603824  59%
/orabin
/dev/mapper/sanvg-lvbackups          73364480  62947536  10416944  86%
/backups
```

Rechercher à nouveau les périphériques ASM

Les périphériques ASMlib auraient dû être redécouverts lorsque les périphériques SCSI ont été renumérisés. La redécouverte peut être vérifiée en ligne en redémarrant ASMlib puis en analysant les disques.

Cette étape concerne uniquement les configurations ASM où ASMlib est utilisé.

Attention : lorsque ASMlib n'est pas utilisé, le `/dev/mapper` les périphériques doivent avoir été recréés automatiquement. Cependant, les autorisations peuvent ne pas être correctes. Vous devez définir des autorisations spéciales sur les périphériques sous-jacents pour ASM en l'absence d'ASMlib. Cette opération est généralement réalisée par des entrées spéciales dans l'un ou l'autre des `/etc/multipath.conf` ou `udev` ou éventuellement dans les deux jeux de règles. Ces fichiers peuvent avoir besoin d'être mis à jour pour refléter les modifications de l'environnement en termes de WWN ou de numéros de série afin de s'assurer que les périphériques ASM disposent toujours des autorisations appropriées.

Dans cet exemple, le redémarrage d'ASMlib et l'analyse des disques affichent les 10 mêmes LUN ASM que l'environnement d'origine.

```
[root@host1 /]# oracleasm exit
Unmounting ASMlib driver filesystem: /dev/oracleasm
Unloading module "oracleasm": oracleasm
[root@host1 /]# oracleasm init
Loading module "oracleasm": oracleasm
Configuring "oracleasm" to use device physical block size
Mounting ASMlib driver filesystem: /dev/oracleasm
[root@host1 /]# oracleasm scandisks
Reloading disk partitions: done
Cleaning any stale ASM disks...
Scanning system for ASM disks...
Instantiating disk "ASM0"
Instantiating disk "ASM1"
Instantiating disk "ASM2"
Instantiating disk "ASM3"
Instantiating disk "ASM4"
Instantiating disk "ASM5"
Instantiating disk "ASM6"
Instantiating disk "ASM7"
Instantiating disk "ASM8"
Instantiating disk "ASM9"
```

Redémarrez les services de grille

Maintenant que les périphériques LVM et ASM sont en ligne et disponibles, les services de grille peuvent être redémarrés.

```
[root@host1 /]# cd /orabin/product/12.1.0/grid/bin
[root@host1 bin]# ./crsctl start has
```

Redémarrez la base de données

Une fois les services de grille redémarrés, la base de données peut être ouverte. Il peut être nécessaire d'attendre quelques minutes que les services ASM soient entièrement disponibles avant d'essayer de démarrer la base de données.

```
[root@host1 bin]# su - oracle
[oracle@host1 ~]$ . oraenv
ORACLE_SID = [oracle] ? FLIDB
The Oracle base has been set to /orabin
[oracle@host1 ~]$ sqlplus / as sysdba
SQL*Plus: Release 12.1.0.2.0
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> startup
ORACLE instance started.
Total System Global Area 3221225472 bytes
Fixed Size                  4502416 bytes
Variable Size            1207962736 bytes
Database Buffers         1996488704 bytes
Redo Buffers               12271616 bytes
Database mounted.
Database opened.
SQL>
```

Achèvement

Du point de vue de l’hôte, la migration est terminée, mais les E/S sont toujours servies depuis la baie étrangère jusqu’à ce que les relations d’importation soient supprimées.

Avant de supprimer les relations, vous devez confirmer que le processus de migration est terminé pour toutes les LUN.

```
Cluster01::*> lun import show -vserver vserver1 -fields foreign-
disk,path,operational-state
vserver   foreign-disk path              operational-state
--------- ------------ ----------------- -----------------
vserver1 800DT$HuVWB/ /vol/new_asm/LUN4 completed
vserver1 800DT$HuVWBW /vol/new_asm/LUN0 completed
vserver1 800DT$HuVWBX /vol/new_asm/LUN1 completed
vserver1 800DT$HuVWBY /vol/new_asm/LUN2 completed
vserver1 800DT$HuVWBZ /vol/new_asm/LUN3 completed
vserver1 800DT$HuVWBa /vol/new_asm/LUN5 completed
vserver1 800DT$HuVWBb /vol/new_asm/LUN6 completed
vserver1 800DT$HuVWBc /vol/new_asm/LUN7 completed
vserver1 800DT$HuVWBd /vol/new_asm/LUN8 completed
vserver1 800DT$HuVWBe /vol/new_asm/LUN9 completed
vserver1 800DT$HuVWBf /vol/new_lvm/LUN0 completed
vserver1 800DT$HuVWBg /vol/new_lvm/LUN1 completed
vserver1 800DT$HuVWBh /vol/new_lvm/LUN2 completed
vserver1 800DT$HuVWBi /vol/new_lvm/LUN3 completed
vserver1 800DT$HuVWBj /vol/new_lvm/LUN4 completed
vserver1 800DT$HuVWBk /vol/new_lvm/LUN5 completed
vserver1 800DT$HuVWBl /vol/new_lvm/LUN6 completed
vserver1 800DT$HuVWBm /vol/new_lvm/LUN7 completed
vserver1 800DT$HuVWBn /vol/new_lvm/LUN8 completed
vserver1 800DT$HuVWBo /vol/new_lvm/LUN9 completed
20 entries were displayed.
```

Supprimer les relations d'importation

Une fois le processus de migration terminé, supprimez la relation de migration. Une fois que vous avez terminé, les E/S sont servies exclusivement à partir des disques sur ONTAP.

```
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_asm/LUN0
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_asm/LUN1
...
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_lvm/LUN8
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_lvm/LUN9
```

Désenregistrer des LUN étrangères

Enfin, modifiez le disque pour retirer le `is-foreign` désignation.

```
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBW} -is
-foreign false
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBX} -is
-foreign false
...
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBn} -is
-foreign false
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBo} -is
-foreign false
Cluster01::*>
```

Conversion de protocoles

La modification du protocole utilisé pour accéder à une LUN est une exigence courante.

Dans certains cas, cela fait partie d'une stratégie globale de migration des données vers le cloud. Le protocole TCP/IP est le protocole du cloud. En passant de FC à iSCSI, vous simplifiez la migration vers divers environnements cloud. Dans d'autres cas, il peut être souhaitable de tirer parti de la réduction des coûts d'un SAN IP. Il arrive qu'une migration utilise un protocole différent comme mesure temporaire. Par exemple, si une baie étrangère et des LUN ONTAP ne peuvent pas coexister sur les mêmes HBA, vous pouvez utiliser des LUN iSCSI suffisamment longues pour copier les données de l'ancienne baie. Vous pouvez ensuite reconvertir en FC après le retrait des anciennes LUN du système.

La procédure suivante illustre la conversion de FC en iSCSI, mais les principes généraux s'appliquent à une conversion iSCSI inverse en FC.

Installez l'initiateur iSCSI

La plupart des systèmes d'exploitation incluent par défaut un initiateur iSCSI logiciel, mais si celui-ci n'est pas inclus, il peut être facilement installé.

```
[root@host1 /]# yum install -y iscsi-initiator-utils
Loaded plugins: langpacks, product-id, search-disabled-repos,
subscription-
              : manager
Resolving Dependencies
--> Running transaction check
---> Package iscsi-initiator-utils.x86_64 0:6.2.0.873-32.el7 will be
updated
--> Processing Dependency: iscsi-initiator-utils = 6.2.0.873-32.el7 for
package: iscsi-initiator-utils-iscsiuio-6.2.0.873-32.el7.x86_64
---> Package iscsi-initiator-utils.x86_64 0:6.2.0.873-32.0.2.el7 will be
an update
--> Running transaction check
---> Package iscsi-initiator-utils-iscsiuio.x86_64 0:6.2.0.873-32.el7 will
be updated
---> Package iscsi-initiator-utils-iscsiuio.x86_64 0:6.2.0.873-32.0.2.el7
```

```
will be an update
--> Finished Dependency Resolution
Dependencies Resolved
================================================================================
===
 Package                         Arch   Version              Repository
Size
================================================================================
===
Updating:
 iscsi-initiator-utils           x86_64 6.2.0.873-32.0.2.el7 ol7_latest 416
k
Updating for dependencies:
 iscsi-initiator-utils-iscsiuio x86_64 6.2.0.873-32.0.2.el7 ol7_latest  84
k
Transaction Summary
================================================================================
===
Upgrade  1 Package (+1 Dependent package)
Total download size: 501 k
Downloading packages:
No Presto metadata available for ol7_latest
(1/2): iscsi-initiator-utils-6.2.0.873-32.0.2.el7.x86_6 | 416 kB   00:00
(2/2): iscsi-initiator-utils-iscsiuio-6.2.0.873-32.0.2. |  84 kB   00:00
--------------------------------------------------------------------------------
---
Total                                           2.8 MB/s | 501 kB
00:00Cluster01
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  Updating   : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.0.2.el7.x86
1/4
  Updating   : iscsi-initiator-utils-6.2.0.873-32.0.2.el7.x86_64
2/4
  Cleanup    : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.el7.x86_64
3/4
  Cleanup    : iscsi-initiator-utils-6.2.0.873-32.el7.x86_64
4/4
rhel-7-server-eus-rpms/7Server/x86_64/productid       | 1.7 kB     00:00
rhel-7-server-rpms/7Server/x86_64/productid           | 1.7 kB     00:00
  Verifying  : iscsi-initiator-utils-6.2.0.873-32.0.2.el7.x86_64
1/4
  Verifying  : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.0.2.el7.x86
2/4
```

```
  Verifying  : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.el7.x86_64
3/4
  Verifying  : iscsi-initiator-utils-6.2.0.873-32.el7.x86_64
4/4
Updated:
  iscsi-initiator-utils.x86_64 0:6.2.0.873-32.0.2.el7
Dependency Updated:
  iscsi-initiator-utils-iscsiuio.x86_64 0:6.2.0.873-32.0.2.el7
Complete!
[root@host1 /]#
```

Identifiez le nom de l'initiateur iSCSI

Un nom d'initiateur iSCSI unique est généré lors du processus d'installation. Sous Linux, il se trouve dans le `/etc/iscsi/initiatorname.iscsi` fichier. Ce nom permet d'identifier l'hôte sur le SAN IP.

```
[root@host1 /]# cat /etc/iscsi/initiatorname.iscsi
InitiatorName=iqn.1992-05.com.redhat:497bd66ca0
```

Créer un nouveau groupe initiateur

Un groupe initiateur (igroup) fait partie de l'architecture de masquage des LUN ONTAP. L'accès à une LUN nouvellement créée n'est pas accessible à moins qu'un hôte ne bénéficie au préalable d'un accès. Cette étape est effectuée en créant un groupe initiateur qui répertorie les WWN FC ou les noms d'initiateurs iSCSI nécessitant un accès.

Dans cet exemple, un groupe initiateur contenant l'initiateur iSCSI de l'hôte Linux est créé.

```
Cluster01::*> igroup create -igroup linuxiscsi -protocol iscsi -ostype
linux -initiator iqn.1994-05.com.redhat:497bd66ca0
```

Arrêtez l'environnement

Avant de modifier le protocole LUN, les LUN doivent être complètement suspendues. Toute base de données de l'une des LUN en cours de conversion doit être arrêtée, les systèmes de fichiers doivent être démontés et les groupes de volumes doivent être désactivés. Si ASM est utilisé, assurez-vous que le groupe de disques ASM est démonté et arrêtez tous les services de grille.

Annulez le mappage des LUN à partir du réseau FC

Une fois les LUN entièrement suspendues, supprimez les mappages du groupe initiateur FC d'origine.

```
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_asm/LUN0 -igroup
linuxhost
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_asm/LUN1 -igroup
linuxhost
...
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_lvm/LUN8 -igroup
linuxhost
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_lvm/LUN9 -igroup
linuxhost
```

Remappez les LUN sur le réseau IP

Accordez l'accès à chaque LUN au nouveau groupe initiateur iSCSI.

```
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN0 -igroup
linuxiscsi
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN1 -igroup
linuxiscsi
...
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN8 -igroup
linuxiscsi
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN9 -igroup
linuxiscsi
Cluster01::*>
```

Découvrez les cibles iSCSI

La découverte iSCSI se déroule en deux phases. Le premier consiste à découvrir les cibles, qui n'équivaut pas à détecter une LUN. Le `iscsiadm` la commande illustrée ci-dessous sonde le groupe de portails spécifié par le `-p argument` Et stocke une liste de toutes les adresses IP et de tous les ports qui offrent des services iSCSI. Dans ce cas, quatre adresses IP disposent de services iSCSI sur le port par défaut 3260.

> Cette commande peut prendre plusieurs minutes si l'une des adresses IP cibles ne peut pas être atteinte.

```
[root@host1 ~]# iscsiadm -m discovery -t st -p fas8060-iscsi-public1
10.63.147.197:3260,1033 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
10.63.147.198:3260,1034 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
172.20.108.203:3260,1030 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
172.20.108.202:3260,1029 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
```

Découverte des LUN iSCSI

Une fois les cibles iSCSI détectées, redémarrez le service iSCSI pour découvrir les LUN iSCSI disponibles et construire les périphériques associés tels que les périphériques multivoies ou ASMlib.

```
[root@host1 ~]# service iscsi restart
Redirecting to /bin/systemctl restart  iscsi.service
```

Redémarrez l'environnement

Redémarrez l'environnement en réactivant les groupes de volumes, en remontant les systèmes de fichiers, en redémarrant les services RAC, etc. Par mesure de précaution, NetApp vous recommande de redémarrer le serveur une fois le processus de conversion terminé afin de vous assurer que tous les fichiers de configuration sont corrects et que tous les périphériques obsolètes sont supprimés.

Attention : avant de redémarrer un hôte, assurez-vous que toutes les entrées dans `/etc/fstab` Les ressources SAN migrées de cette référence sont commentées. Si cette étape n'est pas effectuée et qu'il y a des problèmes avec l'accès aux LUN, le système d'exploitation ne s'amorce pas. Ce problème n'endommage pas les données. Cependant, il peut être très peu commode de démarrer en mode de secours ou un mode similaire et correct `/etc/fstab` Afin que le système d'exploitation puisse être démarré pour permettre aux efforts de dépannage de commencer.

Exemples de scripts

Les scripts présentés sont fournis sous forme d'exemples de script de diverses tâches du système d'exploitation et de la base de données. Ils sont fournis en l'état. Si une assistance est requise pour une procédure particulière, contactez NetApp ou un revendeur NetApp.

Arrêt de la base de données

Le script Perl suivant prend un seul argument du SID Oracle et arrête une base de données. Il peut être exécuté en tant qu'utilisateur Oracle ou en tant que root.

```
#! /usr/bin/perl
use strict;
use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
my $uid=$<;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
77 Migration of Oracle Databases to NetApp Storage Systems © 2021 NetApp,
Inc. All rights reserved
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
shutdown immediate;
EOF2
'
`;}
else {
@out=`. oraenv << EOF1
$oraclesid
EOF4
sqlplus / as sysdba << EOF2
shutdown immediate;
EOF2
`;};
print @out;
if ("@out" =~ /ORACLE instance shut down/) {
print "$oraclesid shut down\n";
exit 0;}
elsif ("@out" =~ /Connected to an idle instance/) {
print "$oraclesid already shut down\n";
exit 0;}
else {
print "$oraclesid failed to shut down\n";
exit 1;}
```

Démarrage de la base de données

Le script Perl suivant prend un seul argument du SID Oracle et arrête une base de données. Il peut être exécuté en tant qu'utilisateur Oracle ou en tant que root.

```
#! /usr/bin/perl
use strict;
use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
my $uid=$<;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
startup;
EOF2
'
`;}
else {
@out=`. oraenv << EOF3
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
startup;
EOF2
`;};
print @out;
if ("@out" =~ /Database opened/) {
print "$oraclesid started\n";
exit 0;}
elsif ("@out" =~ /cannot start already-running ORACLE/) {
print "$oraclesid already started\n";
exit 1;}
else {
78 Migration of Oracle Databases to NetApp Storage Systems © 2021 NetApp,
Inc. All rights reserved
print "$oraclesid failed to start\n";
exit 1;}
```

Convertir le système de fichiers en lecture seule

Le script suivant prend un argument de système de fichiers et tente de le démonter et de le remonter en lecture seule. Cette opération est utile lors des processus de migration au cours desquels un système de fichiers doit être mis à disposition pour répliquer des données, tout en étant protégé contre les dommages accidentels.

```
#! /usr/bin/perl
use strict;
#use warnings;
my $filesystem=$ARGV[0];
my @out=`umount '$filesystem'`;
if ($? == 0) {
 print "$filesystem unmounted\n";
 @out = `mount -o ro '$filesystem'`;
 if ($? == 0) {
 print "$filesystem mounted read-only\n";
 exit 0;}}
else {
 print "Unable to unmount $filesystem\n";
 exit 1;}
print @out;
```

Remplacer le système de fichiers

L'exemple de script suivant est utilisé pour remplacer un système de fichiers par un autre. Comme il modifie le fichier `/etc/fstab `file, il doit être exécuté en tant que root. Il accepte un seul argument délimité par des virgules pour les anciens et les nouveaux systèmes de fichiers.

1. Pour remplacer le système de fichiers, exécutez le script suivant :

```
#! /usr/bin/perl
use strict;
#use warnings;
my $oldfs;
my $newfs;
my @oldfstab;
my @newfstab;
my $source;
my $mountpoint;
my $leftover;
my $oldfstabentry='';
my $newfstabentry='';
my $migratedfstabentry='';
($oldfs, $newfs) = split (',',$ARGV[0]);
open(my $filehandle, '<', '/etc/fstab') or die "Could not open
/etc/fstab\n";
while (my $line = <$filehandle>) {
 chomp $line;
 ($source, $mountpoint, $leftover) = split(/[ , ]/,$line, 3);
 if ($mountpoint eq $oldfs) {
 $oldfstabentry = "#Removed by swap script $source $oldfs $leftover";}
```

```
 elsif ($mountpoint eq $newfs) {
 $newfstabentry = "#Removed by swap script $source $newfs $leftover";
 $migratedfstabentry = "$source $oldfs $leftover";}
 else {
 push (@newfstab, "$line\n")}}
79 Migration of Oracle Databases to NetApp Storage Systems © 2021
NetApp, Inc. All rights reserved
push (@newfstab, "$oldfstabentry\n");
push (@newfstab, "$newfstabentry\n");
push (@newfstab, "$migratedfstabentry\n");
close($filehandle);
if ($oldfstabentry eq ''){
 die "Could not find $oldfs in /etc/fstab\n";}
if ($newfstabentry eq ''){
 die "Could not find $newfs in /etc/fstab\n";}
my @out=`umount '$newfs'`;
if ($? == 0) {
 print "$newfs unmounted\n";}
else {
 print "Unable to unmount $newfs\n";
 exit 1;}
@out=`umount '$oldfs'`;
if ($? == 0) {
 print "$oldfs unmounted\n";}
else {
 print "Unable to unmount $oldfs\n";
 exit 1;}
system("cp /etc/fstab /etc/fstab.bak");
open ($filehandle, ">", '/etc/fstab') or die "Could not open /etc/fstab
for writing\n";
for my $line (@newfstab) {
 print $filehandle $line;}
close($filehandle);
@out=`mount '$oldfs'`;
if ($? == 0) {
 print "Mounted updated $oldfs\n";
 exit 0;}
else{
 print "Unable to mount updated $oldfs\n";
 exit 1;}
exit 0;
```

Comme exemple d'utilisation de ce script, supposons que les données dans `/oradata` est migré vers `/neworadata` et `/logs` est migré vers `/newlogs`. L'une des méthodes les plus simples pour effectuer cette tâche consiste à utiliser une simple opération de copie de fichier pour replacer le nouveau périphérique sur le point de montage d'origine.

2. Supposons que l'ancien et le nouveau système de fichiers sont présents dans le `/etc/fstab` classer comme suit :

```
cluster01:/vol_oradata /oradata nfs rw,bg,vers=3,rsize=65536,wsize=65536
0 0
cluster01:/vol_logs /logs nfs rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_neworadata /neworadata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_newlogs /newlogs nfs rw,bg,vers=3,rsize=65536,wsize=65536
0 0
```

3. Lors de son exécution, ce script démonte le système de fichiers actuel et le remplace par le nouveau :

```
[root@jfsc3 scripts]# ./swap.fs.pl /oradata,/neworadata
/neworadata unmounted
/oradata unmounted
Mounted updated /oradata
[root@jfsc3 scripts]# ./swap.fs.pl /logs,/newlogs
/newlogs unmounted
/logs unmounted
Mounted updated /logs
```

4. Le script met également à jour le `/etc/fstab` classez-les en conséquence. Dans l'exemple illustré ici, il inclut les modifications suivantes :

```
#Removed by swap script cluster01:/vol_oradata /oradata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
#Removed by swap script cluster01:/vol_neworadata /neworadata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_neworadata /oradata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
#Removed by swap script cluster01:/vol_logs /logs nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
#Removed by swap script cluster01:/vol_newlogs /newlogs nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_newlogs /logs nfs rw,bg,vers=3,rsize=65536,wsize=65536 0
0
```

Migration automatisée des bases de données

Cet exemple illustre l'utilisation de scripts d'arrêt, de démarrage et de remplacement de système de fichiers pour automatiser complètement la migration.

```
#! /usr/bin/perl
use strict;
#use warnings;
my $oraclesid=$ARGV[0];
my @oldfs;
my @newfs;
my $x=1;
while ($x < scalar(@ARGV)) {
   ($oldfs[$x-1], $newfs[$x-1]) = split (',',$ARGV[$x]);
   $x+=1;}
my @out=`./dbshut.pl '$oraclesid'`;
print @out;
if ($? ne 0) {
   print "Failed to shut down database\n";
   exit 0;}
$x=0;
while ($x < scalar(@oldfs)) {
   my @out=`./mk.fs.readonly.pl '$oldfs[$x]'`;
   if ($? ne 0) {
      print "Failed to make filesystem $oldfs[$x] readonly\n";
      exit 0;}
   $x+=1;}
$x=0;
while ($x < scalar(@oldfs)) {
   my @out=`rsync -rlpogt --stats --progress --exclude='.snapshot'
'$oldfs[$x]/' '/$newfs[$x]/'`;
   print @out;
   if ($? ne 0) {
      print "Failed to copy filesystem $oldfs[$x] to $newfs[$x]\n";
      exit 0;}
   else {
      print "Succesfully replicated filesystem $oldfs[$x] to
$newfs[$x]\n";}
   $x+=1;}
$x=0;
while ($x < scalar(@oldfs)) {
   print "swap $x $oldfs[$x] $newfs[$x]\n";
   my @out=`./swap.fs.pl '$oldfs[$x],$newfs[$x]'`;
   print @out;
   if ($? ne 0) {
      print "Failed to swap filesystem $oldfs[$x] for $newfs[$x]\n";
      exit 1;}
   else {
      print "Swapped filesystem $oldfs[$x] for $newfs[$x]\n";}
   $x+=1;}
my @out=`./dbstart.pl '$oraclesid'`;
```

```
print @out;
```

Afficher les emplacements des fichiers

Ce script collecte un certain nombre de paramètres de base de données critiques et les imprime dans un format facile à lire. Ce script peut être utile lors de la révision des dispositions de données. En outre, le script peut être modifié pour d'autres utilisations.

```
#! /usr/bin/perl
#use strict;
#use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
sub dosql{
        my $command = @_[0];
        my @lines;
        my $uid=$<;
        if ($uid == 0) {
        @lines=`su - $oracleuser -c "export ORAENV_ASK=NO;export
ORACLE_SID=$oraclesid;. oraenv -s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
"
        `;}
        else {
        $command=~s/\\\\\\/\\/g;
        @lines=`export ORAENV_ASK=NO;export ORACLE_SID=$oraclesid;. oraenv
-s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
        `;};
return @lines}
print "\n";
@out=dosql('select name from v\\\\\$datafile;');
print "$oraclesid datafiles:\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
```

```
@out=dosql('select member from v\\\\\$logfile;');
print "$oraclesid redo logs:\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
@out=dosql('select name from v\\\\\$tempfile;');
print "$oraclesid temp datafiles:\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
@out=dosql('show parameter spfile;');
print "$oraclesid spfile\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
@out=dosql('select name||\' \'||value from v\\\\\$parameter where
isdefault=\'FALSE\';');
print "$oraclesid key parameters\n";
for $line (@out) {
        chomp($line);
        if ($line =~ /control_files/) {print "$line\n";}
        if ($line =~ /db_create/) {print "$line\n";}
        if ($line =~ /db_file_name_convert/) {print "$line\n";}
        if ($line =~ /log_archive_dest/) {print "$line\n";}}
        if ($line =~ /log_file_name_convert/) {print "$line\n";}
        if ($line =~ /pdb_file_name_convert/) {print "$line\n";}
        if ($line =~ /spfile/) {print "$line\n";}
print "\n";
```

Nettoyage de la migration ASM

```
#! /usr/bin/perl
#use strict;
#use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
sub dosql{
        my $command = @_[0];
        my @lines;
        my $uid=$<;
        if ($uid == 0) {
```

```
        @lines=`su - $oracleuser -c "export ORAENV_ASK=NO;export
ORACLE_SID=$oraclesid;. oraenv -s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
"
        `;}
        else {
        $command=~s/\\\\\\\/\\/g;
        @lines=`export ORAENV_ASK=NO;export ORACLE_SID=$oraclesid;. oraenv
-s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
        `;}
return @lines}
print "\n";
@out=dosql('select name from v\\\\\$datafile;');
print @out;
print "shutdown immediate;\n";
print "startup mount;\n";
print "\n";
for $line (@out) {
        if (length($line) > 1) {
                chomp($line);
                ($first, $second,$third,$fourth)=split('_',$line);
                $fourth =~ s/^TS-//;
                $newname=lc("$fourth.dbf");
                $path2file=$line;
                $path2file=~ /(^.*.\/)/;
                print "host mv $line $1$newname\n";}}
print "\n";
for $line (@out) {
        if (length($line) > 1) {
                chomp($line);
                ($first, $second,$third,$fourth)=split('_',$line);
                $fourth =~ s/^TS-//;
                $newname=lc("$fourth.dbf");
                $path2file=$line;
                $path2file=~ /(^.*.\/)/;
                print "alter database rename file '$line' to
'$1$newname';\n";}}
```

```
print "alter database open;\n";
print "\n";
```

Conversion du nom ASM en nom de système de fichiers

```
set serveroutput on;
set wrap off;
declare
    cursor df is select file#, name from v$datafile;
    cursor tf is select file#, name from v$tempfile;
    cursor lf is select member from v$logfile;
    firstline boolean := true;
begin
    dbms_output.put_line(CHR(13));
    dbms_output.put_line('Parameters for log file conversion:');
    dbms_output.put_line(CHR(13));
    dbms_output.put('*.log_file_name_convert = ');
    for lfrec in lf loop
        if (firstline = true) then
            dbms_output.put('''' || lfrec.member || ''', ');
            dbms_output.put('''/NEW_PATH/' ||
regexp_replace(lfrec.member,'^.*/','') || '''');
        else
            dbms_output.put(',''' || lfrec.member || ''', ');
            dbms_output.put('''/NEW_PATH/' ||
regexp_replace(lfrec.member,'^.*/','') || '''');
        end if;
        firstline:=false;
    end loop;
    dbms_output.put_line(CHR(13));
    dbms_output.put_line(CHR(13));
    dbms_output.put_line('rman duplication script:');
    dbms_output.put_line(CHR(13));
    dbms_output.put_line('run');
    dbms_output.put_line('{');
    for dfrec in df loop
        dbms_output.put_line('set newname for datafile ' ||
            dfrec.file# || ' to ''' || dfrec.name ||''';');
    end loop;
    for tfrec in tf loop
        dbms_output.put_line('set newname for tempfile ' ||
            tfrec.file# || ' to ''' || tfrec.name ||''';');
    end loop;
    dbms_output.put_line('duplicate target database for standby backup
location INSERT_PATH_HERE;');
    dbms_output.put_line('}');
end;
/
```

Relire les journaux sur la base de données

Ce script accepte un seul argument d'un SID Oracle pour une base de données en mode montage et tente de relire tous les journaux d'archives actuellement disponibles.

```
#! /usr/bin/perl
use strict;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
84 Migration of Oracle Databases to NetApp Storage Systems © 2021 NetApp,
Inc. All rights reserved
my $uid = $<;
my @out;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover database until cancel;
auto
EOF2
'
`;}
else {
@out=`. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover database until cancel;
auto
EOF2
`;
}
print @out;
```

Relire les journaux sur la base de données de secours

Ce script est identique au script précédent, sauf qu'il est conçu pour une base de données de secours.

```
#! /usr/bin/perl
use strict;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my $uid = $<;
my @out;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover standby database until cancel;
auto
EOF2
'
`;}
else {
@out=`. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover standby database until cancel;
auto
EOF2
`;
}
print @out;
```

Remarques supplémentaires

Optimisation des performances et analyse comparative

Il est extrêmement compliqué de tester précisément les performances du stockage des bases de données. Il faut comprendre les problèmes suivants :

- IOPS et débit
- Différence entre les opérations d'E/S au premier plan et en arrière-plan
- Effet de la latence sur la base de données
- Nombreux paramètres du système d'exploitation et du réseau qui affectent également les performances du stockage

En outre, il faut tenir compte des tâches qui ne relèvent pas du domaine du stockage dans les bases de données. L'optimisation de la performance du stockage ne présente plus d'avantages, car la performance du stockage n'est plus un facteur limitant.

La majorité des clients de base de données choisissent désormais des baies 100 % Flash, ce qui entraîne d'autres considérations. Prenons l'exemple des tests de performances sur un système AFF A900 à deux nœuds :

- Avec un ratio de lecture/écriture de 80/20, deux nœuds A900 peuvent fournir plus de 1 million d'IOPS de base de données aléatoires avant que la latence ne dépasse même le seuil de 150 µs. Au-delà des exigences de performance actuelles de la plupart des bases de données, il est difficile de prévoir l'amélioration attendue. Le stockage serait largement effacé comme un goulot d'étranglement.
- La bande passante réseau est une source de plus en plus courante de limites de performances. Par exemple, les solutions sur disque mécanique constituent souvent des goulots d'étranglement pour les performances des bases de données, car la latence d'E/S est très élevée. Lorsque les limites de latence sont éliminées par un système 100 % Flash, le obstacle est fréquemment basculer vers le réseau. Ceci est particulièrement notable dans les environnements virtualisés et les systèmes lames où la véritable connectivité réseau est difficile à visualiser. Les tests de performances peuvent ainsi être plus complexes si le système de stockage lui-même ne peut pas être pleinement utilisé en raison des limitations de bande passante.
- Il est généralement impossible de comparer les performances d'une baie 100 % Flash à celles d'une baie contenant des disques rotatifs en raison de la latence considérablement améliorée des baies 100 % Flash. Les résultats des tests ne sont généralement pas significatifs.
- Généralement, comparer les pics de performance d'IOPS avec un système 100 % Flash n'est pas utile, car les bases de données ne sont pas limitées par les E/S de stockage Supposons par exemple qu'une baie peut supporter 500 000 IOPS aléatoires, tandis qu'une autre peut supporter 300 000. La différence n'est pas pertinente en situation réelle si une base de données consacre 99 % de son temps au traitement du processeur. Ces charges de travail n'exploitent jamais toutes les capacités de la baie de stockage. À l'inverse, les pics d'activité d'E/S par seconde peuvent s'avérer critiques pour une plateforme de consolidation sur laquelle la baie de stockage doit être chargée au maximum de ses capacités.
- Lors de tout test de stockage, la latence et les IOPS sont systématiquement prises en compte. De nombreuses baies de stockage sur le marché revendiquant des niveaux extrêmes d'IOPS, mais avec la latence, ces IOPS deviennent inutiles à de tels niveaux. La cible type avec des baies 100 % Flash est le millième de seconde, Une meilleure approche lors de ces tests n'est pas de mesurer les IOPS maximales mais de déterminer le nombre d'IOPS qu'une baie de stockage peut supporter avant que la latence moyenne ne soit supérieure à 1 ms.

Référentiel automatique de workloads Oracle et banc d'essai

Pour les comparaisons de performances Oracle, il est référence dans le rapport Oracle Automatic Workload Repository (AWR).

Il existe plusieurs types de rapports AWR. Du point de vue du stockage, un rapport généré par l'exécution de `awrrpt.sql` La commande est la plus complète et la plus utile, car elle cible une instance de base de données spécifique et inclut des histogrammes détaillés qui décomposent les événements d'E/S de stockage en fonction de la latence.

Dans l'idéal, comparer deux baies de performances implique d'exécuter la même charge de travail sur chaque baie et de produire un rapport AWR qui cible précisément la charge de travail. Dans le cas d'une charge de travail très longue, il est possible d'utiliser un seul rapport AWR avec un temps écoulé couvrant le temps de début et de fin, mais il est préférable de séparer les données AWR sous forme de plusieurs rapports. Par exemple, si une tâche par lots s'est exécutée de minuit à 6 h, créez une série de rapports AWR d'une heure de minuit à 1 h, de 1 h à 2 h, etc.

Dans d'autres cas, une requête très courte doit être optimisée. La meilleure option est un rapport AWR basé sur un instantané AWR créé au début de la requête et un deuxième instantané AWR créé à la fin de la requête. Le serveur de base de données doit être silencieux pour réduire au minimum l'activité en arrière-plan

qui pourrait masquer l'activité de la requête en cours d'analyse.

> Lorsque les rapports AWR ne sont pas disponibles, les rapports Oracle statspack constituent une bonne alternative. Ils contiennent la plupart des mêmes statistiques d'E/S qu'un rapport AWR.

Oracle AWR et dépannage

Un rapport AWR est également l'outil le plus important pour analyser un problème de performances.

Comme pour les bancs d'essai, la résolution des problèmes de performances nécessite que vous mesuriez précisément une charge de travail particulière. Dans la mesure du possible, fournissez des données AWR lorsque vous signalez un problème de performance au centre de support NetApp ou lorsque vous travaillez avec une équipe NetApp ou un partenaire responsable de compte concernant une nouvelle solution.

Lorsque vous fournissez des données AWR, tenez compte des exigences suivantes :

- Exécutez le `awrrpt.sql` pour générer le rapport. La sortie peut être texte ou HTML.
- Si Oracle Real application clusters (RAC) est utilisé, générez des rapports AWR pour chaque instance du cluster.
- Cibler l'heure précise à laquelle le problème a existé. La durée maximale acceptable d'un rapport AWR est généralement d'une heure. Si un problème persiste pendant plusieurs heures ou implique une opération sur plusieurs heures, par exemple un traitement par lots, fournissez plusieurs rapports AWR d'une heure qui couvrent l'ensemble de la période à analyser.
- Si possible, réglez l'intervalle d'instantané AWR sur 15 minutes. Ce paramètre permet d'effectuer une analyse plus détaillée. Cela nécessite également des exécutions supplémentaires de `awrrpt.sql` fournir un rapport pour chaque intervalle de 15 minutes.
- Si le problème est une requête en cours très courte, fournissez un rapport AWR basé sur un instantané AWR créé au début de l'opération et un second instantané AWR créé à la fin de l'opération. Le serveur de base de données doit être silencieux pour minimiser l'activité en arrière-plan qui pourrait masquer l'activité de l'opération en cours d'analyse.
- Si un problème de performance est signalé à certains moments mais pas à d'autres, fournissez des données AWR supplémentaires qui démontrent de bonnes performances pour la comparaison.

etalonnez_io

Le `calibrate_io` command ne doit jamais être utilisé pour tester, comparer ou tester les systèmes de stockage. Comme indiqué dans la documentation Oracle, cette procédure permet d'étalonner les capacités d'E/S du stockage.

L'étalonnage n'est pas le même que l'étalonnage. L'objectif de cette commande est d'émettre des E/S pour aider à étalonner les opérations de la base de données et améliorer leur efficacité en optimisant le niveau d'E/S émis pour l'hôte. Car le type d'E/S effectué par le `calibrate_io` Le fonctionnement ne représente pas les E/S réelles de l'utilisateur de la base de données, les résultats ne sont pas prévisibles et ne sont souvent même pas reproductibles.

SLOB2

SLOB2, le très petit banc d'essai Oracle, est devenu l'outil privilégié pour évaluer les performances des bases de données. Il a été développé par Kevin Closson et est disponible à l'adresse "https://kevinclosson.net/slob/". L'installation et la configuration ne prennent que quelques minutes et une base de données Oracle génère des modèles d'E/S sur un espace de table définissable par l'utilisateur. Il s'agit de l'une des rares options de test

disponibles permettant de saturer une baie 100 % Flash par E/S. Il est également utile de générer des niveaux d'E/S beaucoup plus bas pour simuler des charges de travail de stockage qui font partie des IOPS faibles, mais qui sont sensibles à la latence.

Swingbench

Swingbench peut être utile pour tester les performances des bases de données, mais il est extrêmement difficile d'utiliser Swingbench sous un contrainte de stockage. NetApp n'a constaté aucun test de Swingbench ayant produit suffisamment d'E/S pour être une charge significative sur n'importe quelle baie AFF. Dans certains cas limités, le test OET (Order Entry Test) peut être utilisé pour évaluer le stockage du point de vue de la latence. Cela peut s'avérer utile lorsqu'une base de données a une dépendance connue en termes de latence pour des requêtes particulières. Assurez-vous que l'hôte et le réseau sont correctement configurés pour atteindre les potentiels de latence d'une baie 100 % Flash.

HammerDB

HammerDB est un outil de test de base de données qui simule les bancs d'essai TPC-C et TPC-H, entre autres. La construction d'un jeu de données suffisamment volumineux pour exécuter correctement un test peut prendre beaucoup de temps, mais elle peut constituer un outil efficace pour évaluer les performances des applications OLTP et d'entrepôt de données.

Orion

L'outil Oracle Orion a été couramment utilisé avec Oracle 9, mais il n'a pas été maintenu pour assurer la compatibilité avec les modifications apportées aux différents systèmes d'exploitation hôtes. Il est rarement utilisé avec Oracle 10 ou Oracle 11 en raison d'incompatibilités avec le système d'exploitation et la configuration du stockage.

Oracle a réécrit l'outil, qui est installé par défaut dans Oracle 12c. Bien que ce produit ait été amélioré et utilise la plupart des appels qu'une véritable base de données Oracle utilise, il n'utilise pas exactement le même chemin de code ou le même comportement d'E/S que celui utilisé par Oracle. Par exemple, la plupart des E/S Oracle sont exécutées de manière synchrone, ce qui signifie que la base de données s'arrête jusqu'à ce que les E/S soient terminées lorsque l'opération d'E/S se termine au premier plan. Le simple fait d'inonder un système de stockage d'E/S aléatoires n'est pas une reproduction de véritables E/S Oracle et n'offre pas de méthode directe pour comparer les baies de stockage ou mesurer l'impact des modifications de configuration.

Cela étant, Orion est souvent associé à des cas d'usage, comme l'évaluation générale des performances maximales d'une configuration de stockage hôte-réseau ou encore l'évaluation de l'état d'un système de stockage. Grâce à des tests rigoureux, nous pouvons concevoir des tests Orion exploitables afin de comparer les baies de stockage ou d'évaluer l'effet d'une modification de la configuration, dans la mesure où les paramètres tiennent compte des IOPS, du débit et de la latence, et tenter de répliquer fidèlement une charge de travail réaliste.

Les verrous NFSv3 obsolètes

Si un serveur de base de données Oracle tombe en panne, des verrous NFS obsolètes peuvent se présenter au redémarrage. Ce problème peut être évité en portant une attention particulière à la configuration de la résolution de nom sur le serveur.

Ce problème survient parce que la création d'un verrou et l'effacement d'un verrou utilisent deux méthodes légèrement différentes de résolution de nom. Deux processus sont impliqués, Network Lock Manager (NLM) et le client NFS. Le NLM utilise `uname n` pour déterminer le nom d'hôte, pendant que le système `rpc.statd` utilisations des processus `gethostbyname()`. Ces noms d'hôte doivent correspondre pour que le système d'exploitation efface correctement les verrous obsolètes. Par exemple, l'hôte peut rechercher des verrous

appartenant à `dbserver5`, mais les verrous ont été enregistrés par l'hôte comme `dbserver5.mydomain.org`. Si `gethostbyname()` ne renvoie pas la même valeur que `uname -a`, le processus de déverrouillage n'a pas réussi.

L'exemple de script suivant vérifie si la résolution des noms est parfaitement cohérente :

```
#! /usr/bin/perl
$uname=`uname -n`;
chomp($uname);
($name, $aliases, $addrtype, $length, @addrs) = gethostbyname $uname;
print "uname -n yields: $uname\n";
print "gethostbyname yields: $name\n";
```

Si `gethostbyname` ne correspond pas `uname`, des verrous obsolètes sont probables. Par exemple, ce résultat indique un problème potentiel :

```
uname -n yields: dbserver5
gethostbyname yields: dbserver5.mydomain.org
```

La solution est généralement trouvée en modifiant l'ordre dans lequel les hôtes apparaissent dans `/etc/hosts`. Par exemple, supposons que le fichier hosts inclut l'entrée suivante :

```
10.156.110.201  dbserver5.mydomain.org dbserver5 loghost
```

Pour résoudre ce problème, modifiez l'ordre dans lequel le nom de domaine complet et le nom d'hôte court apparaissent :

```
10.156.110.201  dbserver5 dbserver5.mydomain.org loghost
```

`gethostbyname()` renvoie maintenant le court `dbserver5` nom d'hôte, qui correspond à la sortie de `uname`. Les verrous sont donc effacés automatiquement après une panne de serveur.

Vérification de l'alignement WAFL

Un alignement WAFL correct est essentiel pour de bonnes performances. Même si ONTAP gère des blocs dans des unités de 4 Ko, ONTAP ne réalise pas forcément toutes les opérations dans des unités de 4 Ko. ONTAP prend en charge les opérations en mode bloc de différentes tailles, mais la comptabilité sous-jacente est gérée par WAFL en unités de 4 Ko.

Le terme « alignement » fait référence à la manière dont les E/S Oracle correspondent à ces unités de 4 Ko. Pour optimiser les performances, un bloc Oracle de 8 Ko doit résider sur deux blocs physiques WAFL de 4 Ko sur un disque. Si un bloc est décalé de 2 Ko, ce bloc réside dans la moitié d'un bloc de 4 Ko, dans un bloc séparé complet de 4 Ko, puis dans la moitié d'un troisième bloc de 4 Ko. Cette configuration entraîne une dégradation des performances.

L'alignement n'est pas un problème avec les systèmes de fichiers NAS. Les fichiers de données Oracle sont alignés sur le début du fichier en fonction de la taille du bloc Oracle. Par conséquent, les tailles de bloc de 8 Ko, 16 Ko et 32 Ko sont toujours alignées. Toutes les opérations de bloc sont décalées par rapport au début du fichier en unités de 4 kilo-octets.

Les LUN, en revanche, contiennent généralement au départ un type d'en-tête de pilote ou de métadonnées de système de fichiers qui crée un décalage. L'alignement est rarement un problème dans les systèmes d'exploitation modernes, car ces systèmes d'exploitation sont conçus pour des disques physiques pouvant utiliser un secteur natif de 4 Ko. De plus, ils requièrent l'alignement des E/S sur les limites de 4 Ko pour des performances optimales.

Il y a toutefois quelques exceptions. Une base de données a peut-être été migrée à partir d'un système d'exploitation plus ancien qui n'a pas été optimisé pour les E/S de 4 Ko, ou une erreur de l'utilisateur lors de la création de la partition a pu entraîner un décalage qui ne se situe pas dans des unités de 4 Ko.

Les exemples suivants sont spécifiques à Linux, mais la procédure peut être adaptée à n'importe quel système d'exploitation.

Aligné

L'exemple suivant montre une vérification d'alignement sur une seule LUN avec une seule partition.

Tout d'abord, créez la partition qui utilise toutes les partitions disponibles sur le lecteur.

```
[root@host0 iscsi]# fdisk /dev/sdb
Device contains neither a valid DOS partition table, nor Sun, SGI or OSF
disklabel
Building a new DOS disklabel with disk identifier 0xb97f94c1.
Changes will remain in memory only, until you decide to write them.
After that, of course, the previous content won't be recoverable.
The device presents a logical sector size that is smaller than
the physical sector size. Aligning to a physical sector (or optimal
I/O) size boundary is recommended, or performance may be impacted.
Command (m for help): n
Command action
   e   extended
   p   primary partition (1-4)
p
Partition number (1-4): 1
First cylinder (1-10240, default 1):
Using default value 1
Last cylinder, +cylinders or +size{K,M,G} (1-10240, default 10240):
Using default value 10240
Command (m for help): w
The partition table has been altered!
Calling ioctl() to re-read partition table.
Syncing disks.
[root@host0 iscsi]#
```

L'alignement peut être vérifié mathématiquement à l'aide de la commande suivante :

```
[root@host0 iscsi]# fdisk -u -l /dev/sdb
Disk /dev/sdb: 10.7 GB, 10737418240 bytes
64 heads, 32 sectors/track, 10240 cylinders, total 20971520 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 4096 bytes
I/O size (minimum/optimal): 4096 bytes / 65536 bytes
Disk identifier: 0xb97f94c1
   Device Boot      Start         End      Blocks   Id  System
/dev/sdb1              32    20971519    10485744   83  Linux
```

Le résultat indique que les unités sont de 512 octets et que le début de la partition est de 32 unités. Il s'agit d'un total de 32 x 512 = 16,834 octets, soit un ensemble de blocs WAFL de 4 Ko. Cette partition est correctement alignée.

Pour vérifier que l'alignement est correct, procédez comme suit :

1. Identifier l'UUID (identifiant universel unique) de la LUN

```
FAS8040SAP::> lun show -v /vol/jfs_luns/lun0
              Vserver Name: jfs
                  LUN UUID: ed95d953-1560-4f74-9006-85b352f58fcd
                    Mapped: mapped`                `
```

2. Entrez le shell du nœud sur le contrôleur ONTAP.

```
FAS8040SAP::> node run -node FAS8040SAP-02
Type 'exit' or 'Ctrl-D' to return to the CLI
FAS8040SAP-02> set advanced
set not found.  Type '?' for a list of commands
FAS8040SAP-02> priv set advanced
Warning: These advanced commands are potentially dangerous; use
         them only when directed to do so by NetApp
         personnel.
```

3. Démarrer les collections statistiques sur l'UUID cible identifié dans la première étape.

```
FAS8040SAP-02*> stats start lun:ed95d953-1560-4f74-9006-85b352f58fcd
Stats identifier name is 'Ind0xffffff08b9536188'
FAS8040SAP-02*>
```

4. Certaines E/S. Il est important d'utiliser le `iflag` Argument permettant de s'assurer que les E/S sont synchrones et non mises en tampon.

Faites très attention avec cette commande. Inversion du `if` et `of` les arguments détruisent les données.

```
[root@host0 iscsi]# dd if=/dev/sdb1 of=/dev/null iflag=dsync count=1000
bs=4096
1000+0 records in
1000+0 records out
4096000 bytes (4.1 MB) copied, 0.0186706 s, 219 MB/s
```

5. Arrêtez les statistiques et affichez l'histogramme d'alignement. Toutes les E/S doivent se trouver dans le `.0` Bucket, qui indique les E/S alignées sur les limites d'un bloc de 4 Ko.

```
FAS8040SAP-02*> stats stop
StatisticsID: Ind0xffffff08b9536188
lun:ed95d953-1560-4f74-9006-85b352f58fcd:instance_uuid:ed95d953-1560-
4f74-9006-85b352f58fcd
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.0:186%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.1:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.2:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.3:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.4:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.5:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.6:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.7:0%
```

Mauvais alignement

L'exemple suivant illustre un mauvais alignement des E/S :

1. Créez une partition qui ne s'aligne pas sur une limite de 4 Ko. Il ne s'agit pas d'un comportement par défaut sur les systèmes d'exploitation modernes.

```
[root@host0 iscsi]# fdisk -u /dev/sdb
Command (m for help): n
Command action
   e   extended
   p   primary partition (1-4)
p
Partition number (1-4): 1
First sector (32-20971519, default 32): 33
Last sector, +sectors or +size{K,M,G} (33-20971519, default 20971519):
Using default value 20971519
Command (m for help): w
The partition table has been altered!
Calling ioctl() to re-read partition table.
Syncing disks.
```

2. La partition a été créée avec un décalage de 33 secteurs au lieu du décalage de 32 par défaut. Répétez la procédure décrite à la section "Aligné". L'histogramme s'affiche comme suit :

```
FAS8040SAP-02*> stats stop
StatisticsID: Ind0xffffff0468242e78
lun:ed95d953-1560-4f74-9006-85b352f58fcd:instance_uuid:ed95d953-1560-
4f74-9006-85b352f58fcd
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.0:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.1:136%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.2:4%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.3:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.4:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.5:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.6:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.7:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_partial_blocks:31%
```

Le mauvais alignement est clair. Les E/S tombent principalement dans le* *`.1` godet, qui correspond au décalage attendu. Lorsque la partition a été créée, elle a été déplacée de 512 octets plus loin dans le périphérique que la valeur par défaut optimisée, ce qui signifie que l'histogramme est décalé de 512 octets.

De plus, le `read_partial_blocks` Ces statistiques ne sont pas égales à zéro, ce qui signifie que des E/S n'ont pas rempli un bloc de 4 Ko entier.

Fichiers de reprise

Les procédures décrites ici s'appliquent aux fichiers de données. Les journaux de reprise et d'archivage Oracle ont différents modèles d'E/S. Par exemple, la journalisation de reprise est un remplacement circulaire d'un seul fichier. Si la taille de bloc par défaut de 512 octets est utilisée, les statistiques d'écriture se ressemblent à ceci :

```
FAS8040SAP-02*> stats stop
StatisticsID: Ind0xffffff0468242e78
lun:ed95d953-1560-4f74-9006-85b352f58fcd:instance_uuid:ed95d953-1560-4f74-
9006-85b352f58fcd
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.0:12%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.1:8%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.2:4%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.3:10%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.4:13%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.5:6%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.6:8%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.7:10%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_partial_blocks:85%
```

Les E/S sont réparties dans tous les compartiments de l'histogramme, mais cela n'est pas un problème de performances. Toutefois, des taux de journalisation de reprise extrêmement élevés peuvent bénéficier d'une taille de bloc de 4 Ko. Dans ce cas, il est conseillé de vérifier que les LUN de journalisation de reprise sont correctement alignées. Cependant, cette condition n'est pas aussi importante pour de bonnes performances que l'alignement des fichiers de données.