Disaster Recovery mit Oracle

Enterprise applications

NetApp
January 12, 2026

Inhalt

Disaster Recovery mit Oracle

Überblick

Disaster Recovery bezieht sich auf die Wiederherstellung von Datenservices nach einem schwerwiegenden Ereignis, beispielsweise durch einen Brand, der ein Storage-System oder sogar einen kompletten Standort zerstört.

Diese Dokumentation ersetzt zuvor veröffentlichte technische Berichte *TR-4591: Oracle Data Protection* und *TR-4592: Oracle on MetroCluster.*

Disaster Recovery kann durch eine einfache Datenreplizierung mit SnapMirror durchgeführt werden, wobei viele Kunden gespiegelte Replikate natürlich so oft wie stündlich aktualisieren.

Bei den meisten Kunden benötigt DR mehr als nur einen Remote-Kopiervorgang, sondern muss in der Lage sein, diese Daten schnell zu nutzen. NetApp bietet zwei Technologien zur Erfüllung dieser Anforderungen: MetroCluster und SnapMirror Active Sync

MetroCluster bezieht sich auf ONTAP in einer Hardwarekonfiguration mit synchron gespiegelten Storage auf niedriger Ebene und zahlreichen zusätzlichen Funktionen. Integrierte Lösungen wie MetroCluster vereinfachen die heutigen komplizierten, horizontal skalierbaren Datenbanken, Applikationen und Virtualisierungsinfrastrukturen. Sie ersetzt mehrere externe Datensicherungsprodukte und -Strategien durch ein einfaches, zentrales Storage-Array. Sie bietet außerdem integriertes Backup, Recovery, Disaster Recovery und Hochverfügbarkeit (HA) in einem einzigen geclusterten Storage-System.

Die SnapMirror Active Sync (SM-AS) basiert auf SnapMirror Synchronous. Mit MetroCluster ist jeder ONTAP Controller für die Replizierung seiner Laufwerksdaten an einen Remote-Standort verantwortlich. Bei SnapMirror Active Sync haben Sie im Grunde zwei verschiedene ONTAP-Systeme, die unabhängige Kopien Ihrer LUN-Daten führen, aber zusammenarbeiten, um eine einzige Instanz dieser LUN zu präsentieren. Auf Host-Ebene handelt es sich um eine einzelne LUN-Einheit.

Vergleich von SM-AS und MCC

SM-AS und MetroCluster sind in der Gesamtfunktionalität ähnlich, es gibt jedoch wichtige Unterschiede in der Art und Weise, wie die RPO=0-Replikation implementiert und gemanagt wird. Die asynchronen und synchronen SnapMirror können auch im Rahmen eines DR-Plans eingesetzt werden, sind aber nicht als Technologien für die HA-Replizierung konzipiert.

- Eine MetroCluster-Konfiguration ähnelt eher einem integrierten Cluster mit über mehrere Standorte verteilten Nodes. SM-AS verhält sich wie zwei ansonsten unabhängige Cluster, die zusammenarbeiten, um ausgewählte RPO=0 synchron replizierte LUNs bereitzustellen.
- Die Daten in einer MetroCluster-Konfiguration sind zu einem bestimmten Zeitpunkt nur von einem bestimmten Standort aus zugänglich. Eine zweite Kopie der Daten befindet sich am gegenüberliegenden Standort, die Daten sind jedoch passiv. Ohne Failover des Speichersystems ist der Zugriff nicht möglich.
- MetroCluster und SM-As führen die Spiegelung auf verschiedenen Ebenen durch. Die MetroCluster Spiegelung wird auf der RAID-Schicht durchgeführt. Die Low-Level-Daten werden mithilfe von SyncMirror in einem gespiegelten Format gespeichert. Die Verwendung der Spiegelung ist in den LUN-, Volume- und Protokollebenen praktisch unsichtbar.
- Im Gegensatz dazu erfolgt die SM-AS-Spiegelung auf der Protokollebene. Die beiden Cluster sind insgesamt unabhängige Cluster. Sobald die beiden Datenkopien synchron sind, müssen die beiden Cluster

nur noch Schreibvorgänge spiegeln. Wenn ein Schreibvorgang auf einem Cluster stattfindet, wird er in das andere Cluster repliziert. Der Schreibvorgang wird dem Host nur dann bestätigt, wenn der Schreibvorgang auf beiden Seiten abgeschlossen ist. Anders als dieses Verhalten bei der Protokollaufteilung sind die beiden Cluster ansonsten normale ONTAP-Cluster.

- Die Hauptrolle bei MetroCluster ist die umfangreiche Replizierung. Sie können ein gesamtes Array mit RPO=0 und RTO von nahezu null replizieren. Dies vereinfacht den Failover-Prozess, da es nur eine „Sache“ für Failover gibt und lässt sich hinsichtlich Kapazität und IOPS extrem gut skalieren.
- Ein wichtiger Anwendungsfall für SM-AS ist die granulare Replizierung. Manchmal möchten Sie nicht alle Daten als eine Einheit replizieren oder bestimmte Workloads selektiv ausfallsicher durchführen.
- Ein weiterer wichtiger Anwendungsfall für SM-As ist der aktiv/aktiv-Betrieb. Dort sollen vollständig nutzbare Datenkopien auf zwei verschiedenen Clustern verfügbar sein, die sich an zwei verschiedenen Standorten mit identischen Performance-Merkmalen befinden und auf Wunsch nicht über Standorte verteilt werden müssen. Sie können Ihre Applikationen bereits auf beiden Standorten ausführen, wodurch sich die RTO während eines Failover verringert.

MetroCluster

Disaster Recovery mit MetroCluster

MetroCluster ist eine ONTAP-Funktion, die Ihre Oracle Datenbanken mit einer standortübergreifenden synchronen RPO=0-Spiegelung sichern kann. Sie lässt sich auf bis zu Hunderte von Datenbanken auf einem einzigen MetroCluster System skalieren.

Darüber hinaus ist die Bedienung einfach. Die Verwendung von MetroCluster trägt nicht notwendigerweise zur Ergänzung oder Änderung der besten Racks für den Betrieb von Enterprise-Applikationen und -Datenbanken bei.

Die üblichen Best Practices gelten weiterhin, und wenn Ihre Bedürfnisse nur RPO=0 Datensicherung erfordern, wird diese Anforderung mit MetroCluster erfüllt. Die meisten Kunden verwenden MetroCluster jedoch nicht nur für die RPO=0-Datensicherung, sondern auch zur Verbesserung der RTO in Notfallszenarien sowie zur Gewährleistung eines transparenten Failovers im Rahmen der Wartungsarbeiten an den Standorten.

Physische Architektur

Um zu verstehen, wie Oracle Datenbanken in einer MetroCluster-Umgebung arbeiten, ist eine Erläuterung des physischen Designs eines MetroCluster-Systems erforderlich.

Diese Dokumentation ersetzt den zuvor veröffentlichten technischen Bericht *TR-4592: Oracle on MetroCluster.*

MetroCluster ist in 3 verschiedenen Konfigurationen erhältlich

- HA-Paare mit IP-Konnektivität
- HA-Paare mit FC-Konnektivität
- Single Controller mit FC-Konnektivität

Der Begriff „Konnektivität“ bezieht sich auf die Clusterverbindung, die für die standortübergreifende Replizierung verwendet wird. Er bezieht sich nicht auf die Host-Protokolle. Unabhängig von der Art der Verbindung, die für die Kommunikation zwischen den Clustern verwendet wird, werden alle Host-seitigen Protokolle wie gewohnt in einer MetroCluster-Konfiguration unterstützt.

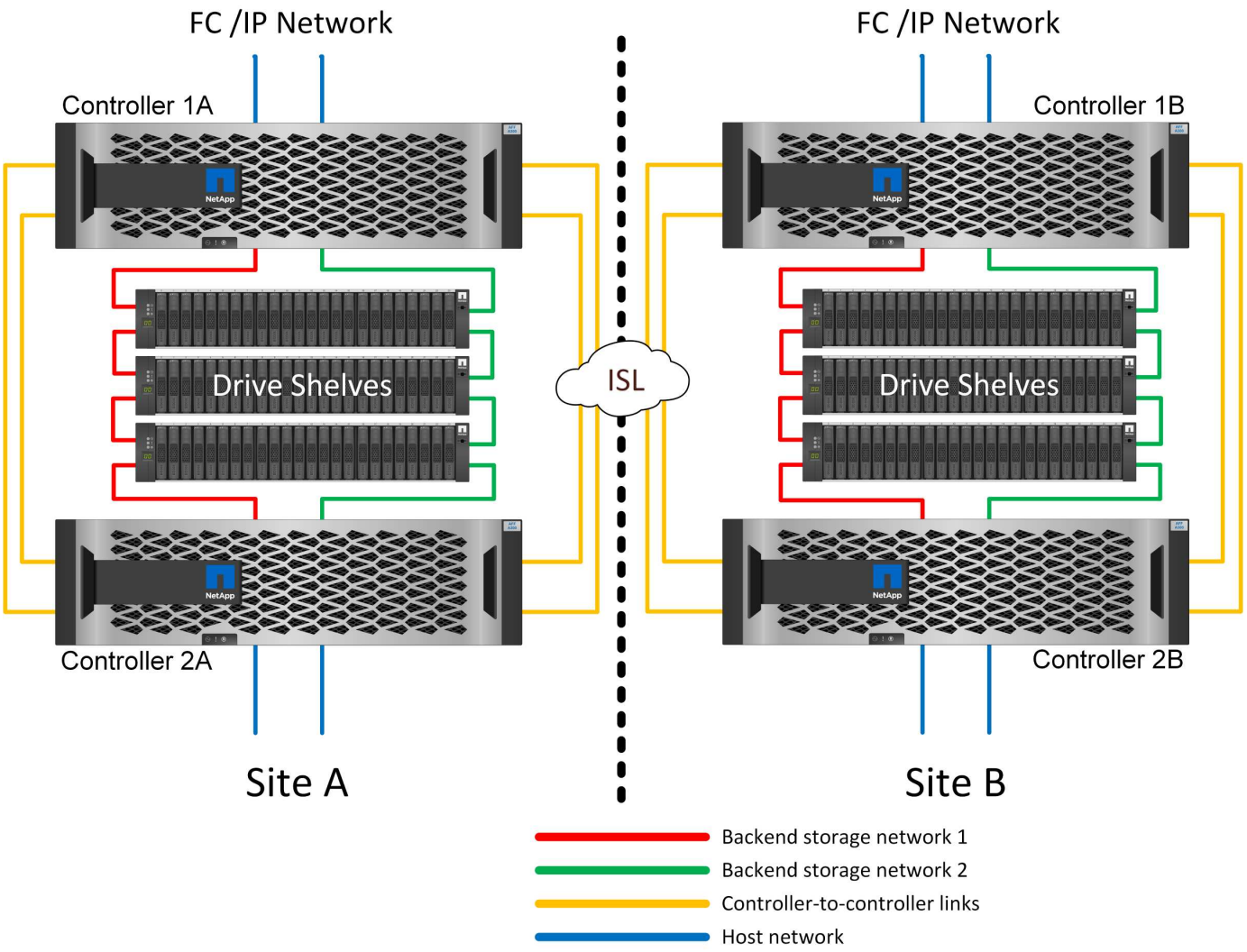
MetroCluster IP

Die HA-Paar-MetroCluster IP-Konfiguration nutzt zwei oder vier Nodes pro Standort. Diese Konfigurationsoption erhöht die Komplexität und die Kosten im Vergleich zur Option mit zwei Nodes, bietet aber einen wichtigen Vorteil: intrasite-Redundanz. Bei einem einfachen Controller-Ausfall ist kein Datenzugriff über das WAN erforderlich. Der Datenzugriff bleibt über den alternativen lokalen Controller lokal.

Die meisten Kunden entscheiden sich für IP-Konnektivität, da die Infrastrukturanforderungen einfacher sind. In der Vergangenheit war die Bereitstellung von ultraschnellen standortübergreifenden Verbindungen über Dark Fibre und FC Switches im Allgemeinen einfacher, heute sind jedoch ultraschnelle IP-Verbindungen mit niedriger Latenz schneller verfügbar.

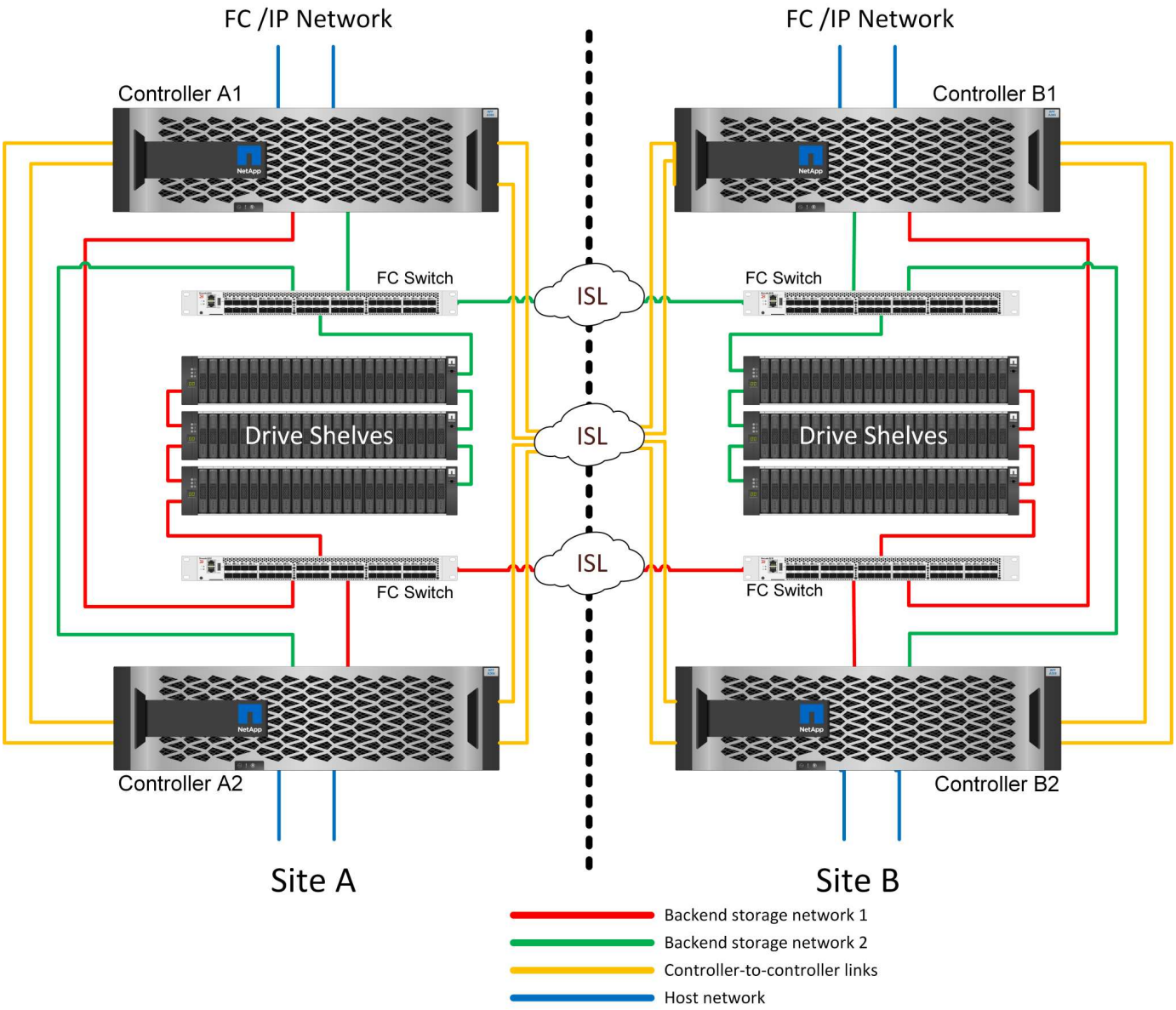
Auch die Architektur ist einfacher, da die einzigen standortübergreifenden Verbindungen für die Controller gelten. Bei FC SAN Attached MetroCluster schreibt ein Controller direkt auf die Laufwerke am entgegengesetzten Standort und benötigt somit zusätzliche SAN-Verbindungen, Switches und Bridges. Ein Controller in einer IP-Konfiguration hingegen schreibt über den Controller auf die entgegengesetzten Laufwerke.

Weitere Informationen finden Sie in der offiziellen ONTAP-Dokumentation und "Architektur und Design der MetroCluster IP-Lösung".



HA-Paar FC SAN Attached MetroCluster

Die HA-Paar-Konfiguration von MetroCluster FC nutzt zwei oder vier Nodes pro Standort. Diese Konfigurationsoption erhöht die Komplexität und die Kosten im Vergleich zur Option mit zwei Nodes, bietet aber einen wichtigen Vorteil: intrasite-Redundanz. Bei einem einfachen Controller-Ausfall ist kein Datenzugriff über das WAN erforderlich. Der Datenzugriff bleibt über den alternativen lokalen Controller lokal.



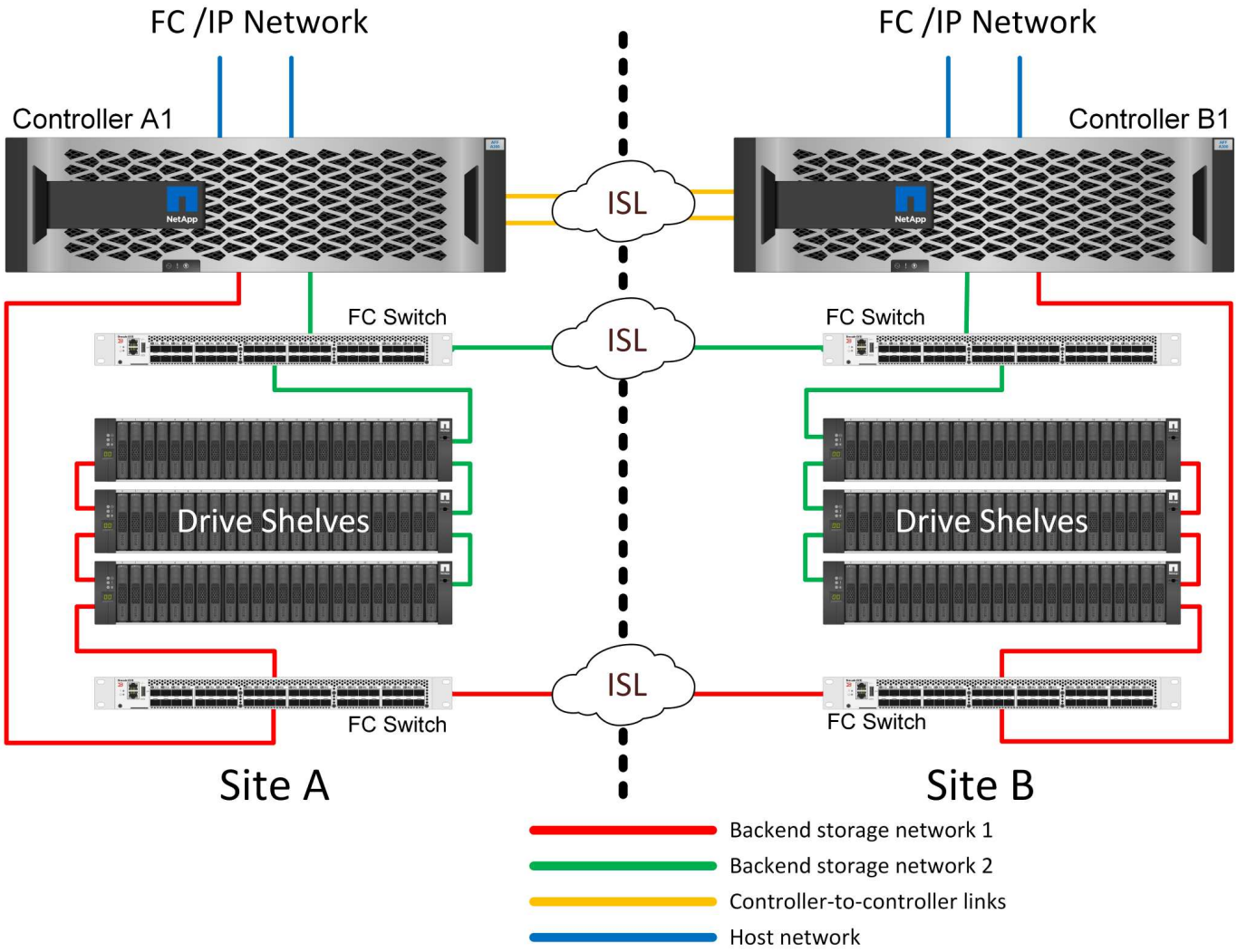
Einige Infrastrukturen mehrerer Standorte sind nicht für den aktiv/aktiv-Betrieb konzipiert, sondern werden eher als primärer Standort und Disaster-Recovery-Standort genutzt. In dieser Situation ist eine MetroCluster-Option für HA-Paare aus den folgenden Gründen im Allgemeinen vorzuziehen:

- Obwohl es sich bei einem MetroCluster Cluster mit zwei Nodes um ein HA-System handelt, müssen für einen unerwarteten Ausfall eines Controllers oder einer geplanten Wartung die Datenservices am anderen Standort online geschaltet werden. Wenn die Netzwerkverbindung zwischen Standorten die erforderliche Bandbreite nicht unterstützen kann, ist die Performance beeinträchtigt. Die einzige Option wäre auch ein Failover der verschiedenen Host-Betriebssysteme und der damit verbundenen Services zum alternativen Standort. Das HA-Paar MetroCluster Cluster eliminiert dieses Problem, da der Verlust eines Controllers zu einfachem Failover innerhalb desselben Standorts führt.
- Einige Netzwerktopologien sind nicht für den standortübergreifenden Zugriff ausgelegt, sondern verwenden stattdessen unterschiedliche Subnetze oder isolierte FC-SANs. In diesen Fällen fungiert der MetroCluster Cluster mit zwei Nodes nicht mehr als HA-System, da der alternative Controller keine Daten für die Server am gegenüberliegenden Standort bereitstellen kann. Um vollständige Redundanz zu gewährleisten, ist die MetroCluster Option für das HA-Paar erforderlich.
- Wird eine Infrastruktur mit zwei Standorten als eine einzelne hochverfügbare Infrastruktur angesehen, eignet sich die MetroCluster Konfiguration mit zwei Nodes. Falls das System jedoch nach einem

Standortausfall über einen längeren Zeitraum hinweg funktionieren muss, ist ein HA-Paar vorzuziehen, da es weiterhin HA innerhalb eines einzelnen Standorts bereitstellen muss.

FC SAN-Attached MetroCluster mit zwei Nodes

Die MetroCluster Konfiguration mit zwei Nodes verwendet nur einen Node pro Standort. Dieses Design ist einfacher als die Option für HA-Paare, da weniger Komponenten konfiguriert und gewartet werden müssen. Zudem wurden die Infrastrukturanforderungen hinsichtlich Verkabelung und FC-Switching gesenkt. Und schließlich senkt es die Kosten.



Ein solches Design hat ganz offensichtlich zur Folge, dass der Controller-Ausfall an einem einzigen Standort dazu führt, dass die Daten am entgegengesetzten Standort verfügbar sind. Diese Einschränkung ist nicht unbedingt ein Problem. Viele Unternehmen verfügen über standortübergreifende Datacenter-Betriebsabläufe mit verteilten, schnellen Netzwerken mit niedriger Latenz, die im Wesentlichen als eine einzige Infrastruktur fungieren. In diesen Fällen ist die MetroCluster Version mit zwei Nodes die bevorzugte Konfiguration. Systeme mit zwei Nodes werden derzeit im Petabyte-Bereich von mehreren Service-Providern eingesetzt.

Funktionen zur Ausfallsicherheit von MetroCluster

Es gibt keine Single Points of Failure in einer MetroCluster Lösung:

- Jeder Controller verfügt über zwei unabhängige Pfade zu den Laufwerk-Shelfs am lokalen Standort.

- Jeder Controller verfügt über zwei unabhängige Pfade zu den Laufwerk-Shelfs am Remote-Standort.
- Jeder Controller verfügt über zwei unabhängige Pfade zu den Controllern am gegenüberliegenden Standort.
- In der HA-Paar-Konfiguration besitzt jeder Controller zwei Pfade zu seinem lokalen Partner.

Zusammenfassend lässt sich sagen, dass jede Komponente der Konfiguration entfernt werden kann, ohne dass die Fähigkeit von MetroCluster zur Datenbereitstellung beeinträchtigt wird. Der einzige Unterschied in Bezug auf die Ausfallsicherheit zwischen den beiden Optionen ist, dass die HA-Paar-Version nach einem Standortausfall weiterhin ein insgesamt HA-Storage-System ist.

Logische Architektur

Um zu verstehen, wie Oracle-Datenbanken in einer MetroCluster-Umgebung funktionieren, bedarf es einer Erklärung der logischen Funktionalität eines MetroCluster-Systems.

Schutz vor Standortausfällen: NVRAM und MetroCluster

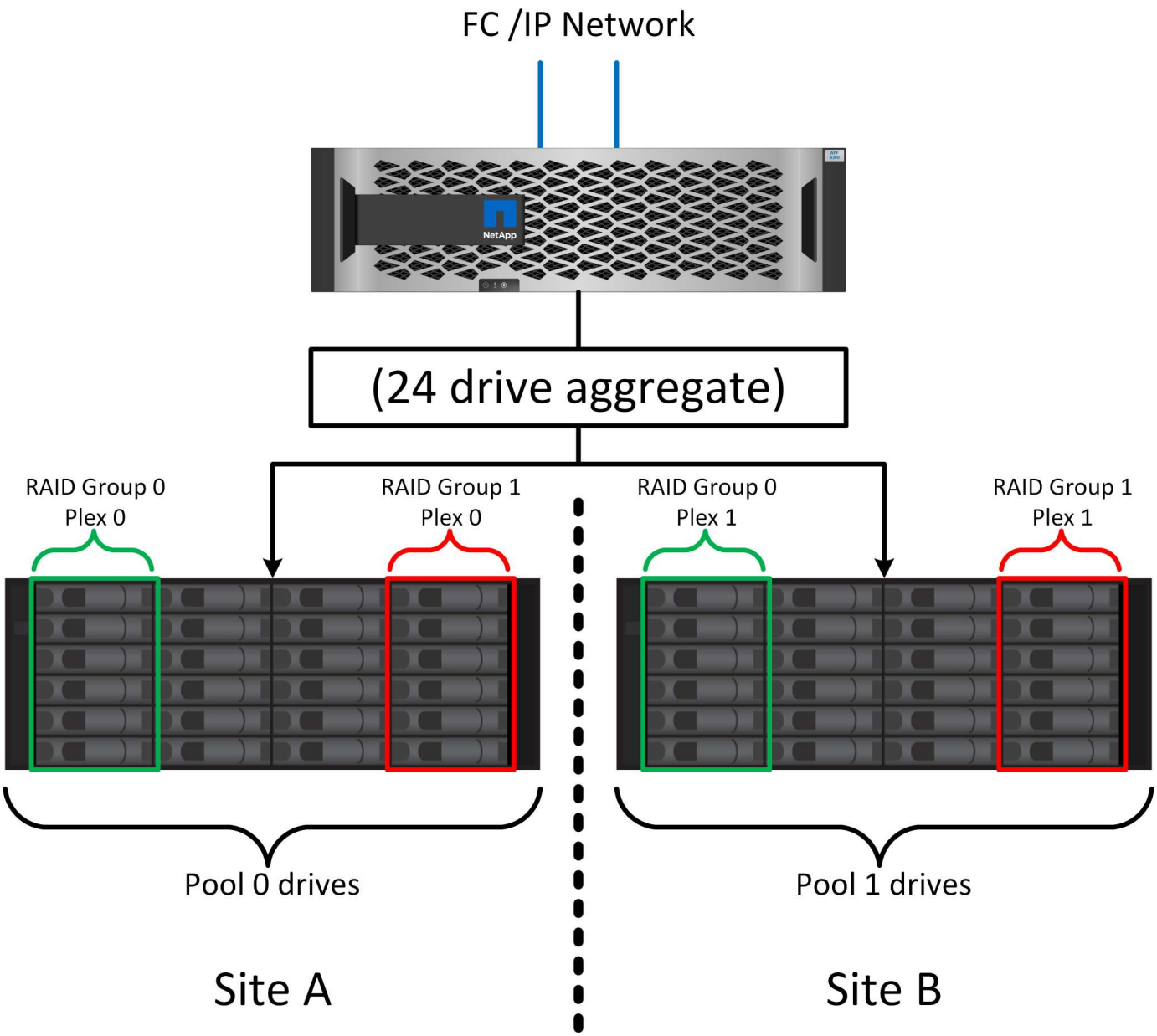
MetroCluster erweitert die NVRAM-Datensicherung auf folgende Weise:

- In einer Konfiguration mit zwei Nodes werden NVRAM-Daten mithilfe von Inter-Switch Links (ISLs) zum Remote-Partner repliziert.
- In einer HA-Paar-Konfiguration werden NVRAM-Daten sowohl auf den lokalen Partner als auch auf einen Remote-Partner repliziert.
- Ein Schreibvorgang wird erst bestätigt, wenn er für alle Partner repliziert wird. Diese Architektur schützt aktive I/O-Vorgänge vor Standortausfällen, indem NVRAM-Daten zu einem Remote-Partner repliziert werden. Dieser Prozess ist nicht mit der Datenreplizierung auf Laufwerksebene verbunden. Der Controller, der die Aggregate besitzt, ist für die Datenreplizierung verantwortlich, indem er auf beide Plexe im Aggregat schreibt. Bei einem Standortausfall muss jedoch weiterhin ein Schutz vor inaktiven I/O-Datenverlusten gewährleistet sein. Replizierte NVRAM-Daten werden nur verwendet, wenn ein Partner-Controller für einen ausgefallenen Controller übernehmen muss.

Schutz vor Standort- und Shelf-Ausfällen: SyncMirror und Plexe

SyncMirror ist eine Spiegelungstechnologie, die RAID DP oder RAID-TEC verbessert, aber nicht ersetzt. Es spiegelt den Inhalt von zwei unabhängigen RAID-Gruppen. Die logische Konfiguration ist wie folgt:

1. Laufwerke werden je nach Standort in zwei Pools konfiguriert. Ein Pool besteht aus allen Laufwerken an Standort A und der zweite Pool besteht aus allen Laufwerken an Standort B
2. Ein gemeinsamer Storage Pool, auch bekannt als Aggregat, wird dann auf der Basis gespiegelter Gruppen von RAID-Gruppen erstellt. Von jedem Standort wird eine gleiche Anzahl von Laufwerken gezogen. Ein SyncMirror Aggregat für 20 Laufwerke würde beispielsweise aus 10 Laufwerken an Standort A und 10 Laufwerken an Standort B bestehen
3. Jeder Laufwerkssatz an einem bestimmten Standort wird automatisch als eine oder mehrere vollständig redundante RAID DP- oder RAID-TEC-Gruppen konfiguriert, und zwar unabhängig von der Verwendung von Spiegelung. Diese Verwendung von RAID unter der Spiegelung bietet Datensicherheit auch nach dem Verlust eines Standorts.



Die Abbildung oben zeigt eine Beispiel-SyncMirror-Konfiguration. Es wurde ein Aggregat mit 24 Laufwerken auf dem Controller mit 12 Laufwerken aus einem an Standort A zugewiesenen Shelf und 12 Laufwerken aus einem an Standort B zugewiesenen Shelf erstellt Die Laufwerke wurden in zwei gespiegelte RAID-Gruppen gruppiert. RAID-Gruppe 0 enthält einen Plex mit 6 Laufwerken an Standort A, der auf einen Plex mit 6 Laufwerken an Standort B gespiegelt wurde Ebenso enthält die RAID-Gruppe 1 einen Plex mit 6 Laufwerken an Standort A, der auf einen Plex mit 6 Laufwerken an Standort B gespiegelt wird

Normalerweise wird SyncMirror für die Remote-Spiegelung bei MetroCluster Systemen verwendet, wobei eine Kopie der Daten an jedem Standort vorhanden ist. Gelegentlich wurde es verwendet, um eine zusätzliche Redundanz in einem einzigen System bereitzustellen. Insbesondere bietet sie Redundanz auf Shelf-Ebene. Ein Festplatten-Shelf enthält bereits duale Netzteile und Controller und ist im Großen und Ganzen etwas mehr als Bleche, doch in einigen Fällen ist möglicherweise der zusätzliche Schutz gewährleistet. Ein NetApp Kunde beispielsweise hat SyncMirror für eine mobile Echtzeitanalyse-Plattform für Automobiltests implementiert. Das System wurde in zwei physische Racks mit unabhängigen Stromeinspeisungen und unabhängigen USV-Systemen getrennt.

Redundanzfehler: NVFAIL

Wie zuvor bereits erläutert, wird ein Schreibvorgang erst bestätigt, wenn er in lokalem NVRAM und NVRAM auf mindestens einem anderen Controller angemeldet wurde. Dieser Ansatz stellt sicher, dass ein Hardware-Ausfall oder ein Stromausfall nicht zum Verlust der aktiven I/O führen Wenn der lokale NVRAM ausfällt oder die Verbindung zu anderen Nodes ausfällt, werden die Daten nicht mehr gespiegelt.

Wenn der lokale NVRAM einen Fehler meldet, wird der Node heruntergefahren. Dieses Herunterfahren führt zu einem Failover auf einen Partner-Controller, wenn HA-Paare verwendet werden. Bei MetroCluster hängt das Verhalten von der gewählten Gesamtkonfiguration ab, kann jedoch zu einem automatischen Failover auf die entfernte Notiz führen. In jedem Fall gehen keine Daten verloren, da der Controller den Schreibvorgang nicht bestätigt hat.

Komplizierter wird dies, wenn die Verbindung zwischen Standorten ausfällt, die die NVRAM-Replizierung auf Remote-Nodes blockiert. Schreibvorgänge werden nicht mehr auf die Remote-Nodes repliziert. Dadurch besteht die Möglichkeit eines Datenverlusts, falls ein schwerwiegender Fehler auf einem Controller auftritt. Noch wichtiger ist, dass der Versuch, während dieser Bedingungen ein Failover auf einen anderen Node durchzuführen, zu Datenverlust führt.

Der Steuerungsfaktor ist, ob NVRAM synchronisiert wird. Bei NVRAM-Synchronisierung kann ein Node-to-Node Failover ohne das Risiko eines Datenverlusts fortgesetzt werden. Wenn in einer MetroCluster Konfiguration NVRAM und die zugrunde liegenden Aggregat-Plexe synchron sind, kann ohne das Risiko eines Datenverlusts eine Umschaltung durchgeführt werden.

ONTAP lässt kein Failover oder Switchover zu, wenn die Daten nicht synchron sind, es sei denn, das Failover oder die Umschaltung ist erzwungen. Durch das Erzwingen einer solchen Änderung der Bedingungen wird bestätigt, dass Daten im ursprünglichen Controller zurückgelassen werden können und dass ein Datenverlust akzeptabel ist.

Datenbanken und andere Applikationen sind besonders anfällig für Beschädigungen, wenn ein Failover oder Switchover erzwungen wird, da sie größere interne Daten-Caches auf Festplatten beibehalten. Wenn ein erzwungenes Failover oder eine Umschaltung auftritt, werden zuvor bestätigte Änderungen effektiv verworfen. Der Inhalt des Storage Arrays springt effektiv zurück in die Zeit, und der Cache-Status gibt nicht mehr den Status der Daten auf der Festplatte wieder.

Um dies zu verhindern, können Volumes mit ONTAP für speziellen Schutz vor NVRAM-Ausfällen konfiguriert werden. Wenn dieser Schutzmechanismus ausgelöst wird, gelangt ein Volume in den Status „NVFAIL“. Dieser Zustand führt zu I/O-Fehlern, die einen Absturz der Applikation verursachen. Dieser Absturz führt dazu, dass die Applikationen heruntergefahren werden, damit keine veralteten Daten verwendet werden. Daten dürfen nicht verloren gehen, da alle festzugebenden Transaktionsdaten in den Protokollen vorhanden sein sollten. Als Nächstes muss ein Administrator die Hosts vollständig herunterfahren, bevor die LUNs und Volumes manuell wieder online geschaltet werden. Obwohl diese Schritte etwas Arbeit erfordern können, ist dieser Ansatz der sicherste Weg, um die Datenintegrität zu gewährleisten. Nicht alle Daten erfordern diesen Schutz. Daher kann ein NVFAIL-Verhalten auf Volume-Basis konfiguriert werden.

HA-Paare und MetroCluster

MetroCluster ist in zwei Konfigurationen erhältlich: Zwei Nodes und ein HA-Paar. Die Konfiguration mit zwei Nodes verhält sich in Bezug auf NVRAM wie ein HA-Paar. Im Falle eines plötzlichen Ausfalls kann der Partner-Node NVRAM-Daten wiedergeben, um die Laufwerke konsistent zu machen und sicherzustellen, dass keine bestätigten Schreibvorgänge verloren gegangen sind.

Die HA-Paar-Konfiguration repliziert NVRAM auch auf den lokalen Partner-Node. Ein einfacher Controller-Ausfall führt zu einer NVRAM-Wiedergabe auf dem Partner-Node, wie dies bei einem Standalone HA-Paar ohne MetroCluster der Fall ist. Bei einem plötzlichen vollständigen Standortausfall verfügt der Remote Standort

außerdem über den NVRAM, der erforderlich ist, um die Laufwerke konsistent zu gestalten und Daten bereitzustellen.

Ein wichtiger Aspekt von MetroCluster ist, dass die Remote Nodes unter normalen Betriebsbedingungen keinen Zugriff auf Partnerdaten haben. Jeder Standort funktioniert im Wesentlichen als ein unabhängiges System, das die Persönlichkeit des gegenüberliegenden Standorts übernehmen kann. Dieser Prozess wird als Umschaltung bezeichnet und umfasst ein geplantes Switchover, bei dem Standortvorgänge unterbrechungsfrei zum anderen Standort migriert werden. Auch ungeplante Situationen, in denen ein Standort verloren geht und bei der Disaster Recovery ein manuelles oder automatisches Switchover erforderlich ist, werden berücksichtigt.

Umschaltung und Switchback

Die Begriffe Switchover und Switchback beziehen sich auf den Prozess, bei dem Volumes zwischen Remote Controllern in einer MetroCluster Konfiguration migriert werden. Dieser Vorgang gilt nur für die Remote-Knoten. Wenn MetroCluster in einer Konfiguration mit vier Volumes zum Einsatz kommt, entspricht das lokale Node Failover dem zuvor beschriebenen Takeover- und Giveback-Prozess.

Geplante Umschaltung und Umschaltung

Ein geplanter Switchover oder Switchback ähnelt einer Übernahme oder einem Giveback zwischen Nodes. Der Prozess umfasst mehrere Schritte und scheint möglicherweise mehrere Minuten zu erfordern. Aber was wirklich geschieht, ist eine mehrstufige Übertragung der Storage- und Netzwerkressourcen. Der Moment, in dem Kontrolltransfers schneller erfolgen, als der vollständige Befehl ausgeführt werden muss.

Der Hauptunterschied zwischen Takeover/Giveback und Switchover/Switchback besteht in den Auswirkungen auf die FC SAN-Konnektivität. Durch lokale Übernahme/Giveback wird der Verlust aller FC-Pfade zum lokalen Node durch den Host erlebbar und verlässt sich auf natives MPIO, um auf verfügbare alternative Pfade umzusteigen. Ports werden nicht verlegt. Mit Switchover und Switchback werden die virtuellen FC-Ziel-Ports der Controller zum anderen Standort übertragen. Sie existieren praktisch einen Moment lang nicht mehr auf dem SAN und werden dann auf einem alternativen Controller wieder angezeigt.

SyncMirror-Timeouts

Bei SyncMirror handelt es sich um eine ONTAP-Spiegelungstechnologie, die Schutz vor Shelf-Ausfällen bietet. Wenn Shelfs über eine Entfernung voneinander getrennt sind, führt dies zu einer Remote-Datensicherung.

SyncMirror bietet kein universelles synchrones Spiegeln. Das Ergebnis ist eine höhere Verfügbarkeit. Einige Speichersysteme nutzen eine konstante Spiegelung alles oder nichts, die manchmal auch Domino-Modus genannt wird. Diese Form der Spiegelung ist in der Anwendung beschränkt, da alle Schreibaktivitäten unterbrochen werden müssen, wenn die Verbindung zum Remote-Standort verloren geht. Andernfalls würde ein Schreiben an einer Stelle, aber nicht an der anderen existieren. Solche Umgebungen sind normalerweise so konfiguriert, dass LUNs offline geschaltet werden, wenn die Verbindung zwischen Standorten länger als einen kurzen Zeitraum (wie etwa 30 Sekunden) unterbrochen wird.

Dieses Verhalten ist für eine kleine Untermenge von Umgebungen wünschenswert. Die meisten Anwendungen benötigen jedoch eine Lösung, die eine garantierte synchrone Replikation unter normalen Betriebsbedingungen bietet, aber die Replikation unterbrechen kann. Ein vollständiger Verlust der Verbindung zwischen Standorten wird häufig als nahezu katastrophennahe Situation betrachtet. In der Regel werden solche Umgebungen online gehalten und stellen Daten bereit, bis die Konnektivität repariert wird oder eine formale Entscheidung getroffen wird, die Umgebung zum Schutz der Daten herunterzufahren. Eine Notwendigkeit für das automatische Herunterfahren der Anwendung allein aufgrund eines Fehlers bei der Remote-Replikation ist ungewöhnlich.

SyncMirror unterstützt Anforderungen an die synchrone Spiegelung mit der Flexibilität einer

Zeitüberschreitung. Wenn die Verbindung zum Remote-Controller und/oder Plex unterbrochen wird, beginnt ein 30-Sekunden-Timer zu zählen. Wenn der Zähler 0 erreicht, wird die Schreib-I/O-Verarbeitung mithilfe der lokalen Daten fortgesetzt. Die Remote-Kopie der Daten ist nutzbar, wird aber rechtzeitig eingefroren, bis die Verbindung wiederhergestellt ist. Die Neusynchronisierung nutzt Snapshots auf Aggregatebene, um das System so schnell wie möglich in den synchronen Modus zurückzuversetzen.

Bemerkenswert ist, dass in vielen Fällen diese Art universeller Domino-Modus-Replikation auf Anwendungsebene besser implementiert wird. Beispielsweise verfügt Oracle DataGuard über einen maximalen Schutzmodus, der unter allen Umständen eine Replizierung mit einer langen Instanz garantiert. Wenn die Replikationsverbindung für einen Zeitraum fehlschlägt, der ein konfigurierbares Timeout überschreitet, werden die Datenbanken heruntergefahren.

Automatische, unbeaufsichtigte Umschaltung mit Fabric Attached MetroCluster

AUSO (Automatic unbeaufsichtigter Switchover) ist eine Fabric Attached MetroCluster Funktion, die eine Form standortübergreifender Hochverfügbarkeit bietet. Wie zuvor erläutert, gibt es bei MetroCluster zwei Typen: Einen einzigen Controller an jedem Standort oder ein HA-Paar an jedem Standort. Der Hauptvorteil der HA-Option besteht darin, dass bei geplanter oder ungeplanter Controller-Abschaltung alle I/O-Vorgänge weiterhin lokal ausgeführt werden können. Der Vorteil der Single-Node-Option liegt in der Reduzierung der Kosten, der Komplexität und der Infrastruktur.

Der wichtigste Vorteil von AUSO ist die Verbesserung der Hochverfügbarkeitsfunktionen von Fabric Attached MetroCluster Systemen. Jeder Standort überwacht den Zustand des anderen Standorts. Falls kein Node mehr vorhanden ist, um Daten bereitzustellen, ermöglicht AUSO ein schnelles Switchover. Dieser Ansatz erweist sich insbesondere für MetroCluster Konfigurationen mit nur einem einzigen Node pro Standort, da er die Konfiguration in Bezug auf die Verfügbarkeit näher an ein HA-Paar bringt.

AUSO kann auf Ebene eines HA-Paars kein umfassendes Monitoring bieten. Ein HA-Paar kann für eine extrem hohe Verfügbarkeit sorgen, da es zwei redundante physische Kabel für eine direkte Kommunikation zwischen den Nodes umfasst. Darüber hinaus haben beide Nodes in einem HA-Paar Zugriff auf den gleichen Satz an Festplatten in redundanten Loops, die einen weiteren Weg für einen Node zur Überwachung des Systemzustands eines anderen bereitstellen.

MetroCluster Cluster sind über Standorte verteilt, bei denen sowohl die Node-to-Node-Kommunikation als auch der Festplattenzugriff auf die Site-to-Site-Netzwerkverbindung angewiesen sind. Die Fähigkeit, den Heartbeat des restlichen Clusters zu überwachen, ist begrenzt. AUSO muss zwischen Situationen unterscheiden, in denen der andere Standort aufgrund eines Netzwerkproblems nicht verfügbar ist, sondern tatsächlich ausgefallen ist.

So kann ein Controller in einem HA-Paar eine Übernahme veranlassen, wenn ein Controller-Ausfall erkannt wird, der aus einem bestimmten Grund, wie z. B. einem Systempanik, aufgetreten ist. Es kann auch zu einem Takeover führen, wenn ein vollständiger Verbindungsverlust besteht, manchmal auch als verlorener Herzschlag bezeichnet.

Ein MetroCluster System kann eine automatische Umschaltung nur sicher durchführen, wenn ein bestimmter Fehler am ursprünglichen Standort erkannt wird. Darüber hinaus muss der Controller, der das Storage-System übernimmt, in der Lage sein, die Synchronisierung von Festplatten- und NVRAM-Daten zu gewährleisten. Der Controller kann die Sicherheit einer Umschaltung nicht garantieren, nur weil er den Kontakt zum Quellstandort verloren hat, der noch betriebsbereit sein könnte. Weitere Optionen zur Automatisierung einer Umschaltung finden Sie im nächsten Abschnitt zur MetroCluster Tiebreaker Lösung (MCTB).

MetroCluster Tiebreaker mit Fabric Attached MetroCluster

Die "NetApp MetroCluster Tiebreaker" Software kann an einem dritten Standort ausgeführt werden, um den Zustand der MetroCluster Umgebung zu überwachen, Benachrichtigungen zu senden und in einer

Notfallsituation optional ein Switchover zu erzwingen. Eine vollständige Beschreibung des Tiebreaker finden Sie auf dem "NetApp Support Website", aber der primäre Zweck des MetroCluster Tiebreaker ist das Erkennen von Standortausfällen. Außerdem muss zwischen Standortausfällen und Verbindungsverlust unterschieden werden. So sollte beispielsweise keine Umschaltung erfolgen, da der primäre Standort nicht erreichbar war. Aus diesem Grund überwacht Tiebreaker auch die Fähigkeit des Remote-Standorts, mit dem primären Standort in Kontakt zu treten.

Die automatische Umschaltung mit AUSO ist auch mit der MCTB kompatibel. AUSO reagiert sehr schnell, da es darauf ausgelegt ist, bestimmte Fehlerereignisse zu erkennen und dann die Umschaltung nur dann aufzurufen, wenn NVRAM und SyncMirror Plexe synchron sind.

Im Gegensatz dazu befindet sich das Tiebreaker Remote und muss daher warten, bis ein Timer verstrichen ist, bevor ein Standort für tot erklärt wird. Über Tiebreaker wird schließlich festgestellt, wie ein Controller-Ausfall von AUSO abgedeckt ist, doch im Allgemeinen hat AUSO bereits die Umschaltung gestartet und möglicherweise die Umschaltung abgeschlossen, bevor es Tiebreaker wirkt. Der resultierende zweite Switchover-Befehl aus dem Tiebreaker würde abgelehnt.

Die MCTB-Software überprüft nicht, ob NVRAM und/oder Plexe synchronisiert sind, wenn eine Umschaltung erzwungen wird. Sofern konfiguriert, sollte die automatische Umschaltung während Wartungsaktivitäten deaktiviert werden, die zu einem Verlust der Synchronisierung von NVRAM- oder SyncMirror-Plexen führen.

Darüber hinaus geht die MCTB möglicherweise nicht bei einem rollierenden Notfall ein, der zu der folgenden Ereignisabfolge führt:

1. Die Konnektivität zwischen Standorten wird für mehr als 30 Sekunden unterbrochen.
2. Die SyncMirror-Replizierung ist zeitgemäß, und der Betrieb wird am primären Standort fortgesetzt, sodass das Remote-Replikat nicht mehr zeitgemäß ist.
3. Der primäre Standort geht verloren.das Ergebnis sind nicht replizierte Änderungen am primären Standort. Eine Umschaltung könnte dann aus verschiedenen Gründen unerwünscht sein, unter anderem aus folgenden Gründen:
    - Am primären Standort befinden sich möglicherweise kritische Daten, und diese Daten können nach und nach wiederhergestellt werden. Mit einer Umschaltung, die eine Weiterführung des Betriebs der Applikation ermöglichte, würden die kritischen Daten praktisch verworfen.
    - Möglicherweise haben Daten im Cache einer Applikation gespeichert, die am verbleibenden Standort zum Zeitpunkt des Standortverlusts die Storage-Ressourcen am primären Standort nutzte. Durch ein Switchover würde eine veraltete Version der Daten eingeführt, die nicht mit dem Cache übereinstimmt.
    - Möglicherweise haben Daten im Cache eines Betriebssystems, das auf dem verbleibenden Standort zum Zeitpunkt eines Standortausfalls Speicherressourcen am primären Standort genutzt hat, gespeichert. Durch ein Switchover würde eine veraltete Version der Daten eingeführt, die nicht mit dem Cache übereinstimmt. Am sichersten ist es, dass Sie Tiebreaker so konfigurieren, dass eine Warnmeldung ausgegeben wird, wenn ein Standortausfall erkannt wird und anschließend eine Person Entscheidungen darüber treffen muss, ob eine Umschaltung erzwungen werden soll. Applikationen und/oder Betriebssysteme müssen möglicherweise zunächst heruntergefahren werden, um zwischengespeicherte Daten zu löschen. Darüber hinaus können die NVFAIL-Einstellungen verwendet werden, um einen zusätzlichen Schutz zu bieten und den Failover-Prozess zu rationalisieren.

ONTAP Mediator mit MetroCluster IP

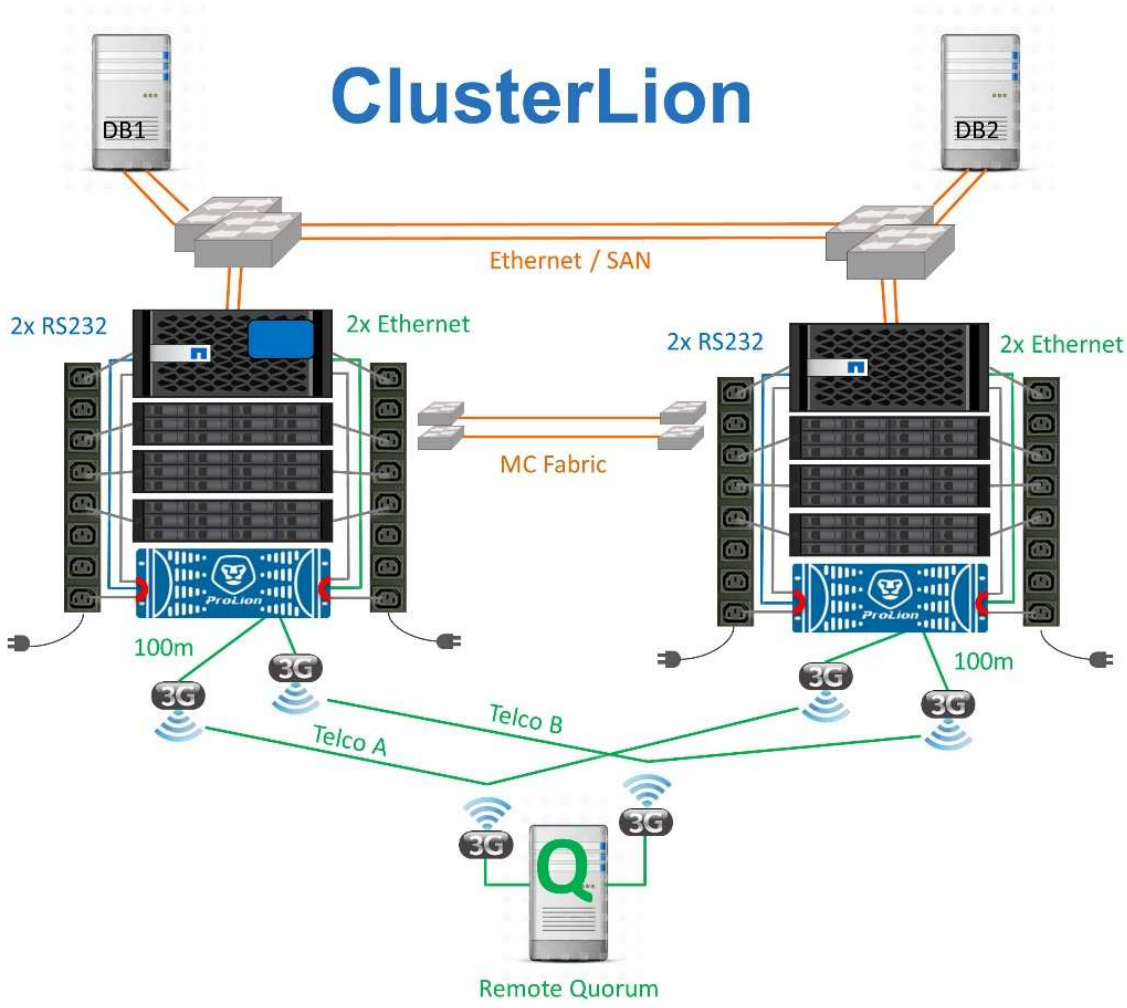
Der ONTAP Mediator wird mit MetroCluster IP und bestimmten anderen ONTAP-Lösungen verwendet. Es fungiert als herkömmlicher Tiebreaker Service, ähnlich wie die oben beschriebene MetroCluster Tiebreaker Software, verfügt aber auch über eine wichtige Funktion zum automatisierten, unbeaufsichtigten Switchover.

Ein Fabric-Attached MetroCluster hat direkten Zugriff auf die Storage-Geräte am gegenüberliegenden Standort. Dadurch kann ein MetroCluster-Controller den Zustand der anderen Controller überwachen, indem er die Heartbeat-Daten von den Laufwerken liest. So kann ein Controller den Ausfall eines anderen Controllers erkennen und eine Umschaltung durchführen.

Im Gegensatz dazu leitet die MetroCluster IP Architektur alle I/O ausschließlich über die Controller-Controller-Verbindung weiter; es besteht kein direkter Zugriff auf Speichergeräte am Remote-Standort. Dadurch wird die Fähigkeit eines Controllers eingeschränkt, Ausfälle zu erkennen und eine Umschaltung durchzuführen. Der ONTAP Mediator ist daher als Tiebreaker-Gerät erforderlich, um Standortverluste zu erkennen und automatisch eine Umschaltung durchzuführen.

Virtueller dritter Standort mit ClusterLion

ClusterLion ist eine fortschrittliche MetroCluster Monitoring-Appliance, die als virtueller dritter Standort fungiert. Dieser Ansatz ermöglicht die sichere Implementierung von MetroCluster in einer Konfiguration mit zwei Standorten und einer vollständig automatisierten Umschaltfunktion. Des Weiteren kann ClusterLion zusätzliche Überwachung auf Netzwerkebene durchführen und Vorgänge nach der Umschaltung ausführen. Die vollständige Dokumentation ist bei ProLion erhältlich.



- Die ClusterLion Appliances überwachen den Zustand der Controller mit direkt angeschlossenem Ethernet und seriellen Kabeln.
- Die beiden Geräte sind über redundante 3G-Wireless-Verbindungen miteinander verbunden.

- Die Stromversorgung des ONTAP-Controllers erfolgt über interne Relais. Bei einem Standortausfall trennt ClusterLion, das ein internes USV-System enthält, die Stromanschlüsse, bevor eine Umschaltung initiiert wird. Dieser Prozess stellt sicher, dass kein Split-Brain-Zustand auftritt.
- ClusterLion führt eine Umschaltung innerhalb der SyncMirror-Zeitüberschreitung von 30 Sekunden oder überhaupt nicht aus.
- ClusterLion führt nur eine Umschaltung durch, wenn die Zustände NVRAM und SyncMirror Plexe synchron sind.
- Da ClusterLion nur umgeschaltet wird, wenn die MetroCluster vollständig synchron ist, ist das NVFAIL nicht erforderlich. Diese Konfiguration ermöglicht es, standortübergreifende Umgebungen wie beispielsweise einen erweiterten Oracle RAC auch während einer ungeplanten Umschaltung online zu bleiben.
- Die Unterstützung umfasst sowohl Fabric-Attached MetroCluster als auch MetroCluster IP

SyncMirror

Die Grundlage für die Oracle Datensicherung mit einem MetroCluster System ist SyncMirror, eine Technologie für die synchrone Spiegelung, die maximale Performance und horizontale Skalierbarkeit bietet.

Datensicherung mit SyncMirror

Auf der einfachsten Ebene bedeutet synchrone Replikation, dass jede Änderung an beiden Seiten des gespiegelten Speichers vorgenommen werden muss, bevor sie bestätigt wird. Wenn beispielsweise eine Datenbank ein Protokoll schreibt oder ein VMware Gast gepatcht wird, darf ein Schreibvorgang nie verloren gehen. Als Protokollebene darf das Storage-System den Schreibvorgang erst dann bestätigen, wenn es auf nichtflüchtigen Medien an beiden Standorten gespeichert wurde. Nur dann ist es sicher, ohne das Risiko eines Datenverlusts zu gehen.

Die Verwendung einer Technologie zur synchronen Replizierung ist der erste Schritt beim Entwurf und Management einer Lösung zur synchronen Replizierung. Die wichtigste Überlegung ist, zu verstehen, was in verschiedenen geplanten und ungeplanten Ausfallszenarien passieren könnte. Nicht alle Lösungen zur synchronen Replizierung bieten dieselben Funktionen. Wenn Sie eine Lösung benötigen, die einen Recovery Point Objective (RPO) von null bietet, d. h. keinen Datenverlust verursacht, müssen alle Ausfallszenarien in Betracht gezogen werden. Welches ist insbesondere das erwartete Ergebnis, wenn die Replikation aufgrund des Verlusts der Verbindung zwischen Standorten nicht möglich ist?

SyncMirror Datenverfügbarkeit

Die MetroCluster-Replizierung basiert auf der NetApp SyncMirror Technologie, mit der effizient in den synchronen Modus bzw. aus dem synchronen Modus gewechselt werden kann. Diese Funktion erfüllt die Anforderungen von Kunden, die synchrone Replizierung benötigen, aber auch Hochverfügbarkeit für ihre Datenservices benötigen. Wenn zum Beispiel die Verbindung zu einem Remote-Standort unterbrochen wird, ist es in der Regel besser, dass das Speichersystem weiterhin in einem nicht replizierten Zustand betrieben wird.

Viele Lösungen zur synchronen Replizierung können nur im synchronen Modus betrieben werden. Diese Art der alles-oder-nichts-Replikation wird manchmal Domino-Modus genannt. Solche Storage-Systeme stellen keine Daten mehr bereit, statt die lokalen und Remote-Kopien der Daten unsynchronisiert zu lassen. Wenn die Replikation gewaltsam unterbrochen wird, kann die Resynchronisierung äußerst zeitaufwendig sein und einen Kunden während der Wiederherstellung der Spiegelung einem vollständigen Datenverlust aussetzen.

SyncMirror kann nicht nur nahtlos aus dem synchronen Modus wechseln, wenn der Remote-Standort nicht erreichbar ist, sondern auch bei der Wiederherstellung der Konnektivität schnell zu einem RPO = 0-Zustand neu synchronisieren. Die veraltete Kopie der Daten am Remote-Standort kann während der

Resynchronisierung auch in einem nutzbaren Zustand aufbewahrt werden. Auf diese Weise ist gewährleistet, dass lokale und Remote-Kopien der Daten jederzeit vorhanden sind.

Wo der Domino-Modus erforderlich ist, bietet NetApp SnapMirror Synchronous (SM-S) an. Darüber hinaus gibt es Optionen auf Applikationsebene wie Oracle DataGuard oder SQL Server Always On Availability Groups. Für die Festplattenspiegelung auf Betriebssystemebene kann eine Option sein. Wenden Sie sich an Ihren NetApp oder Ihr Partner Account Team, um weitere Informationen und Optionen zu erhalten.

MetroCluster und NVFAIL

NVFAIL ist eine allgemeine Datenintegritätsfunktion in ONTAP, die darauf ausgelegt ist, die Datenintegrität in Datenbanken zu maximieren.

Dieser Abschnitt erweitert die Erläuterung der grundlegenden ONTAP NVFAIL, um MetroCluster-spezifische Themen zu behandeln.

Bei MetroCluster wird ein Schreibvorgang erst bestätigt, wenn er in lokalem NVRAM und NVRAM auf mindestens einem anderen Controller angemeldet wurde. Dieser Ansatz stellt sicher, dass ein Hardware-Ausfall oder ein Stromausfall nicht zum Verlust der aktiven I/O führen Wenn der lokale NVRAM ausfällt oder die Verbindung zu anderen Nodes ausfällt, werden die Daten nicht mehr gespiegelt.

Wenn der lokale NVRAM einen Fehler meldet, wird der Node heruntergefahren. Dieses Herunterfahren führt zu einem Failover auf einen Partner-Controller, wenn HA-Paare verwendet werden. Bei MetroCluster hängt das Verhalten von der gewählten Gesamtkonfiguration ab, kann jedoch zu einem automatischen Failover auf die entfernte Notiz führen. In jedem Fall gehen keine Daten verloren, da der Controller den Schreibvorgang nicht bestätigt hat.

Komplizierter wird dies, wenn die Verbindung zwischen Standorten ausfällt, die die NVRAM-Replizierung auf Remote-Nodes blockiert. Schreibvorgänge werden nicht mehr auf die Remote-Nodes repliziert. Dadurch besteht die Möglichkeit eines Datenverlusts, falls ein schwerwiegender Fehler auf einem Controller auftritt. Noch wichtiger ist, dass der Versuch, während dieser Bedingungen ein Failover auf einen anderen Node durchzuführen, zu Datenverlust führt.

Der Steuerungsfaktor ist, ob NVRAM synchronisiert wird. Bei NVRAM-Synchronisierung kann ein Node-to-Node Failover ohne das Risiko eines Datenverlusts fortgesetzt werden. Wenn in einer MetroCluster Konfiguration NVRAM und die zugrunde liegenden Aggregat-Plexe synchron sind, ist es sicher, mit der Umschaltung fortzufahren, ohne das Risiko eines Datenverlusts zu verursachen.

ONTAP lässt kein Failover oder Switchover zu, wenn die Daten nicht synchron sind, es sei denn, das Failover oder die Umschaltung ist erzwungen. Durch das Erzwingen einer solchen Änderung der Bedingungen wird bestätigt, dass Daten im ursprünglichen Controller zurückgelassen werden können und dass ein Datenverlust akzeptabel ist.

Datenbanken sind besonders anfällig für Beschädigungen, wenn ein Failover oder Switchover erzwungen wird, da Datenbanken größere interne Daten-Caches auf der Festplatte beibehalten. Wenn ein erzwungenes Failover oder eine Umschaltung auftritt, werden zuvor bestätigte Änderungen effektiv verworfen. Der Inhalt des Storage Arrays springt effektiv zurück in die Zeit, und der Zustand des Datenbank-Cache entspricht nicht mehr dem Status der Daten auf der Festplatte.

Um Applikationen vor dieser Situation zu schützen, können mit ONTAP Volumes für speziellen Schutz vor NVRAM-Ausfällen konfiguriert werden. Wenn dieser Schutzmechanismus ausgelöst wird, gelangt ein Volume in den Status „NVFAIL“. Dieser Status führt zu I/O-Fehlern, die dazu führen, dass Applikationen heruntergefahren werden, sodass keine veralteten Daten verwendet werden. Daten sollten nicht verloren gehen, da alle bestätigten Schreibvorgänge noch auf dem Speichersystem vorhanden sind, und bei

Datenbanken sollten alle festgeschriebenen Transaktionsdaten in den Protokollen vorhanden sein.

Als Nächstes muss ein Administrator die Hosts vollständig herunterfahren, bevor die LUNs und Volumes manuell wieder online geschaltet werden. Obwohl diese Schritte etwas Arbeit erfordern können, ist dieser Ansatz der sicherste Weg, um die Datenintegrität zu gewährleisten. Nicht alle Daten erfordern diesen Schutz. Daher kann ein NVFAIL-Verhalten auf Volume-Basis konfiguriert werden.

NVFAIL manuell erzwungen

Die sicherste Option, um ein Switchover mit einem Anwendungs-Cluster (einschließlich VMware, Oracle RAC und anderen) zu erzwingen, das über Standorte verteilt ist, ist durch Angabe `-force-nvfail-all` An der Kommandozeile. Diese Option ist als Notfallmaßnahme verfügbar, um sicherzustellen, dass alle zwischengespeicherten Daten gelöscht werden. Wenn ein Host Speicherressourcen verwendet, die sich ursprünglich am Standort mit Notfällen befinden, erhält er entweder I/O-Fehler oder eine veraltete Dateihandle (`ESTALE`) Fehler. Oracle Datenbanken stürzen ab und Dateisysteme gehen entweder vollständig offline oder wechseln in den schreibgeschützten Modus.

Nachdem die Umschaltung abgeschlossen ist, wird der angezeigt `in-nvfailed-state` Flag muss gelöscht werden und die LUNs müssen in den Online-Modus versetzt werden. Nach Abschluss dieser Aktivität kann die Datenbank neu gestartet werden. Diese Aufgaben können automatisiert werden, um die RTO zu reduzieren.

dr-Force-NV-Fehler

Stellen Sie als allgemeine Sicherheitsmaßnahme die ein `dr-force-nvfail` Markieren Sie alle Volumes, auf die während des normalen Betriebs von einem Remote-Standort aus zugegriffen werden kann, d. h. sie sind Aktivitäten, die vor dem Failover verwendet werden. Das Ergebnis dieser Einstellung ist, dass ausgewählte Remote-Volumes beim Aufrufen nicht mehr verfügbar sind `in-nvfailed-state` Während einer Umschaltung. Nachdem die Umschaltung abgeschlossen ist, wird der angezeigt `in-nvfailed-state` Flag muss gelöscht und die LUNs müssen in den Online-Modus versetzt werden. Nach Abschluss dieser Aktivitäten können die Anwendungen neu gestartet werden. Diese Aufgaben können automatisiert werden, um die RTO zu reduzieren.

Das Ergebnis ist wie bei der Verwendung von `-force-nvfail-all` Markierung für manuelle Umschaltung. Die Anzahl der betroffenen Volumes kann jedoch auf die Volumes beschränkt werden, die vor Anwendungen oder Betriebssystemen mit veralteten Caches geschützt werden müssen.

Es gibt zwei entscheidende Anforderungen an eine Umgebung, die nicht verwendet wird `dr-force-nvfail` Auf Anwendungsvolumes:

- Ein erzwungenes Switchover darf nicht mehr als 30 Sekunden nach dem Ausfall des primären Standorts erfolgen.
- Eine Umschaltung darf nicht während Wartungsaufgaben oder unter anderen Bedingungen erfolgen, unter denen SyncMirror Plexe oder NVRAM-Replikation nicht synchron sind. Die erste Anforderung ist über eine Tiebreaker Software möglich, die im Fall eines Standortausfalls innerhalb von 30 Sekunden umgeschaltet wird. Dies bedeutet jedoch nicht, dass die Umschaltung innerhalb von 30 Sekunden nach Erkennung eines Standortausfalls durchgeführt werden muss. Das bedeutet, dass es nicht mehr sicher ist, eine Umschaltung zu erzwingen, wenn 30 Sekunden vergangen sind, seit die Betriebsbereitschaft eines Standorts bestätigt wurde.

Die zweite Anforderung wird teilweise erfüllt, indem alle Funktionen zum automatisierten Switchover deaktiviert werden, wenn bekannt ist, dass die MetroCluster-Konfiguration nicht synchron ist. Eine bessere Option ist die Nutzung einer Tiebreaker Lösung, mit der der Systemzustand der NVRAM-Replizierung und der SyncMirror Plexe überwacht werden kann. Wenn das Cluster nicht vollständig synchronisiert ist, sollte Tiebreaker keine

Umschaltung auslösen.

Die NetApp-MCTB-Software kann den Synchronisierungsstatus nicht überwachen, daher sollte sie deaktiviert werden, wenn MetroCluster aus irgendeinem Grund nicht synchron ist. ClusterLion verfügt über Funktionen zur NVRAM-Überwachung und Plex-Überwachung und kann so konfiguriert werden, dass das Switchover nur ausgelöst wird, wenn für das MetroCluster-System eine vollständige Synchronisierung bestätigt wurde.

Oracle Single Instance

Wie bereits erwähnt, trägt das Vorhandensein eines MetroCluster-Systems nicht notwendigerweise zur Ergänzung oder Änderung von Best Practices für den Betrieb einer Datenbank bei. Bei den meisten Datenbanken, die derzeit auf MetroCluster Kundensystemen ausgeführt werden, handelt es sich um eine Einzelinstanz, und befolgen Sie die Empfehlungen in der Dokumentation zu Oracle auf ONTAP.

Failover mit einem vorkonfigurierten Betriebssystem

SyncMirror liefert eine synchrone Kopie der Daten am Disaster Recovery-Standort. Um diese Daten verfügbar zu machen, sind jedoch ein Betriebssystem und die zugehörigen Applikationen erforderlich. Eine grundlegende Automatisierung kann die Failover-Zeit der gesamten Umgebung deutlich verbessern. Clusterware Produkte wie Veritas Cluster Server (VCS) werden oft verwendet, um einen Cluster standortübergreifend zu erstellen, in vielen Fällen kann der Failover-Prozess mit einfachen Skripten angetrieben werden.

Wenn die primären Knoten verloren gehen, ist die Clusterware (oder Skripte) so konfiguriert, dass die Datenbanken am alternativen Standort online geschaltet werden. Eine Option besteht darin, Standby-Server zu erstellen, die für die NFS- oder SAN-Ressourcen, aus denen die Datenbank besteht, vorkonfiguriert sind. Wenn der primäre Standort ausfällt, führt die Clusterware- oder skriptbasierte Alternative eine Abfolge von Aktionen durch, die der folgenden ähneln:

1. Erzwingen einer MetroCluster-Umschaltung
2. Durchführen der Erkennung von FC-LUNs (nur SAN)
3. Mounten von Dateisystemen und/oder Mounten von ASM-Datenträgergruppen
4. Die Datenbank wird gestartet

Die primäre Anforderung dieses Ansatzes ist ein Betriebssystem, das am Remote Standort ausgeführt wird. Sie muss mit Oracle-Binärdateien vorkonfiguriert sein, was auch bedeutet, dass Aufgaben wie das Patching von Oracle am primären Standort und am Standby-Standort durchgeführt werden müssen. Alternativ können die Oracle Binärdateien auf den Remote-Standort gespiegelt und gemountet werden, wenn ein Notfall deklariert wird.

Die eigentliche Aktivierung ist einfach. Befehle wie die LUN-Erkennung erfordern nur einige wenige Befehle pro FC-Port. Das Mounten des Filesystems ist nichts anderes als ein `mount` Befehl, und sowohl Datenbanken als auch ASM können über die CLI mit einem einzigen Befehl gestartet und gestoppt werden. Wenn die Volumes und Dateisysteme vor dem Switchover nicht am Disaster-Recovery-Standort verwendet werden, müssen Sie sie nicht festlegen `dr-force- nvfail` Auf Volumes.

Failover mit einem virtualisierten Betriebssystem

Der Failover von Datenbankumgebungen kann auf das Betriebssystem selbst erweitert werden. In der Theorie kann dieses Failover mit Boot-LUNs durchgeführt werden, meistens erfolgt es jedoch mit einem virtualisierten Betriebssystem. Das Verfahren ähnelt den folgenden Schritten:

1. Erzwingen einer MetroCluster-Umschaltung
2. Mounten der Datenspeicher, die die virtuellen Maschinen des Datenbankservers hosten
3. Starten der virtuellen Maschinen
4. Manuelles Starten von Datenbanken oder Konfigurieren der virtuellen Maschinen, um die Datenbanken automatisch zu starten, z. B. kann ein ESX-Cluster mehrere Standorte umfassen. Bei einem Notfall können die Virtual Machines nach dem Switchover am Disaster Recovery-Standort online geschaltet werden. Solange die Datastores, die die virtualisierten Datenbankserver hosten, zum Zeitpunkt des Ausfalls nicht verwendet werden, ist keine Einstellung erforderlich `dr-force- nvfail` Auf zugeordneten Volumes.

Oracle Extended RAC

Viele Kunden optimieren ihre RTO, indem sie einen Oracle RAC Cluster über mehrere Standorte verteilen und damit eine vollständig aktiv/aktiv-Konfiguration erzielen. Das gesamte Design wird komplizierter, da es die Quorumverwaltung von Oracle RAC beinhalten muss. Außerdem erfolgt der Datenzugriff von beiden Standorten aus. Ein forcierter Switchover kann dazu führen, dass eine veraltete Kopie der Daten verwendet wird.

Obwohl eine Kopie der Daten auf beiden Standorten vorhanden ist, kann nur der Controller, der derzeit Eigentümer eines Aggregats ist, Daten bereitstellen. Daher müssen bei erweiterten RAC-Clustern die Remote-Knoten I/O über eine Site-to-Site-Verbindung durchführen. Es kommt zu zusätzlicher I/O-Latenz, aber diese Latenz ist im Allgemeinen kein Problem. Das RAC Interconnect-Netzwerk muss auch über mehrere Standorte verteilt sein, was bedeutet, dass ohnehin ein High-Speed-Netzwerk mit niedriger Latenz erforderlich ist. Falls die zusätzliche Latenz ein Problem verursacht, kann das Cluster aktiv/Passiv betrieben werden. I/O-intensive Vorgänge müssten dann zu den RAC-Knoten geleitet werden, die lokal zu dem Controller sind, der die Aggregate besitzt. Die Remote-Knoten führen dann weniger I/O-Vorgänge aus oder werden ausschließlich als Warm-Standby-Server verwendet.

Wenn ein erweiterter aktiv/aktiv-RAC erforderlich ist, sollte die aktive SnapMirror-Synchronisierung anstelle von MetroCluster in Betracht gezogen werden. Die SM-AS-Replikation ermöglicht die bevorzugte Replikation der Daten. Daher kann ein erweiterter RAC-Cluster erstellt werden, in dem alle Lesevorgänge lokal stattfinden. Die Lese-I/O-Vorgänge gehen nie über Standorte hinweg, wodurch die geringstmögliche Latenz erzielt wird. Alle Schreibvorgänge müssen weiterhin die Verbindung zwischen den Standorten übertragen, dieser Traffic ist bei jeder Lösung mit synchroner Spiegelung jedoch unvermeidlich.

Wenn Boot-LUNs, einschließlich virtualisierter Boot-Festplatten, mit Oracle RAC verwendet werden, muss der `misscount` Parameter möglicherweise geändert werden. Weitere Informationen zu RAC-Timeout-Parametern finden Sie unter "Oracle RAC mit ONTAP".

Konfiguration an zwei Standorten

Eine erweiterte RAC-Konfiguration mit zwei Standorten kann aktiv/aktiv-Datenbankservices bereitstellen, die viele, aber nicht alle Ausfallszenarien unterbrechungsfrei überstehen.

RAC-Abstimmungsdateien

Die erste Überlegung bei der Implementierung von Extended RAC auf MetroCluster sollte das Quorum-Management sein. Oracle RAC verfügt über zwei Mechanismen zur Verwaltung des Quorums: Disk Heartbeat und Netzwerk Heartbeat. Der Disk Heartbeat überwacht den Speicherzugriff mithilfe der Abstimmungsdateien. Bei einer RAC-Konfiguration an einem Standort ist eine einzelne Abstimmressource ausreichend, solange das zugrunde liegende Storage-System HA-Funktionen bietet.

In früheren Versionen von Oracle wurden die Abstimmungsdateien auf physischen Speichergeräten abgelegt, aber in aktuellen Versionen von Oracle werden die Abstimmungsdateien in ASM-Diskgroups gespeichert.

> Oracle RAC wird von NFS unterstützt. Während der Grid-Installation wird eine Reihe von ASM-Prozessen erstellt, um den für Grid-Dateien verwendeten NFS-Speicherort als ASM-Diskgruppe darzustellen. Der Prozess ist für den Endbenutzer nahezu transparent und erfordert nach Abschluss der Installation keine laufende ASM-Verwaltung.

In einer Konfiguration mit zwei Standorten ist es als erstes erforderlich, sicherzustellen, dass jeder Standort immer auf mehr als die Hälfte der Abstimmungsdateien zugreifen kann und so einen unterbrechungsfreien Disaster Recovery-Prozess garantiert. Diese Aufgabe war einfach, bevor die Abstimmungsdateien in ASM-Diskgroups gespeichert wurden, aber heute müssen Administratoren grundlegende Prinzipien der ASM-Redundanz verstehen.

ASM-Diskgruppen haben drei Optionen für Redundanz `external`, `normal`, und `high`. Mit anderen Worten: Nicht gespiegelt, gespiegelt und 3-fach gespiegelt. Eine neuere Option namens `Flex` Ist auch verfügbar, aber nur selten verwendet. Die Redundanzstufe und die Platzierung der redundanten Geräte steuern, was in Ausfallszenarien geschieht. Beispiel:

- Platzieren der Abstimmungsdateien auf einem `diskgroup` Mit `external` Bei Ausfall der Verbindung zwischen den Standorten wird durch Redundanzressource die Entfernung eines Standorts garantiert.
- Platzieren der Abstimmungsdateien auf einem `diskgroup` Mit `normal` Redundanz mit nur einer ASM-Festplatte pro Standort garantiert die Entfernung von Knoten auf beiden Standorten, wenn die Verbindung zwischen Standorten verloren geht, da keiner der Standorte ein mehrheitlich Quorum hätte.
- Platzieren der Abstimmungsdateien auf einem `diskgroup` Mit `high` Redundanz mit zwei Festplatten an einem Standort und einer einzigen Festplatte am anderen Standort ermöglicht aktiv-aktiv-Vorgänge, wenn beide Standorte betriebsbereit sind und beide Seiten miteinander erreichbar sind. Wenn der Standort mit einer Festplatte jedoch vom Netzwerk isoliert ist, wird dieser Standort entfernt.

RAC-Netzwerk-Heartbeat

Der Heartbeat des Oracle RAC-Netzwerks überwacht die Erreichbarkeit des Knotens über den Cluster-Interconnect hinweg. Damit ein Node im Cluster verbleiben kann, muss er sich mit mehr als der Hälfte der anderen Nodes in Verbindung setzen können. In einer Architektur mit zwei Standorten werden folgende Auswahlmöglichkeiten für die Anzahl der RAC-Knoten erstellt:

- Die Platzierung einer gleichen Anzahl von Nodes pro Standort führt zu einer Entfernung an einem Standort, falls die Netzwerkverbindung unterbrochen wird.
- Die Platzierung von N Nodes auf einem Standort und N+1 Nodes auf dem anderen Standort garantiert, dass der Verlust der Verbindung zwischen den Standorten zu einer größeren Anzahl von Knoten führt, die im Netzwerk-Quorum verbleiben, und zu einem Standort mit weniger Knoten.

Vor der Einführung von Oracle 12cR2 war es nicht praktikabel zu kontrollieren, auf welcher Seite bei einem Standortausfall eine Entfernung auftreten würde. Wenn jeder Standort über eine gleiche Anzahl von Knoten verfügt, wird die Entfernung vom Master-Knoten gesteuert, der im Allgemeinen der erste RAC-Knoten ist, der gestartet wird.

Oracle 12cR2 bietet Funktionen zur Knotengewichtung. Diese Funktion gibt einem Administrator mehr Kontrolle darüber, wie Oracle Split-Brain-Bedingungen löst. Der folgende Befehl legt als einfaches Beispiel die Präferenz für einen bestimmten Knoten in einem RAC fest:

```
[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.
```

Nach dem Neustart von Oracle High-Availability Services sieht die Konfiguration wie folgt aus:

```
[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
```

Knoten `host-a` Ist jetzt als kritischer Server festgelegt. Wenn die beiden RAC-Knoten isoliert sind, `host-a` Überlebt, und `host-b` Wird entfernt.

Ausführliche Informationen finden Sie im Oracle Whitepaper „Oracle Clusterware 12c Release 2 Technical Overview“. „

Bei Versionen von Oracle RAC vor 12cR2 kann der Master-Knoten identifiziert werden, indem die CRS-Protokolle wie folgt geprüft werden:

```
[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
```

Dieses Protokoll gibt an, dass der Master-Node ist `2` Und dem Knoten `host-a` Hat eine ID von `1`. Diese Tatsache bedeutet das `host-a` Ist nicht der Master-Knoten. Die Identität des Master-Knotens kann mit dem Befehl bestätigt werden `olsnodes -n`.

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

Der Knoten mit der ID von `2` Ist `host-b`, Das ist der Master-Knoten. In einer Konfiguration mit gleicher Anzahl von Knoten an jedem Standort, der Standort mit `host-b` Ist der Standort, der überlebt, wenn die beiden Sets aus irgendeinem Grund die Netzwerkverbindung verlieren.

Der Protokolleintrag, der den Master-Knoten identifiziert, kann möglicherweise aus dem System altern. In diesem Fall können die Zeitstempel der Oracle Cluster Registry (OCR) Backups verwendet werden.

```
[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
```

Dieses Beispiel zeigt, dass der Master-Knoten ist `host-b`. Sie zeigt auch eine Änderung im Master-Knoten von an `host-a` Bis `host-b` Am 4. Mai zwischen 2:05 und 21:39 Uhr. Diese Methode zur Identifizierung des Master-Knotens ist nur dann sicher zu verwenden, wenn die CRS-Protokolle ebenfalls geprüft wurden, da sich der Master-Knoten möglicherweise seit der vorherigen OCR-Sicherung geändert hat. Wenn diese Änderung stattgefunden hat, sollte sie in den OCR-Protokollen sichtbar sein.

Die meisten Kunden wählen eine einzelne Abstimmdiskette, die die gesamte Umgebung und eine gleiche Anzahl von RAC-Knoten an jedem Standort unterstützt. Die Datenträgergruppe sollte auf dem Standort platziert werden, der die Datenbank enthält. Das Ergebnis ist, dass der Verlust der Verbindung zu einer Entfernung am Remote-Standort führt. Der Remote-Standort hätte weder Quorum noch würde er Zugriff auf die Datenbankdateien haben, aber der lokale Standort läuft weiterhin wie gewohnt. Wenn die Konnektivität wiederhergestellt ist, kann die Remote-Instanz wieder online geschaltet werden.

Bei einem Notfall ist eine Umschaltung erforderlich, um die Datenbankdateien und die abstimmende Diskgruppe am verbleibenden Standort online zu schalten. Wenn AUSO die Umschaltung auslösen kann, wird das NVFAIL nicht ausgelöst, da bekannt ist, dass das Cluster synchron ist und die Speicherressourcen ordnungsgemäß online gehen. AUSO ist ein sehr schneller Vorgang und sollte vor dem abgeschlossen werden `disktimeout` Zeitraum läuft ab.

Da es nur zwei Standorte gibt, ist es nicht möglich, eine automatisierte externe Tiebreaking-Software zu verwenden, was bedeutet, dass die erzwungene Umschaltung eine manuelle Operation sein muss.

Konfigurationen mit drei Standorten

Ein erweiterter RAC-Cluster lässt sich mit drei Standorten viel einfacher erstellen. Die beiden Standorte, die jeweils die Hälfte des MetroCluster Systems hosten, unterstützen auch die Datenbank-Workloads, während der dritte Standort als Tiebreaker für die Datenbank und das MetroCluster System dient. Die Oracle Tiebreaker-Konfiguration kann so einfach sein, als ob ein Mitglied der ASM-Diskgroup, die für die Abstimmung

an einem dritten Standort verwendet wird, platziert werden könnte, und kann auch eine Betriebsinstanz am dritten Standort enthalten, um sicherzustellen, dass es eine ungerade Anzahl von Knoten im RAC-Cluster gibt.

Wichtige Informationen zur Verwendung von NFS in einer erweiterten RAC-Konfiguration finden Sie in der Oracle Dokumentation zum Thema „Quorum-Fehlergruppe“. Zusammenfassend kann es sein, dass die NFS-Mount-Optionen geändert werden müssen, um sicherzustellen, dass der Verlust der Verbindung zum dritten Standort, der Quorumressourcen hostet, nicht die primären Oracle-Server oder Oracle RAC-Prozesse hängt.

SnapMirror Active Sync

Überblick

Mit SnapMirror Active Sync können Sie Oracle-Datenbankumgebungen mit extrem hoher Verfügbarkeit aufbauen, in denen LUNs von zwei verschiedenen Storage-Clustern verfügbar sind.

Bei SnapMirror Active Sync gibt es keine „primäre“ und „sekundäre“ Kopie der Daten. Jedes Cluster kann Lese-I/O aus seiner lokalen Kopie der Daten bereitstellen, und jedes Cluster repliziert einen Schreibvorgang auf seinen Partner. Das Ergebnis ist ein symmetrisches I/O-Verhalten.

So können Sie unter anderem Oracle RAC als erweiterten Cluster mit operativen Instanzen an beiden Standorten ausführen. Alternativ können Sie RPO=0 aktiv/Passiv-Datenbank-Cluster erstellen, bei denen Datenbanken mit einer Instanz bei einem Standortausfall zwischen Standorten verschoben werden können. Dieser Prozess kann über Produkte wie Pacemaker oder VMware HA automatisiert werden. Die Grundlage für all diese Optionen ist die synchrone Replizierung, die über SnapMirror Active Sync gemanagt wird.

Synchrone Replizierung

Im normalen Betrieb bietet SnapMirror Active Sync jederzeit ein synchrones RPO=0-Replikat, mit einer Ausnahme. Wenn Daten nicht repliziert werden können, gibt ONTAP die Notwendigkeit zur Replizierung von Daten frei und stellt die E/A-Bereitstellung an einem Standort wieder her, während die LUNs am anderen Standort offline geschaltet werden.

Storage-Hardware

Im Gegensatz zu anderen Disaster Recovery-Lösungen für Storage bietet SnapMirror Active Sync asymmetrische Plattformflexibilität. Die Hardware an den einzelnen Standorten muss nicht identisch sein. Dank dieser Funktion können Sie die Größe der Hardware anpassen, die zur Unterstützung der SnapMirror Active Sync verwendet wird. Das Remote-Storage-System kann identisch mit dem primären Standort sein, wenn es einen vollständigen Produktions-Workload unterstützen muss. Wenn jedoch ein Ausfall zu einer Verringerung der I/O führt, könnte ein kleineres System am Remote-Standort kostengünstiger sein.

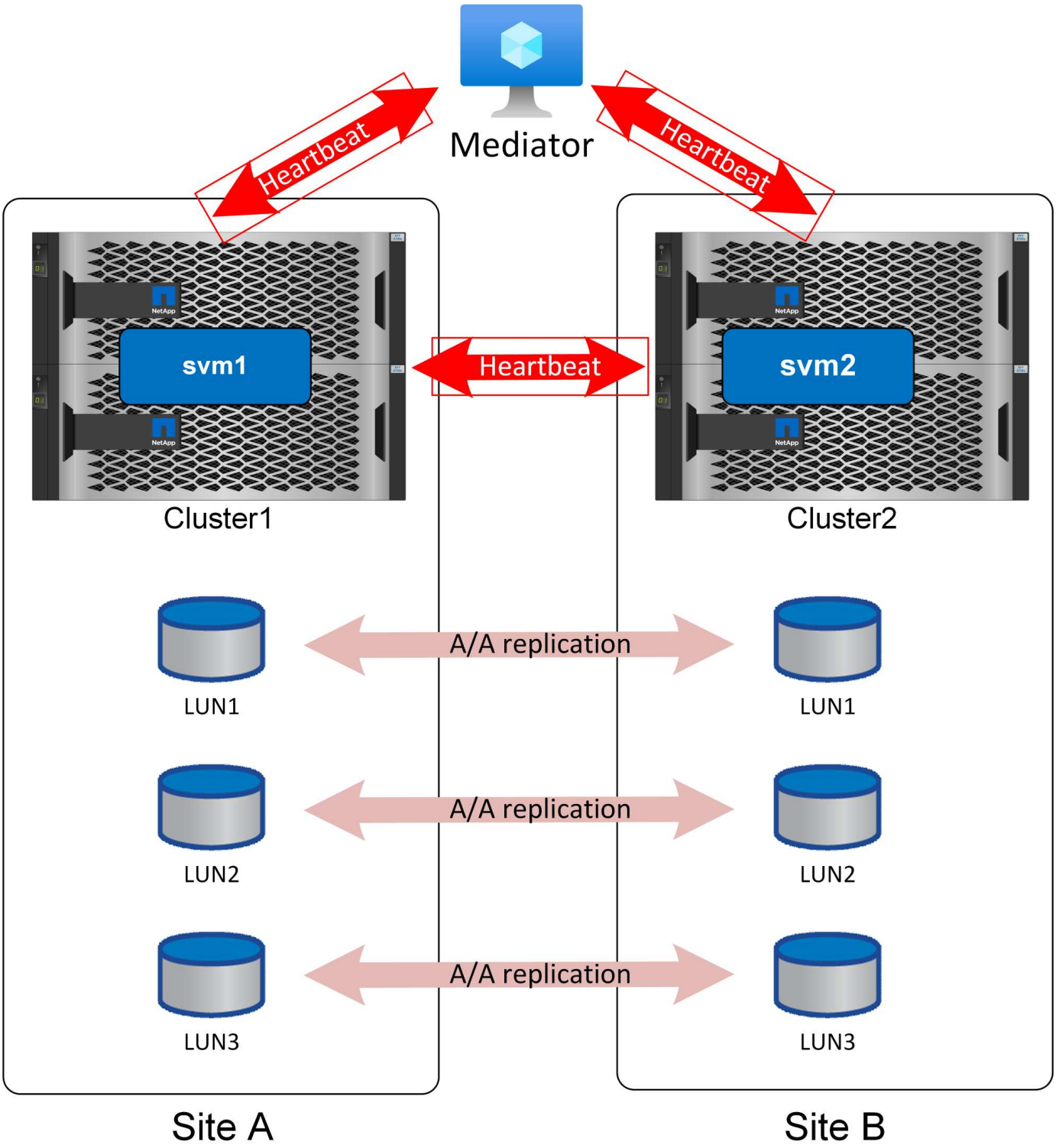
ONTAP Mediator

Der ONTAP Mediator ist eine Softwareanwendung, die von der NetApp-Unterstützung heruntergeladen wird und normalerweise auf einer kleinen virtuellen Maschine bereitgestellt wird. Bei Verwendung mit SnapMirror Active Sync ist der ONTAP Mediator kein Tiebreaker. Es handelt sich um einen alternativen Kommunikationskanal für die beiden Cluster, die an der aktiven synchronen SnapMirror-Replikation beteiligt sind. Der automatisierte Betrieb wird durch ONTAP basierend auf den Antworten gesteuert, die der Partner über direkte Verbindungen und den Mediator erhält.

ONTAP Mediator

Der Mediator ist für die sichere Automatisierung des Failover erforderlich. Idealerweise würde er an einem unabhängigen dritten Standort platziert werden, kann aber dennoch für die meisten Anforderungen funktionieren, wenn er mit einem der an der Replikation beteiligten Cluster kolokiert wird.

Der Mediator ist eigentlich kein Entscheider bei Stimmengleichständen, obwohl er diese Funktion faktisch übernimmt. Der Mediator hilft dabei, den Zustand der Clusterknoten zu ermitteln und unterstützt den automatischen Umschaltprozess im Falle eines Standortausfalls. Der Mediator übermittelt unter keinen Umständen Daten.



Die #1 Herausforderung mit automatisiertem Failover ist das Split-Brain-Problem, und dieses Problem tritt auf, wenn Ihre zwei Standorte die Verbindung miteinander verlieren. Was soll geschehen? Sie möchten nicht, dass sich zwei verschiedene Standorte als verbleibende Kopien der Daten bezeichnen, aber wie kann ein einzelner Standort den Unterschied zwischen dem tatsächlichen Verlust des anderen Standorts und der Unfähigkeit, mit dem gegenüberliegenden Standort zu kommunizieren, erkennen?

Hier betritt der Mediator das Bild. Wenn jeder Standort an einem dritten Standort platziert wird und über eine separate Netzwerkverbindung zu diesem Standort verfügt, haben Sie für jeden Standort einen zusätzlichen Pfad, um den Zustand des anderen zu überprüfen. Sehen Sie sich das Bild oben noch einmal an und

betrachten Sie die folgenden Szenarien.

- Was passiert, wenn der Mediator ausfällt oder von einem oder beiden Standorten nicht erreichbar ist?
  - Die beiden Cluster können weiterhin über dieselbe Verbindung miteinander kommunizieren, die für Replikationsdienste verwendet wird.
  - Für die Daten wird noch eine RPO=0-Sicherung verwendet
- Was passiert, wenn Standort A ausfällt?
  - An Standort B sehen Sie, dass beide Kommunikationskanäle ausgefallen sind.
  - Standort B übernimmt die Datenservices, jedoch ohne RPO=0-Spiegelung
- Was passiert, wenn Standort B ausfällt?
  - An Standort A sehen Sie, dass beide Kommunikationskanäle ausgefallen sind.
  - Standort A übernimmt die Datenservices, aber ohne RPO=0-Spiegelung

Es gibt ein anderes Szenario zu berücksichtigen: Verlust der Datenreplikationsverbindung. Wenn die Replikationsverbindung zwischen Standorten verloren geht, wird eine RPO=0-Spiegelung offensichtlich unmöglich sein. Was soll dann geschehen?

Dies wird durch den bevorzugten Standortstatus gesteuert. In einer SM-AS-Beziehung ist einer der Standorte zweitrangig zum anderen. Dies hat keine Auswirkungen auf den normalen Betrieb, und der gesamte Datenzugriff ist symmetrisch. Wenn die Replikation jedoch unterbrochen wird, muss die Verbindung unterbrochen werden, um den Betrieb wieder aufzunehmen. Das Ergebnis: Der bevorzugte Standort setzt den Betrieb ohne Spiegelung fort und der sekundäre Standort hält die I/O-Verarbeitung an, bis die Replizierungskommunikation wiederhergestellt ist.

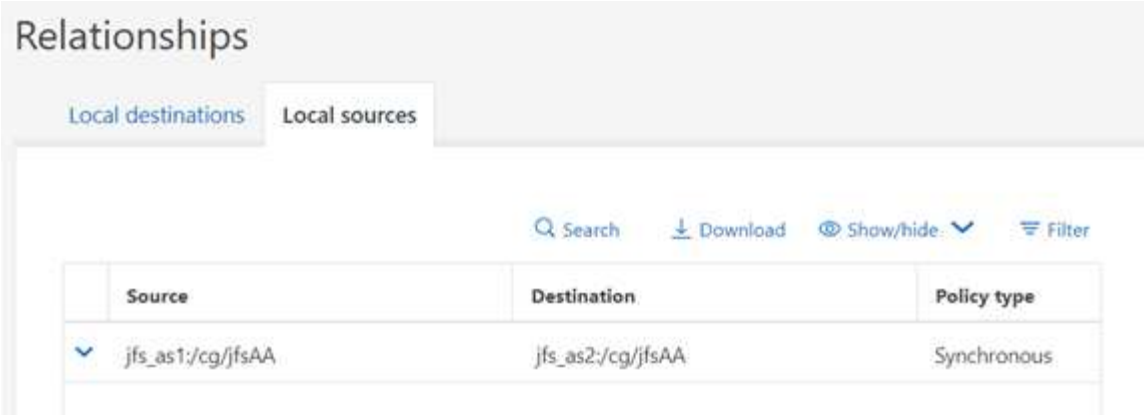
Bevorzugter Standort für SnapMirror Active Sync

Das aktive Synchronisierungsverhalten von SnapMirror ist symmetrisch, mit einer wichtigen Ausnahme: Konfiguration des bevorzugten Standorts.

SnapMirror Active Sync betrachtet einen Standort als „Quelle“ und den anderen als „Ziel“. Dies impliziert eine One-Way-Replikationsbeziehung, aber dies gilt nicht für das IO-Verhalten. Die Replizierung ist bidirektional und symmetrisch, und die I/O-Reaktionszeiten sind auf beiden Seiten der Spiegelung identisch.

Die `source` Bezeichnung steuert den bevorzugten Standort. Wenn die Replizierungsverbindung verloren geht, stellen die LUN-Pfade auf der Quellkopie weiterhin Daten bereit, während die LUN-Pfade auf der Zielkopie erst dann wieder verfügbar sind, wenn die Replikation wiederhergestellt ist und SnapMirror wieder in den synchronen Zustand wechselt. Die Pfade setzen dann das Bereitstellen von Daten fort.

Die Sourcing/Ziel-Konfiguration kann über Systemmanager angezeigt werden:



Oder über die CLI:

```
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
```

Der Schlüssel ist, dass die Quelle die SVM für Cluster1 ist. Wie oben erwähnt, beschreiben die Begriffe „Quelle“ und „Ziel“ nicht den Fluss replizierter Daten. Beide Standorte können einen Schreibvorgang verarbeiten und am anderen Standort replizieren. Beide Cluster sind Quellen und Ziele. Der Effekt der Festlegung eines Clusters als Quelle steuert einfach, welches Cluster als Lese-/Schreib-Speichersystem überlebt, wenn die Replikationsverbindung verloren geht.

Netzwerktopologie

Einheitlicher Zugriff

Ein einheitliches Netzwerk für den Zugriff bedeutet, dass Hosts auf Pfade auf beiden Seiten (oder auf Ausfall-Domains innerhalb desselben Standorts) zugreifen können.

Eine wichtige Funktion von SM-AS ist die Möglichkeit, die Speichersysteme so zu konfigurieren, dass sie wissen, wo sich die Hosts befinden. Wenn Sie die LUNs einem bestimmten Host zuordnen, können Sie angeben, ob sie einem bestimmten Storage-System proximal sind oder nicht.

Annäherungseinstellungen

Proximity bezieht sich auf eine Clusterkonfiguration, die angibt, dass eine bestimmte Host-WWN- oder iSCSI-Initiator-ID zu einem lokalen Host gehört. Dies ist ein zweiter optionaler Schritt für die Konfiguration des LUN-

Zugriffs.

Der erste Schritt ist die übliche igroup-Konfiguration. Jede LUN muss einer Initiatorgruppe zugeordnet werden, die die WWN/iSCSI-IDs der Hosts enthält, die Zugriff auf diese LUN benötigen. Dadurch wird gesteuert, welcher Host *Access* zu einer LUN hat.

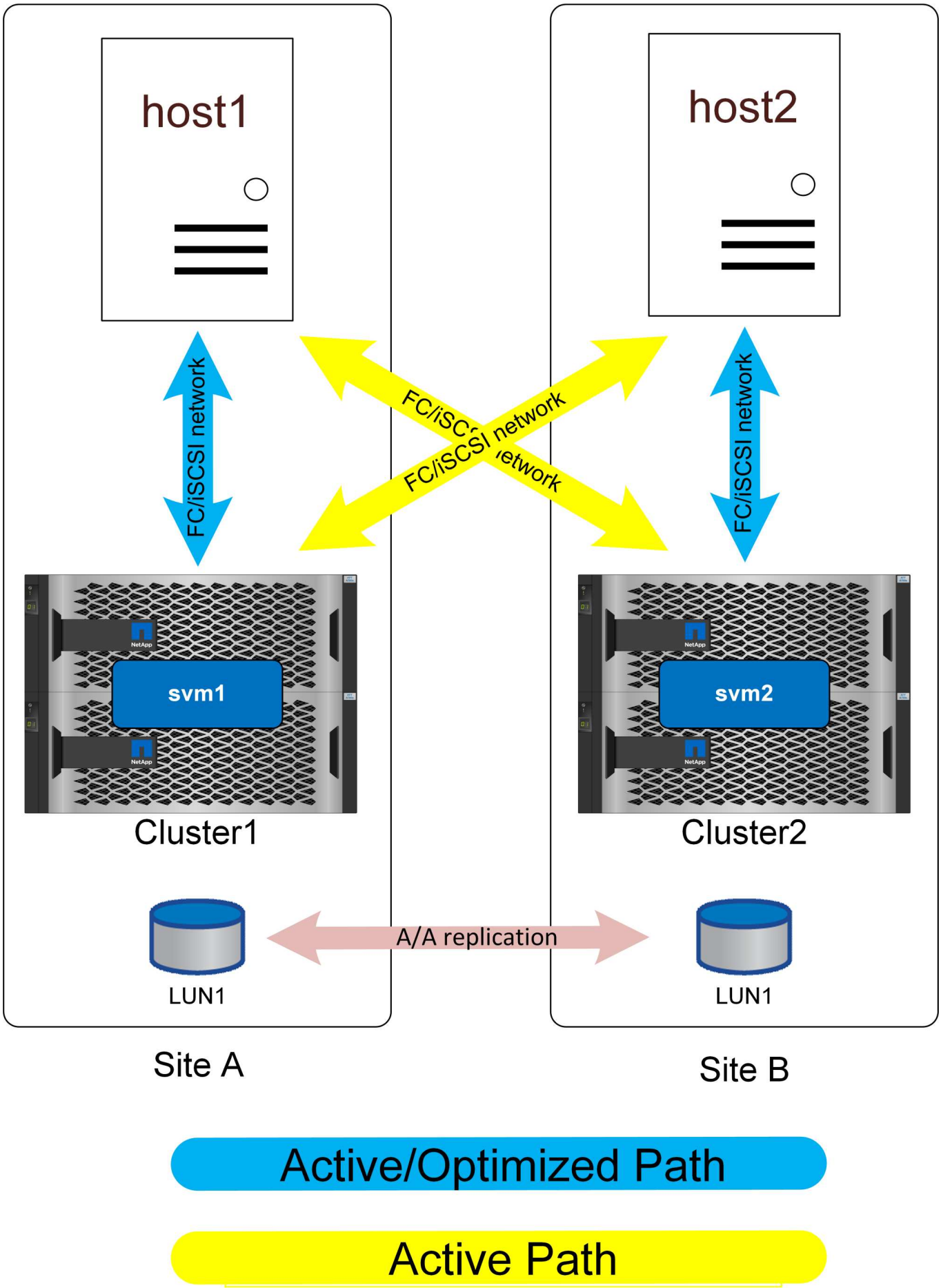
Der zweite, optionale Schritt ist die Konfiguration der Host-Nähe. Dies kontrolliert nicht den Zugriff, es steuert *Priority*.

Beispielsweise kann ein Host an Standort A für den Zugriff auf eine LUN konfiguriert werden, die durch SnapMirror Active Sync geschützt ist. Da das SAN über Standorte erweitert wird, stehen diesem LUN Pfade über Storage an Standort A oder Storage an Standort B zur Verfügung

Ohne Annäherungseinstellungen verwendet der Host beide Speichersysteme gleichmäßig, da beide Speichersysteme aktive/optimierte Pfade anbieten. Wenn die SAN-Latenz und/oder Bandbreite zwischen Standorten begrenzt ist, ist dies möglicherweise nicht erwünscht, und Sie sollten sicherstellen, dass während des normalen Betriebs jeder Host bevorzugt Pfade zum lokalen Speichersystem verwendet. Diese Konfiguration erfolgt durch Hinzufügen der Host-WWN/iSCSI-ID zum lokalen Cluster als proximaler Host. Dies kann unter der CLI oder Systemmanager ausgeführt werden.

AFF

Bei einem AFF System werden die Pfade nach dem Konfigurieren von Host-Nähe wie unten dargestellt angezeigt.


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

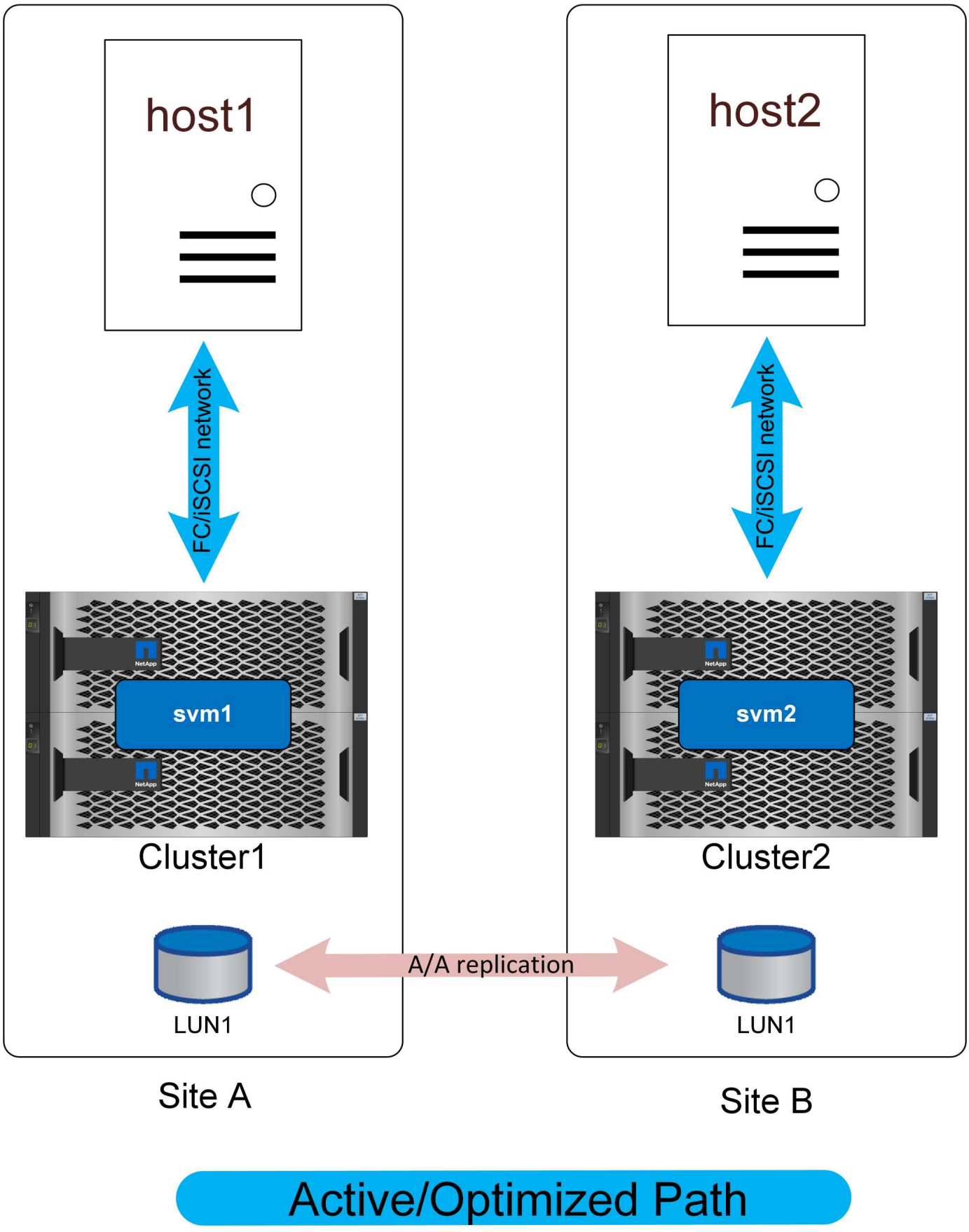
Im normalen Betrieb sind alle E/A-Vorgänge lokal. Lese- und Schreibvorgänge werden vom lokalen Speicher-Array gewartet. Schreib-I/O muss natürlich auch vom lokalen Controller auf das Remote-System repliziert werden, bevor sie bestätigt wird. Alle Lese-I/O-Vorgänge werden jedoch lokal gewartet und es kommt keine zusätzliche Latenz durch das Durchlaufen der SAN-Verbindung zwischen den Standorten zu.

Die nicht optimierten Pfade werden nur dann verwendet, wenn alle aktiven/optimierten Pfade verloren gehen. Wenn beispielsweise das gesamte Array an Standort A Strom verloren hätte, könnten die Hosts an Standort A weiterhin auf Pfade zum Array an Standort B zugreifen und bleiben daher betriebsbereit, obwohl die Latenz höher wäre.

Es gibt redundante Pfade durch den lokalen Cluster, die aus Gründen der Einfachheit nicht auf diesen Diagrammen angezeigt werden. ONTAP Storage-Systeme sind HA selbst, daher sollte ein Controller-Ausfall nicht zu einem Standortausfall führen. Es sollte lediglich zu einer Änderung führen, in der lokale Pfade auf dem betroffenen Standort verwendet werden.

ASA

NetApp ASA Systeme bieten aktiv/aktiv-Multipathing über alle Pfade eines Clusters hinweg. Dies gilt auch für SM-AS Konfigurationen.



Eine ASA-Konfiguration mit nicht-einheitlichem Zugriff würde im Wesentlichen auf die gleiche Weise funktionieren wie mit AFF. Bei einheitlichem Zugriff würde IO das WAN überqueren. Dies kann wünschenswert

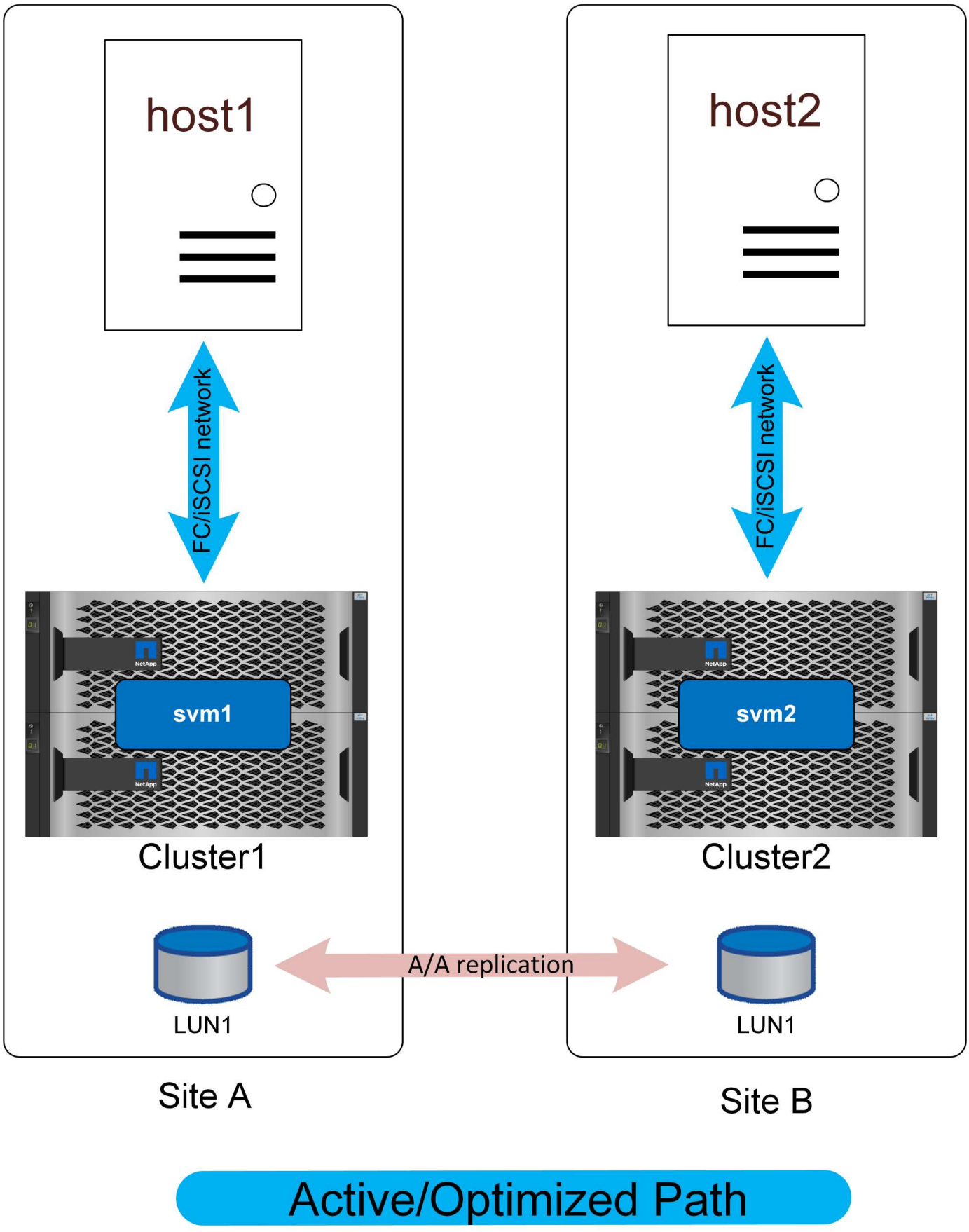
sein oder auch nicht.

Wenn die beiden Standorte mit Glasfaserverbindung 100 Meter voneinander entfernt wären, sollte keine erkennbare zusätzliche Latenz über das WAN entstehen. Wenn jedoch die Standorte weit voneinander entfernt wären, würde die Performance beim Lesen an beiden Standorten darunter leiden. Im Gegensatz dazu würden bei AFF diese WAN-überschneidenden Pfade nur verwendet, wenn keine lokalen Pfade verfügbar wären und die tägliche Performance besser wäre, da alle I/O-Vorgänge lokal wären. ASA mit einem nicht einheitlichen Zugriffsnetzwerk wäre eine Option, um die Kosten- und Funktionsvorteile von ASA zu nutzen, ohne dass sich eine Beeinträchtigung des Zugriffs auf die standortübergreifende Latenz ergeben würde.

ASA mit SM-AS in einer Konfiguration mit niedriger Latenz bietet zwei interessante Vorteile. Zunächst verdoppelt es die Performance bei jedem einzelnen Host *, da IO von doppelt so vielen Controllern mit doppelt so vielen Pfaden gewartet werden kann. Zweitens bietet er in einer Umgebung mit einem einzigen Standort eine extreme Verfügbarkeit, da ein komplettes Storage-System ohne Unterbrechung des Host-Zugriffs verloren gehen könnte.

Uneinheitlicher Zugriff

Uneinheitliches Netzwerk durch Zugriff bedeutet, dass jeder Host nur Zugriff auf Ports im lokalen Storage-System hat. Das SAN wird nicht über Standorte (oder Ausfall-Domains am selben Standort) erweitert.



Der Hauptvorteil dieses Ansatzes ist die SAN-Einfachheit – Sie müssen kein SAN mehr über das Netzwerk erweitern. Einige Kunden verfügen nicht über eine Konnektivität mit niedriger Latenz zwischen den Standorten

und haben nicht die Infrastruktur, um den FC SAN-Datenverkehr über ein standortverbundes Netzwerk zu Tunneln.

Der Nachteil eines uneinheitlichen Zugriffs besteht darin, dass bestimmte Ausfallszenarien, einschließlich des Verlusts der Replikationsverbindung, dazu führen, dass einige Hosts den Zugriff auf den Speicher verlieren. Applikationen, die als einzelne Instanzen ausgeführt werden, wie z. B. eine Datenbank ohne Cluster, die grundsätzlich nur auf einem einzelnen Host bei einem beliebigen Mount ausgeführt wird, würden ausfallen, wenn die lokale Storage-Konnektivität verloren geht. Die Daten bleiben zwar weiterhin geschützt, aber der Datenbankserver würde nicht mehr darauf zugreifen können. Es müsste an einem Remote-Standort neu gestartet werden, vorzugsweise durch einen automatisierten Prozess. VMware HA kann beispielsweise eine heruntergefahrenen Pfade auf einem Server erkennen und eine VM auf einem anderen Server neu starten, auf dem Pfade verfügbar sind.

Im Gegensatz dazu kann eine Cluster-Anwendung wie Oracle RAC einen Service bereitstellen, der gleichzeitig an zwei verschiedenen Standorten verfügbar ist. Der Verlust eines Standorts bedeutet nicht, dass der Anwendungsdienst als Ganzes verloren geht. Instanzen sind nach wie vor verfügbar und werden am verbleibenden Standort ausgeführt.

In vielen Fällen wäre die zusätzliche Latenz, wenn eine Applikation, die auf den Storage über eine Site-to-Site-Verbindung zugreift, nicht akzeptabel. Dies bedeutet, dass die verbesserte Verfügbarkeit von einheitlichem Netzwerk minimal ist, da der Verlust von Speicher an einem Standort dazu führen würde, dass die Dienste auf diesem ausgefallenen Standort sowieso heruntergefahren werden müssen.

Es gibt redundante Pfade durch den lokalen Cluster, die aus Gründen der Einfachheit nicht auf diesen Diagrammen angezeigt werden. ONTAP Storage-Systeme sind HA selbst, daher sollte ein Controller-Ausfall nicht zu einem Standortausfall führen. Es sollte lediglich zu einer Änderung führen, in der lokale Pfade auf dem betroffenen Standort verwendet werden.

Oracle Konfigurationen

Überblick

Die Verwendung von SnapMirror Active Sync trägt nicht notwendigerweise zur Ergänzung oder Änderung von Best Practices für den Betrieb einer Datenbank bei.

Die beste Architektur hängt von den geschäftlichen Anforderungen ab. Wenn das Ziel zum Beispiel ist, RPO=0 Schutz gegen Datenverlust zu haben, aber das RTO entspannt ist, dann kann die Verwendung von Oracle Single Instance Datenbanken und die Replikation der LUNs mit SM-AS ausreichen und auch preisgünstiger von einem Oracle Lizenzierungs-Standard sein. Ein Ausfall des Remote-Standorts würde den Betrieb nicht unterbrechen, und der Verlust des primären Standorts würde zu LUNs am noch intakten Standort führen, die online und einsatzbereit sind.

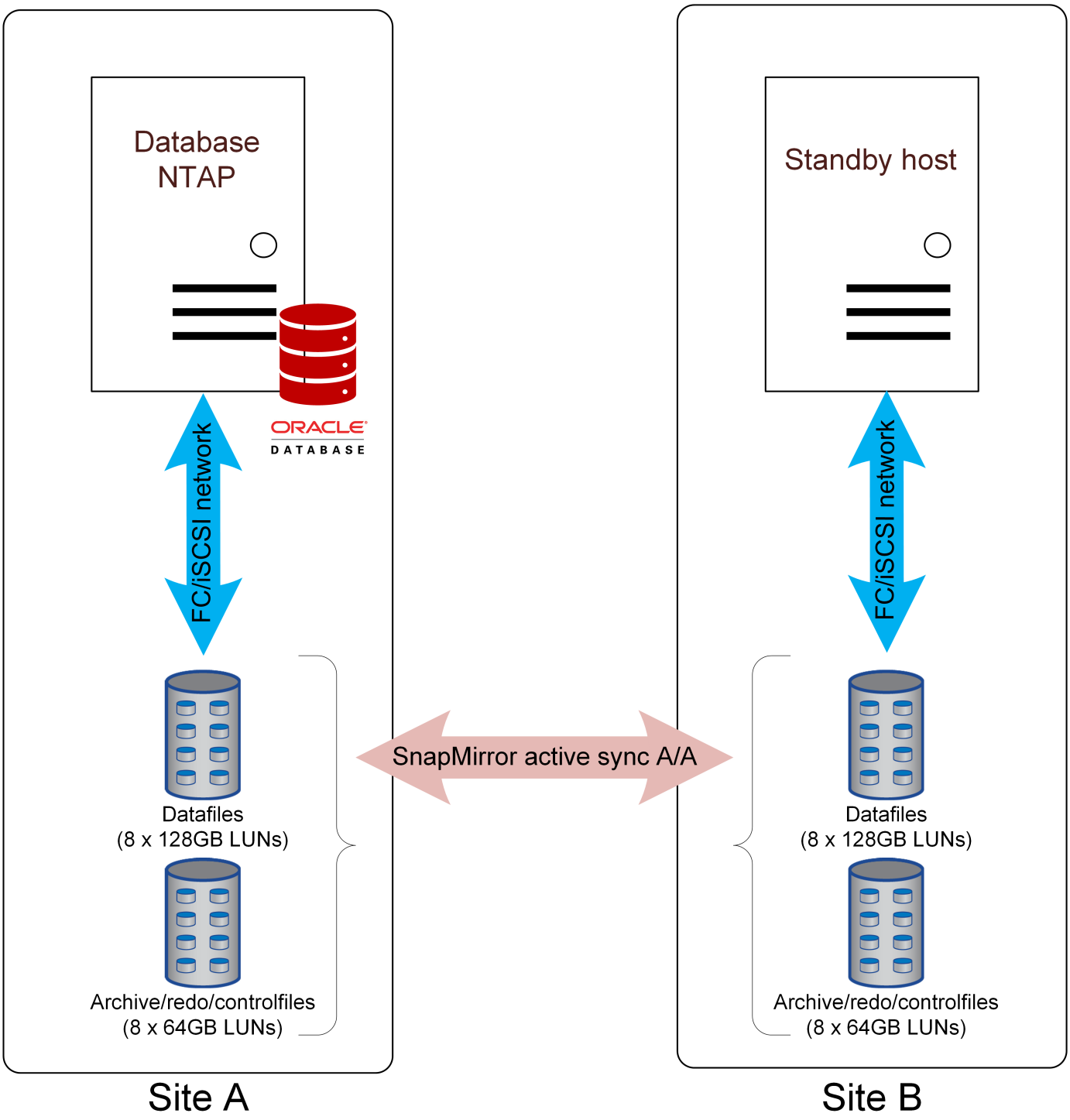
Bei einer strikeren RTO würde die grundlegende aktiv/Passiv-Automatisierung über Skripte oder Clusterware wie Pacemaker oder Ansible die Failover-Zeit verbessern. Beispielsweise könnte VMware HA konfiguriert werden, um den VM-Ausfall am primären Standort zu erkennen und die VM am Remote-Standort zu aktivieren.

Für ein extrem schnelles Failover konnte Oracle RAC über alle Standorte hinweg implementiert werden. Die RTO wäre im Grunde null, da die Datenbank jederzeit online und auf beiden Standorten verfügbar wäre.

Oracle Single Instance

Die unten erläuterten Beispiele zeigen einige der zahlreichen Optionen für die Bereitstellung von Oracle Single-Instance-Datenbanken mit SnapMirror Active Sync-

Replikation.



Failover mit einem vorkonfigurierten Betriebssystem

SnapMirror Active Sync liefert eine synchrone Kopie der Daten am Disaster Recovery-Standort. Für die Verfügbarkeit dieser Daten sind jedoch ein Betriebssystem und die zugehörigen Applikationen erforderlich. Eine grundlegende Automatisierung kann die Failover-Zeit der gesamten Umgebung deutlich verbessern. Clusterware Produkte wie Pacemaker werden häufig eingesetzt, um über die Standorte hinweg einen Cluster zu erstellen, in vielen Fällen ist der Failover-Prozess mit einfachen Skripten angesteuert.

Wenn die primären Knoten verloren gehen, stellt die Clusterware (oder Skripte) die Datenbanken am

alternativen Standort online. Eine Option besteht darin, Standby-Server zu erstellen, die für die SAN-Ressourcen, aus denen die Datenbank besteht, vorkonfiguriert sind. Wenn der primäre Standort ausfällt, führt die Clusterware- oder skriptbasierte Alternative eine Abfolge von Aktionen durch, die der folgenden ähneln:

1. Fehler am primären Standort erkennen
2. Führen Sie eine Erkennung von FC- oder iSCSI-LUNs durch
3. Mounten von Dateisystemen und/oder Mounten von ASM-Datenträgergruppen
4. Die Datenbank wird gestartet

Die primäre Anforderung dieses Ansatzes ist ein Betriebssystem, das am Remote Standort ausgeführt wird. Sie muss mit Oracle-Binärdateien vorkonfiguriert sein, was auch bedeutet, dass Aufgaben wie das Patching von Oracle am primären Standort und am Standby-Standort durchgeführt werden müssen. Alternativ können die Oracle Binärdateien auf den Remote-Standort gespiegelt und gemountet werden, wenn ein Notfall deklariert wird.

Die eigentliche Aktivierung ist einfach. Befehle wie die LUN-Erkennung erfordern nur einige wenige Befehle pro FC-Port. Das Mounten von Dateisystemen ist nichts anderes als ein `mount` Befehl, und sowohl Datenbanken als auch ASM können über die CLI mit einem einzigen Befehl gestartet und gestoppt werden.

Failover mit einem virtualisierten Betriebssystem

Der Failover von Datenbankumgebungen kann auf das Betriebssystem selbst erweitert werden. In der Theorie kann dieses Failover mit Boot-LUNs durchgeführt werden, meistens erfolgt es jedoch mit einem virtualisierten Betriebssystem. Das Verfahren ähnelt den folgenden Schritten:

1. Fehler am primären Standort erkennen
2. Mounten der Datenspeicher, die die virtuellen Maschinen des Datenbankservers hosten
3. Starten der virtuellen Maschinen
4. Manuelles Starten von Datenbanken oder Konfigurieren der virtuellen Maschinen, um die Datenbanken automatisch zu starten.

Beispielsweise kann ein ESX Cluster mehrere Standorte umfassen. Bei einem Notfall können die Virtual Machines nach dem Switchover am Disaster Recovery-Standort online geschaltet werden.

Schutz vor Storage-Ausfällen

Das Diagramm oben zeigt die Verwendung von "Uneinheitlicher Zugriff", wo das SAN nicht über Standorte verteilt ist. Dies ist unter Umständen einfacher zu konfigurieren und ist angesichts der aktuellen SAN-Funktionen in einigen Fällen die einzige Option, was aber auch bedeutet, dass ein Ausfall des primären Storage-Systems einen Datenbankausfall bis zum Failover der Applikation zur Folge hätte.

Für zusätzliche Ausfallsicherheit könnte die Lösung mit implementiert werden"Einheitlicher Zugriff". Dies würde es den Anwendungen ermöglichen, den Betrieb mit den Pfaden fortzusetzen, die vom gegenüberliegenden Standort angezeigt werden.

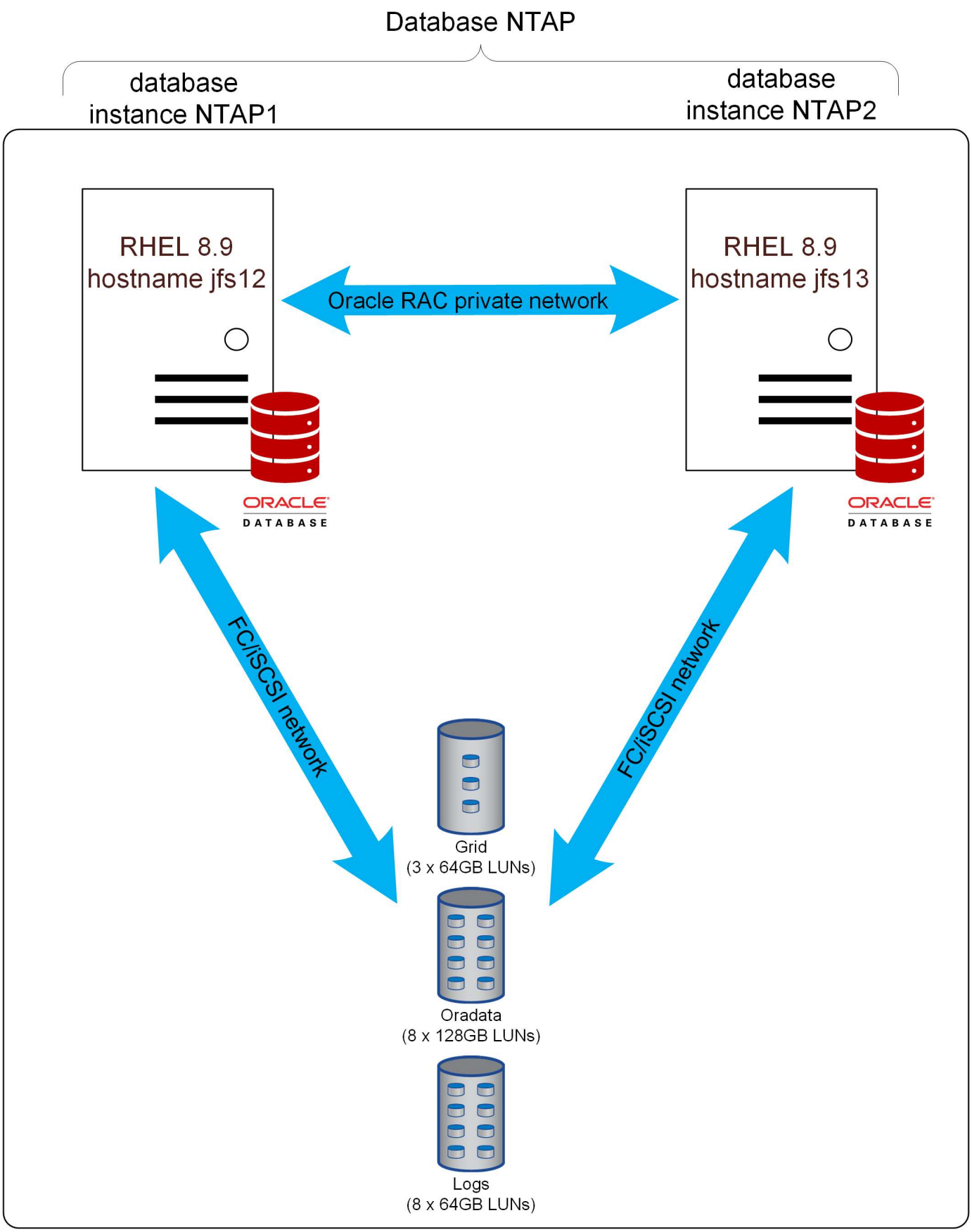
Oracle Extended RAC

Viele Kunden optimieren ihre RTO, indem sie einen Oracle RAC Cluster über mehrere Standorte verteilen und damit eine vollständig aktiv/aktiv-Konfiguration erzielen. Das gesamte Design wird komplizierter, da es die Quorumverwaltung von Oracle RAC beinhalten muss.

Herkömmliche erweiterte RAC-Cluster stützten sich auf ASM-Spiegelung für Datensicherheit. Dieser Ansatz funktioniert, erfordert aber auch zahlreiche manuelle Konfigurationsschritte und führt zu einem Overhead in der Netzwerkinfrastruktur. Da hingegen die SnapMirror Active Sync die Verantwortung für die Datenreplizierung übernehmen kann, wird die Lösung erheblich vereinfacht. Vorgänge wie die Synchronisierung, die Neusynchronisierung nach Unterbrechungen, Failover und das Quorum-Management sind einfacher. Zudem muss das SAN nicht auf mehrere Standorte verteilt werden, was SAN-Design und -Management vereinfacht.

Replizierung

Sie sind für das Verständnis der RAC-Funktionalität auf SnapMirror Active Sync von zentraler Bedeutung, wenn Storage als einheitlicher Satz von LUNs angezeigt wird, die auf gespiegelter Storage gehostet werden. Beispiel:



Es gibt keine primäre Kopie oder gespiegelte Kopie. Logisch gesehen, es gibt nur eine einzige Kopie jeder LUN, und diese LUN ist auf SAN-Pfaden verfügbar, die sich auf zwei verschiedenen Storage-Systemen befinden. Aus Host-Sicht gibt es keine Storage-Failover, sondern Pfadänderungen. Verschiedene

Fehlerereignisse können zum Verlust bestimmter Pfade zum LUN führen, während andere Pfade online bleiben. SnapMirror Active Sync stellt sicher, dass über alle Betriebspfade hinweg dieselben Daten verfügbar sind.

Storage-Konfiguration

In dieser Beispielkonfiguration sind die ASM-Festplatten so konfiguriert, wie sie in jeder RAC-Konfiguration mit einem einzigen Standort auf Enterprise Storage vorhanden wären. Da das Speichersystem Datenschutz bietet, würde ASM externe Redundanz verwendet werden.

Einheitlicher oder uninformierten Zugriff

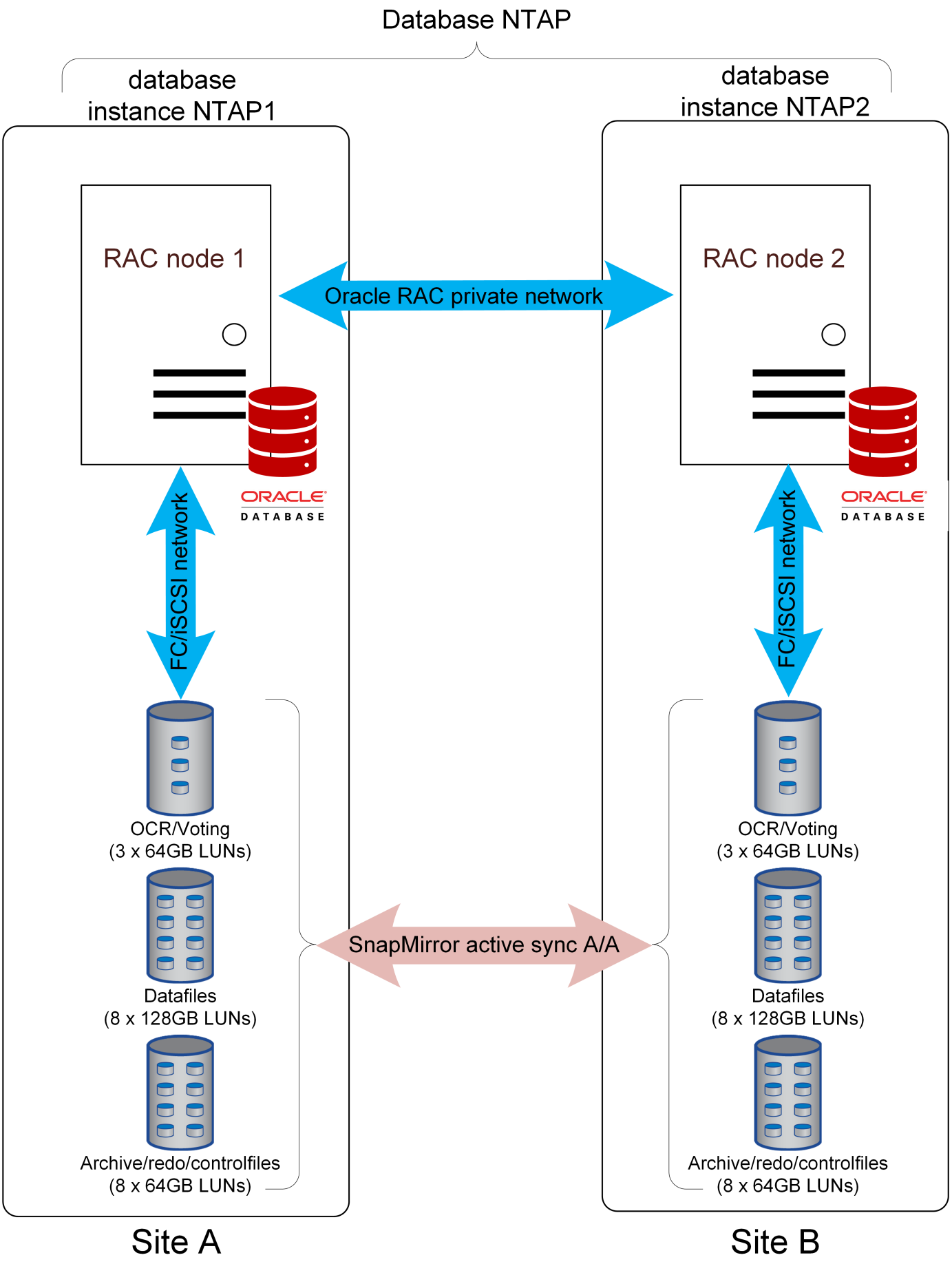
Die wichtigste Überlegung bei Oracle RAC on SnapMirror Active Sync ist, ob ein einheitlicher oder nicht einheitlicher Zugriff verwendet werden soll.

Einheitlicher Zugriff bedeutet, dass jeder Host Pfade auf beiden Clustern sehen kann. Uneinheitlicher Zugriff bedeutet, dass Hosts nur Pfade zum lokalen Cluster sehen können.

Keine Option wird ausdrücklich empfohlen oder abgeraten. Einige Kunden verfügen über Dark Fibre, um Standorte miteinander zu verbinden, andere verfügen entweder über keine solche Konnektivität oder ihre SAN-Infrastruktur unterstützt keine Long-Distance-ISL.

Uneinheitlicher Zugriff

Ein uneinheitlicher Zugriff ist aus SAN-Sicht einfacher zu konfigurieren.


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

Der größte Nachteil des "Uneinheitlicher Zugriff"Ansatzes besteht darin, dass der Verlust der ONTAP-Konnektivität am Standort oder der Verlust eines Storage-Systems zum Verlust der Datenbankinstanzen an einem Standort führen kann. Dies ist natürlich nicht wünschenswert, aber es kann ein akzeptables Risiko im Austausch für eine einfachere SAN-Konfiguration sein.

Einheitlicher Zugriff

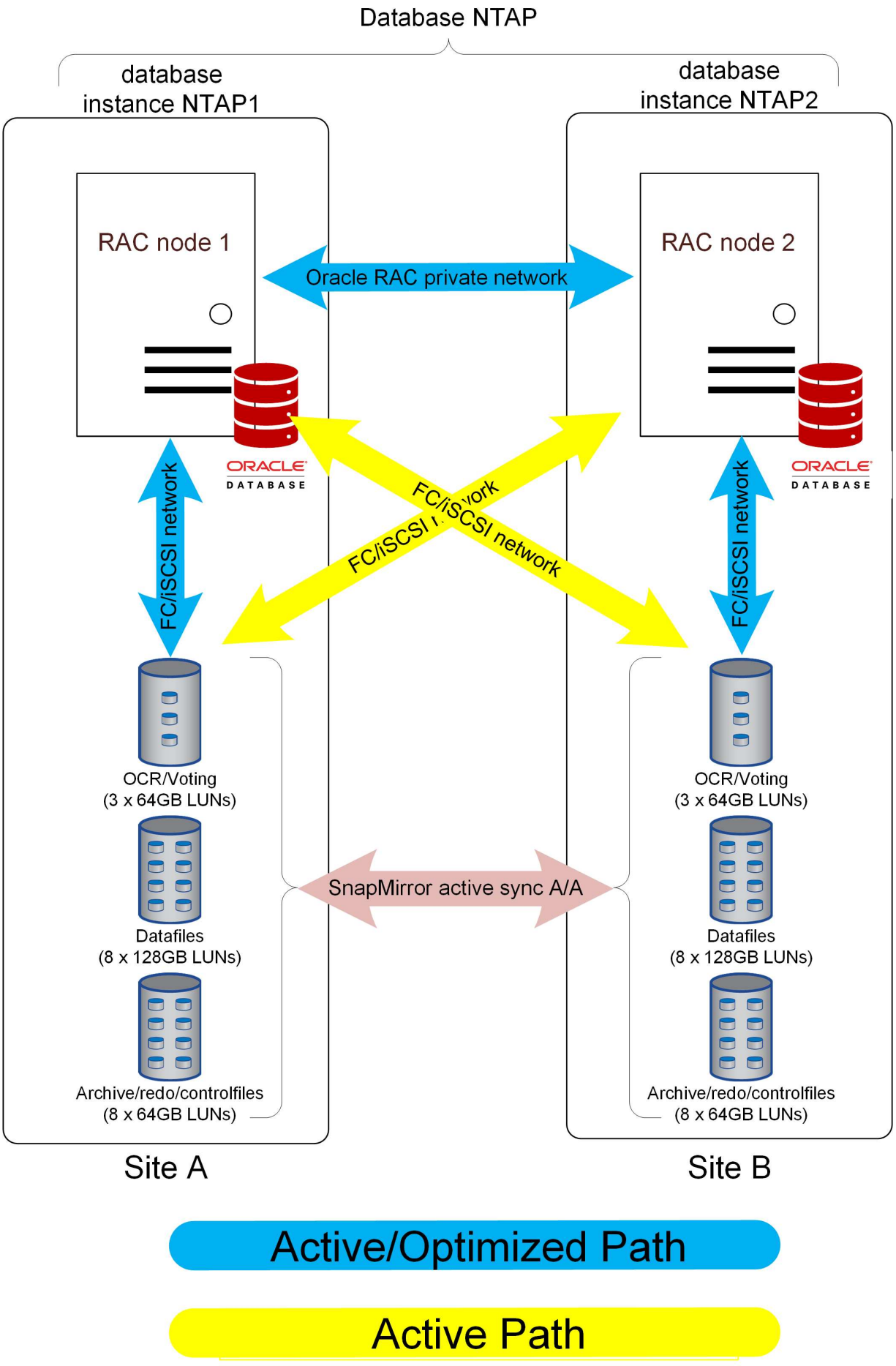
Für einen einheitlichen Zugriff muss das SAN standortübergreifend erweitert werden. Der Hauptvorteil besteht darin, dass der Verlust eines Storage-Systems nicht zum Verlust einer Datenbankinstanz führt. Stattdessen würde dies zu einer Änderung des Multipathing führen, in der Pfade derzeit verwendet werden.

Es gibt mehrere Möglichkeiten, uneinheitlichen Zugriff zu konfigurieren.

> In den Diagrammen unten sind auch aktive, aber nicht optimierte Pfade vorhanden, die bei einfachen Controller-Ausfällen verwendet werden würden. Diese Pfade werden jedoch nicht im Interesse der Vereinfachung der Diagramme angezeigt.

AFF mit Annäherungseinstellungen

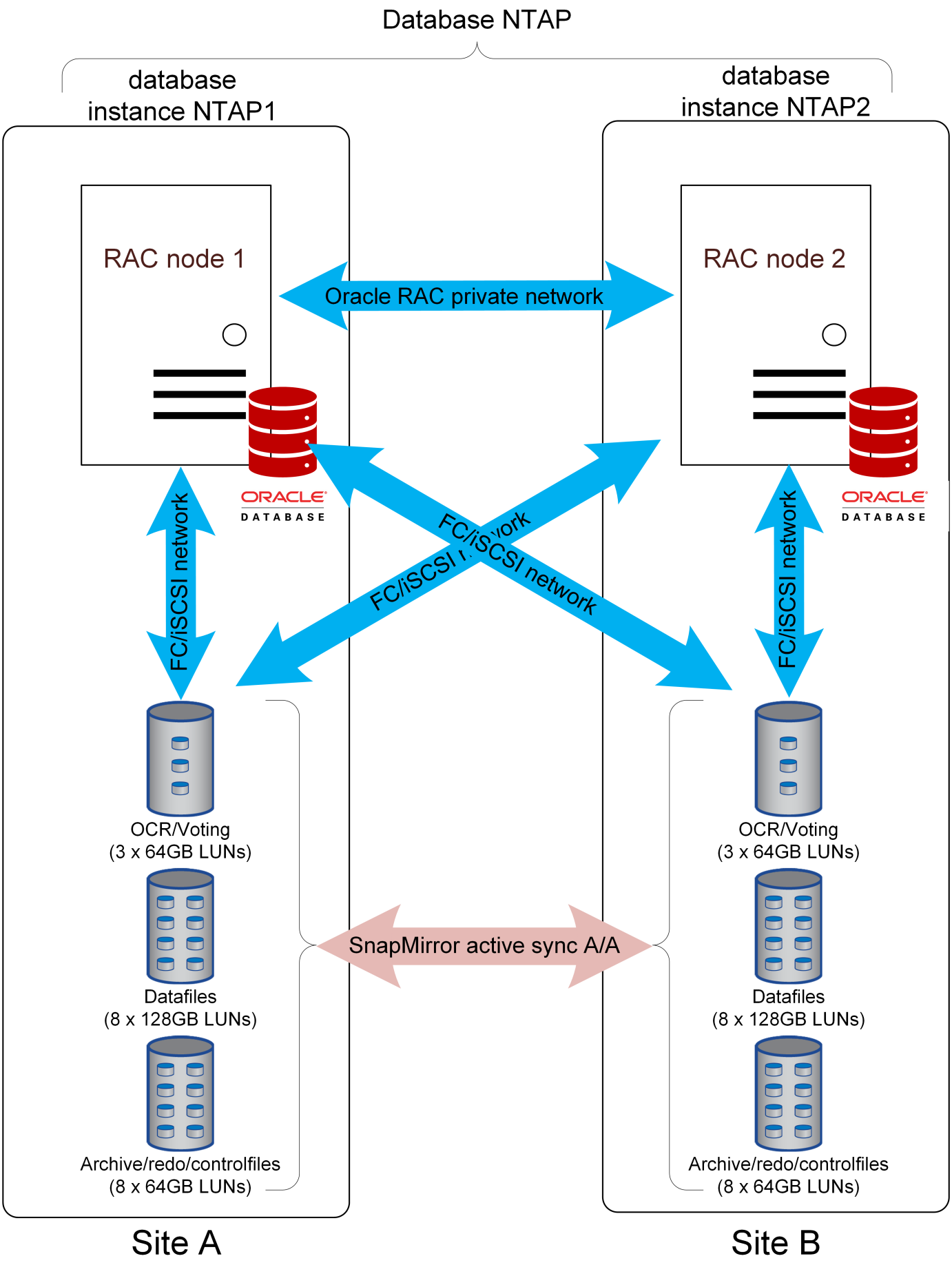
Bei erheblichen Latenzzeiten zwischen den Standorten können AFF Systeme mit Host-Näherungseinstellungen konfiguriert werden. So kann jedes Speichersystem erkennen, welche Hosts lokal und welche Remote sind, und Pfadprioritäten entsprechend zuweisen.


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

Im normalen Betrieb würde jede Oracle-Instanz bevorzugt die lokalen aktiven/optimierten Pfade verwenden. Folglich werden alle Lesezugriffe von der lokalen Kopie der Blöcke bedient. So wird eine möglichst geringe Latenz erzielt. Schreib-I/O wird ähnlich über Pfade zum lokalen Controller gesendet. Die I/O muss noch repliziert werden, bevor sie bestätigt werden kann, und somit würde die zusätzliche Latenz beim Überqueren des Site-to-Site-Netzwerks nach wie vor entstehen. Dies kann in einer Lösung zur synchronen Replizierung jedoch nicht vermieden werden.

ASA / AFF ohne Näherungseinstellungen

Falls keine nennenswerte Latenz zwischen den Standorten erforderlich ist, können AFF Systeme ohne Host-Näherungseinstellungen konfiguriert oder ASA verwendet werden.


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

Jeder Host kann alle Betriebspfade auf beiden Storage-Systemen verwenden. Dies verbessert potenziell die Performance erheblich, da jeder Host das Performance-Potenzial von zwei, nicht nur einem Cluster, nutzen kann.

Mit ASA gelten nicht nur alle Pfade zu beiden Clustern als aktiv und optimiert, sondern auch die Pfade auf den Partner-Controllern wären aktiv. Das Ergebnis wären ständig All-aktiv-SAN-Pfade auf dem gesamten Cluster.

ASA-Systeme können auch in einer uneinheitlichen Zugriffskonfiguration verwendet werden. Da keine standortübergreifenden Pfade vorhanden sind, würde die Performance nicht durch I/O über den ISL beeinträchtigt.

RAC Tiebreaker

Während Extended RAC mit SnapMirror Active Sync eine symmetrische Architektur in Bezug auf IO ist, gibt es eine Ausnahme, die mit Split-Brain-Management verbunden ist.

Was passiert, wenn die Replikationsverbindung verloren geht und keiner der Standorte über ein Quorum verfügt? Was soll geschehen? Diese Frage bezieht sich sowohl auf das Oracle RAC- als auch auf das ONTAP-Verhalten. Wenn Änderungen nicht standortübergreifend repliziert werden können und Sie den Betrieb wieder aufnehmen möchten, muss einer der Standorte überleben und der andere Standort muss nicht mehr verfügbar sein.

Das "ONTAP Mediator" löst diese Anforderung auf ONTAP-Ebene. Es gibt mehrere Optionen für RAC Tiebreaking.

Oracle Tiebreakers

Die beste Methode zur Verwaltung von Split-Brain Oracle RAC-Risiken ist die Verwendung einer ungeraden Anzahl von RAC-Knoten, vorzugsweise unter Verwendung eines Tiebreaker am dritten Standort. Wenn ein dritter Standort nicht verfügbar ist, könnte die Tiebreaker Instanz auf einem Standort der beiden Standorte platziert werden und somit einen bevorzugten Survivor-Standort darstellen.

Oracle und css_Critical

Bei einer geraden Anzahl von Knoten ist das standardmäßige Oracle RAC-Verhalten, dass einer der Knoten im Cluster als wichtiger angesehen wird als die anderen Knoten. Der Standort mit diesem Knoten mit höherer Priorität übersteht die Standortisolierung, während die Knoten am anderen Standort entfernt werden. Die Priorisierung basiert auf mehreren Faktoren, aber Sie können dieses Verhalten auch über die Einstellung steuern `css_critical`.

In der "Beispiel" Architektur sind die Hostnamen für die RAC-Knoten jfs12 und jfs13. Die aktuellen Einstellungen für `css_critical` sind wie folgt:

```
[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.
```

Wenn der Standort mit jfs12 der bevorzugte Standort sein soll, ändern Sie diesen Wert für einen Knoten an Standort A in Ja, und starten Sie die Dienste neu.

```
[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.
```

Ausfallszenarien

Überblick

Die Planung einer vollständigen Applikationsarchitektur für die aktive Synchronisierung von SnapMirror erfordert ein Verständnis dafür, wie SM-AS in verschiedenen geplanten und ungeplanten Failover-Szenarien reagiert.

In den folgenden Beispielen wird davon ausgegangen, dass Standort A als bevorzugter Standort konfiguriert ist.

Verlust der Replikationskonnektivität

Wenn die SM-AS-Replikation unterbrochen wird, kann die Schreib-I/O nicht abgeschlossen werden, da ein Cluster Änderungen nicht auf den anderen Standort replizieren kann.

Standort A (bevorzugte Website)

Das Ergebnis eines Ausfalls der Replikationsverbindung auf dem bevorzugten Standort ist eine ca. 15-Sekunden-Pause bei der Schreib-I/O-Verarbeitung, da ONTAP erneut replizierte Schreibvorgänge versucht, bevor festgestellt wird, dass die Replikationsverbindung wirklich nicht erreichbar ist. Nach 15 Sekunden wird die I/O-Verarbeitung von Lese- und Schreibzugriffen von Standort A fortgesetzt. Die SAN-Pfade ändern sich nicht, und die LUNs bleiben online.

Standort B

Da Standort B nicht der bevorzugte Standort für SnapMirror Active Sync ist, sind die LUN-Pfade nach ca. 15 Sekunden nicht mehr verfügbar.

Ausfall des Storage-Systems

Das Ergebnis eines Storage-Systemausfalls ist nahezu identisch mit dem Ergebnis des Verlusts der Replizierungsverbindung. Am überlebenden Standort sollte eine I/O-Pause von etwa 15 Sekunden stattfinden. Nach Ablauf dieses Zeitraums von 15 Sekunden wird die E/A-Vorgänge wie gewohnt an diesem Standort fortgesetzt.

Verlust des Mediators

Der Mediator hat keine direkte Kontrolle über Storage-Vorgänge. Er fungiert als alternativer Kontrollpfad zwischen Clustern. Die Lösung bietet insbesondere automatisierte Failover-Prozesse ohne Split-Brain-Szenario. Im normalen Betrieb repliziert jedes Cluster Änderungen an seinem Partner. Daher kann jedes Cluster überprüfen, ob das Partner-Cluster online ist und Daten bereitstellt. Wenn die Replikationsverbindung fehlschlägt, wird die Replikation beendet.

Der Grund für einen sicheren automatisierten Failover ist der Mediator, der darauf zurückzuführen ist, dass ein Storage-Cluster andernfalls nicht feststellen kann, ob der Ausfall einer bidirektionalen Kommunikation auf einen Netzwerkausfall oder einen tatsächlichen Storage-Ausfall zurückzuführen ist.

Der Mediator bietet jedem Cluster einen alternativen Pfad zur Überprüfung des Integrität seines Partners. Die Szenarien sind wie folgt:

- Wenn ein Cluster seinen Partner direkt kontaktieren kann, sind die Replizierungsservices betriebsbereit. Keine Aktion erforderlich.
- Wenn ein bevorzugter Standort nicht direkt mit dem Partner oder über den Mediator in Kontakt treten kann, wird davon ausgegangen, dass der Partner entweder tatsächlich nicht verfügbar ist oder isoliert wurde und seine LUN-Pfade offline geschaltet hat. Der bevorzugte Standort setzt dann den Status RPO=0 frei und setzt die Verarbeitung von Lese- und Schreib-I/O fort.
- Wenn ein nicht bevorzugter Standort seinen Partner nicht direkt kontaktieren kann, ihn aber über den Mediator kontaktieren kann, nimmt er seine Pfade offline und wartet auf die Rückkehr der Replikationsverbindung.
- Wenn ein nicht bevorzugter Standort keine direkte Kontaktaufnahme mit dem Partner oder über einen betrieblichen Mediator bietet, nimmt er an, dass der Partner entweder tatsächlich nicht verfügbar ist oder isoliert war und seine LUN-Pfade offline geschaltet hat. Der nicht bevorzugte Standort setzt dann den Status RPO=0 frei und verarbeitet sowohl Lese- als auch Schreib-I/O weiter. Er übernimmt die Rolle der Replikationsquelle und wird der neue bevorzugte Standort.

Wenn der Mediator vollständig nicht verfügbar ist:

- Wenn keine Replizierungsservices aus irgendeinem Grund verfügbar sind, beispielsweise der Ausfall des nicht bevorzugten Standorts oder des Storage-Systems, wird der bevorzugte Standort den Zustand RPO=0 freigeben und die I/O-Verarbeitung für Lese- und Schreibvorgänge wird wieder aufgenommen. Der nicht bevorzugte Standort nimmt seine Pfade offline.
- Ein Ausfall des bevorzugten Standorts führt zu einem Ausfall, da der nicht bevorzugte Standort nicht verifizieren kann, dass der gegenteilige Standort wirklich offline ist. Daher ist es für den nicht bevorzugten Standort nicht sicher, die Services wieder aufzunehmen.

Dienste werden wiederhergestellt

Wenn ein Fehler behoben wurde, wie z. B. die Wiederherstellung der Site-to-Site-Verbindung oder das Einschalten eines ausgefallenen Systems, erkennen die SnapMirror Active Sync-Endpunkte automatisch, dass eine fehlerhafte Replikationsbeziehung vorhanden ist, und versetzen sie wieder in den Zustand RPO=0. Sobald die synchrone Replizierung wiederhergestellt ist, werden die fehlerhaften Pfade wieder online

geschaltet.

In vielen Fällen erkennen Cluster-Applikationen automatisch die Rückgabe ausgefallener Pfade, und diese Applikationen sind ebenfalls wieder online. In anderen Fällen ist möglicherweise ein SAN-Scan auf Host-Ebene erforderlich oder Applikationen müssen manuell wieder online geschaltet werden. Es hängt von der Anwendung und ihrer Konfiguration ab, und im Allgemeinen lassen sich solche Aufgaben leicht automatisieren. ONTAP selbst behebt selbstständig und sollte keinen Benutzereingriff erfordern, um den RPO=0-Storage-Betrieb wiederaufzunehmen.

Manueller Failover

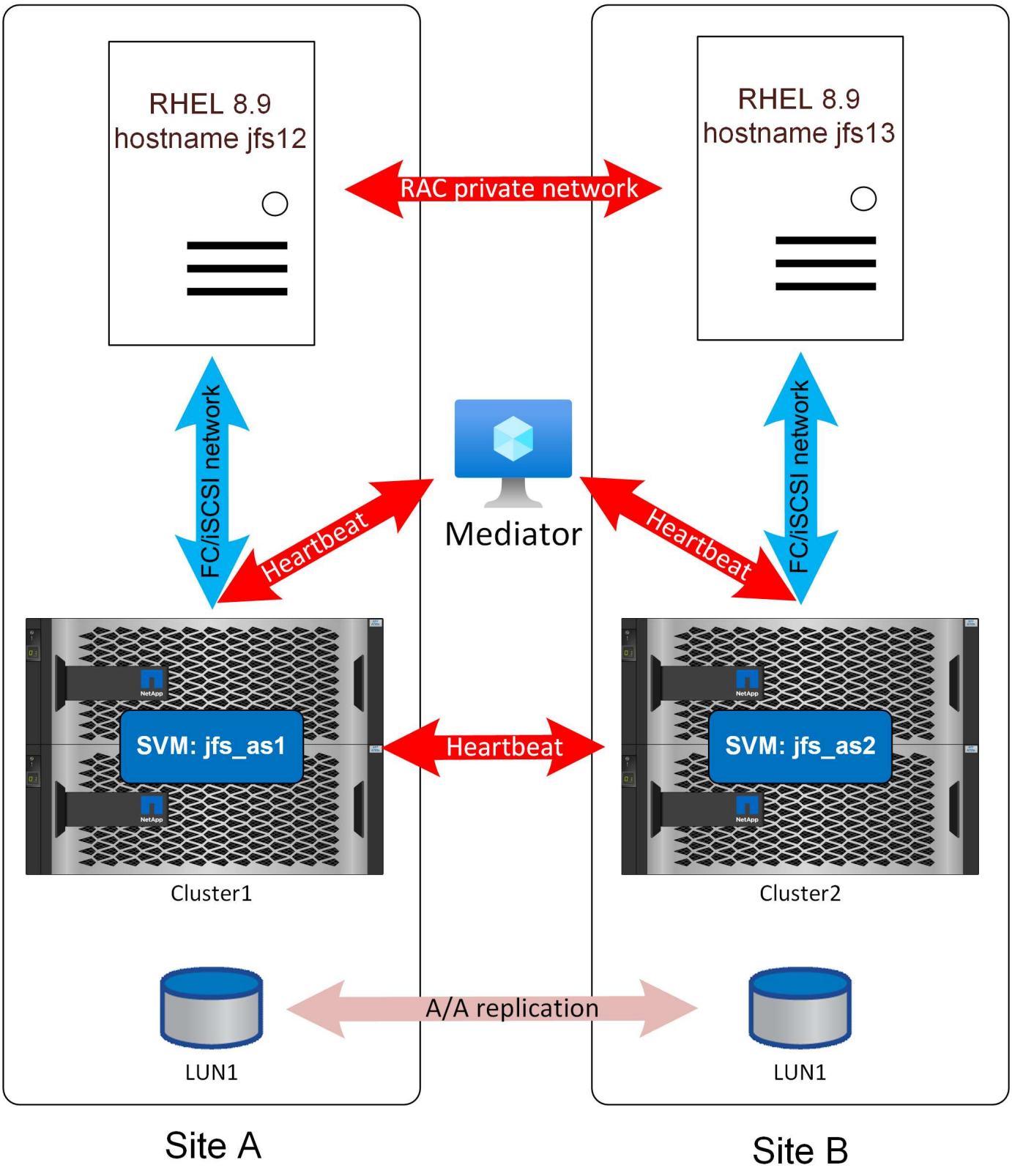
Das Ändern des bevorzugten Standorts erfordert eine einfache Bedienung. I/O-Vorgänge werden für eine oder zwei Sekunden angehalten, da zwischen den Clustern die Berechtigung für das Replikationsverhalten wechselt, die E/A-Vorgänge sind jedoch ansonsten nicht betroffen.

Beispielarchitektur

Die in diesen Abschnitten gezeigten detaillierten Fehlerbeispiele basieren auf der unten dargestellten Architektur.

> Dies ist nur eine von vielen Optionen für Oracle Datenbanken auf SnapMirror Active Sync. Dieses Design wurde gewählt, weil es einige der komplizierteren Szenarien illustriert.

Bei diesem Design wird davon ausgegangen, dass Standort A auf der eingestellt ist"Bevorzugter Standort".



RAC-Verbindungsfehler

Der Verlust der Oracle RAC-Replikationsverbindung führt zu einem ähnlichen Ergebnis wie der Verlust der SnapMirror-Konnektivität, mit Ausnahme der standardmäßig kürzeren Timeouts. In den Standardeinstellungen wartet ein Oracle RAC-Knoten 200 Sekunden

nach Verlust der Speicherverbindung, bevor er entfernt wird, aber er wartet nur 30 Sekunden nach Verlust des RAC-Netzwerk-Heartbeat.

Die CRS-Meldungen ähneln denen unten. Sie können die Zeitlimitüberschreitung von 30 Sekunden sehen. Da css_Critical auf jfs12 gesetzt wurde, befindet sich an Ort A, das wird die Website zu überleben und jfs13 auf Standort B wird entfernt werden.

```
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'.
```

SnapMirror-Kommunikationsfehler

Wenn der SnapMirror Active Sync Replication Link verwendet wird, kann die Schreib-I/O nicht abgeschlossen werden, da ein Cluster Änderungen nicht am anderen Standort replizieren könnte.

Standort A

Das Ergebnis eines Ausfalls einer Replikationsverbindung an Standort A ist eine ca. 15-Sekunden-Pause bei der Schreib-I/O-Verarbeitung, da ONTAP versucht, Schreibvorgänge zu replizieren, bevor es feststellt, dass die Replikationsverbindung wirklich nicht funktionsfähig ist. Nach 15 Sekunden wird das ONTAP Cluster vor Ort A mit der Lese- und Schreib-I/O-Verarbeitung fortgesetzt. Die SAN-Pfade ändern sich nicht, und die LUNs bleiben online.

Standort B

Da Standort B nicht der bevorzugte Standort für SnapMirror Active Sync ist, sind die LUN-Pfade nach ca. 15 Sekunden nicht mehr verfügbar.

Die Replikationsverbindung wurde mit dem Zeitstempel 15:19:44 geschnitten. Die erste Warnung von Oracle RAC kommt 100 Sekunden später, da sich das 200-Sekunden-Timeout (gesteuert durch den Oracle RAC Parameter disktimeout) nähert.

```
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.
```

Sobald das 200-Sekunden-Zeitlimit für Abstimmdatenträger erreicht wurde, wird dieser Oracle RAC-Knoten selbst aus dem Cluster entfernt und neu gestartet.

Totaler Fehler bei der Netzwerkverbindung

Wenn die Replikationsverbindung zwischen den Standorten vollständig unterbrochen wird, werden sowohl die aktive SnapMirror-Synchronisierung als auch die Oracle RAC-Verbindung unterbrochen.

Die Split-Brain-Erkennung von Oracle RAC ist vom Heartbeat des Oracle RAC Storage abhängig. Wenn der Verlust der Site-to-Site-Konnektivität zu einem gleichzeitigen Verlust sowohl des RAC-Netzwerk-Heartbeat als auch der Speicherreplikationsdienste führt, können die RAC-Standorte weder über das RAC-Interconnect noch über die RAC-Abstimmungs-Laufwerke standortübergreifend kommunizieren. Das Ergebnis einer geraden Anzahl von Knoten kann die Entfernung beider Standorte unter den Standardeinstellungen sein. Das genaue Verhalten hängt von der Reihenfolge der Ereignisse und dem Timing des RAC-Netzwerks und der Disk-Heartbeat-Abfragen ab.

Das Risiko eines Ausfalls von 2 Standorten kann auf zwei Arten behoben werden. Zunächst kann eine "Tiebreaker" Konfiguration verwendet werden.

Wenn kein dritter Standort verfügbar ist, kann dieses Risiko durch Anpassung des Parameters für die

Fehlzählung im RAC-Cluster behoben werden. Unter den Standardeinstellungen beträgt das Heartbeat-Timeout des RAC-Netzwerks 30 Sekunden. Dies wird normalerweise von RAC verwendet, um fehlerhafte RAC-Knoten zu identifizieren und aus dem Cluster zu entfernen. Es hat auch eine Verbindung zum Abstimmmedium Heartbeat.

Wenn beispielsweise das Verbindungsrohr, das den Datenverkehr zwischen den Standorten für Oracle RAC und Speicherreplikationsdienste transportiert, durch einen Bagger gekürzt wird, beginnt der 30-Sekunden-Countdown für die Fehlzählung. Wenn der bevorzugte RAC-Standortknoten den Kontakt zum anderen Standort nicht innerhalb von 30 Sekunden wiederherstellen kann und er auch nicht die Abstimmdisks verwenden kann, um zu bestätigen, dass sich der entgegengesetzte Standort innerhalb desselben 30-Sekunden-Fensters befindet, werden die bevorzugten Standortknoten ebenfalls entfernt. Das Ergebnis ist ein vollständiger Ausfall der Datenbank.

Je nachdem, wann die Abfrage der Fehlzählung erfolgt, sind 30 Sekunden möglicherweise nicht genügend Zeit für die SnapMirror Active Sync, um die Zeit zu verkürzen und die Speicherung auf dem bevorzugten Standort zu ermöglichen, um die Dienste wieder aufzunehmen, bevor das 30-Sekunden-Fenster abläuft. Dieses 30-Sekunden-Fenster kann vergrößert werden.

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

Mit diesem Wert kann das Speichersystem am bevorzugten Standort den Betrieb wieder aufnehmen, bevor das Timeout für die Fehlzählung abläuft. Das Ergebnis ist eine Entfernung nur der Knoten am Standort, an dem die LUN-Pfade entfernt wurden. Beispiel unten:

```
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
```

Der Oracle Support rät dringend davon ab, die Parameter „Fehlstellen“ oder „Disktimeout“ zu ändern, um Konfigurationsprobleme zu lösen. Eine Änderung dieser Parameter kann jedoch in vielen Fällen gerechtfertigt und unvermeidbar sein, einschließlich Konfigurationen für SAN-Booting, virtualisierte Konfigurationen und Speicherreplikation. Wenn Sie beispielsweise Stabilitätsprobleme mit einem SAN- oder IP-Netzwerk hatten, das zu RAC-Räumungen führte, sollten Sie das zugrunde liegende Problem beheben und die Werte des Misscount- oder Disktimeout nicht aufladen. Durch das Ändern von Timeouts zur Behebung von Konfigurationsfehlern wird ein Problem maskiert und kein Problem gelöst. Die Änderung dieser Parameter zur ordnungsgemäßen Konfiguration einer RAC-Umgebung basierend auf Designaspekten der zugrunde liegenden Infrastruktur unterscheidet sich und entspricht den Oracle-Support-Anweisungen. Bei SAN-Bootvorgang ist es üblich, Fehlstellen bis zu 200 anzupassen, um Disktimeout zu entsprechen. Weitere Informationen finden Sie unter"Dieser Link".

Standortausfall

Das Ergebnis eines Storage-System- oder Standortausfalls ist nahezu identisch mit dem Ergebnis des Verlusts der Replizierungsverbindung. Am verbleibenden Standort sollte eine I/O-Pause von etwa 15 Sekunden bei Schreibvorgängen stattfinden. Nach Ablauf dieses Zeitraums von 15 Sekunden wird die E/A-Vorgänge wie gewohnt an diesem Standort fortgesetzt.

Wenn nur das Speichersystem betroffen war, gehen die Speicherdienste des Oracle RAC-Knotens am ausgefallenen Standort verloren und führen vor der Entfernung und dem anschließenden Neustart denselben Countdown für die 200-Sekunden-Zeitüberschreitung für die Festplatte ein.

```
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.
```

Der SAN-Pfadstatus auf dem RAC-Knoten, der die Speicherdienste verloren hat, sieht wie folgt aus:

```
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
```

Der linux-Host hat den Verlust der Pfade viel schneller als 200 Sekunden erkannt, aber aus Sicht der Datenbank werden die Clientverbindungen zum Host auf dem ausgefallenen Standort unter den standardmäßigen Oracle RAC-Einstellungen weiterhin 200 Sekunden lang eingefroren. Die vollständigen Datenbankvorgänge werden erst nach Abschluss der Entfernung fortgesetzt.

In der Zwischenzeit zeichnet der Oracle RAC-Knoten am gegenüberliegenden Standort den Verlust des anderen RAC-Knotens auf. Ansonsten funktioniert es wie gewohnt.

```
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
```

Fehler beim Mediator

Der Mediator hat keine direkte Kontrolle über Storage-Vorgänge. Er fungiert als alternativer Kontrollpfad zwischen Clustern. Die Lösung bietet insbesondere automatisierte Failover-Prozesse ohne Split-Brain-Szenario.

Im normalen Betrieb repliziert jedes Cluster Änderungen an seinem Partner. Daher kann jedes Cluster überprüfen, ob das Partner-Cluster online ist und Daten bereitstellt. Wenn die Replikationsverbindung fehlschlägt, wird die Replikation beendet.

Der Grund für den sicheren automatisierten Betrieb ist ein Mediator, der andernfalls nicht feststellen kann, ob der Ausfall einer bidirektionalen Kommunikation auf einen Netzwerkausfall oder einen tatsächlichen Storage-Ausfall zurückzuführen ist.

Der Mediator bietet jedem Cluster einen alternativen Pfad zur Überprüfung des Integrität seines Partners. Die Szenarien sind wie folgt:

- Wenn ein Cluster seinen Partner direkt kontaktieren kann, sind die Replizierungsservices betriebsbereit. Keine Aktion erforderlich.
- Wenn ein bevorzugter Standort nicht direkt mit dem Partner oder über den Mediator in Kontakt treten kann, wird davon ausgegangen, dass der Partner entweder tatsächlich nicht verfügbar ist oder isoliert wurde und seine LUN-Pfade offline geschaltet hat. Der bevorzugte Standort setzt dann den Status RPO=0 frei und setzt die Verarbeitung von Lese- und Schreib-I/O fort.
- Wenn ein nicht bevorzugter Standort seinen Partner nicht direkt kontaktieren kann, ihn aber über den Mediator kontaktieren kann, nimmt er seine Pfade offline und wartet auf die Rückkehr der Replikationsverbindung.

- Wenn ein nicht bevorzugter Standort keine direkte Kontaktaufnahme mit dem Partner oder über einen betrieblichen Mediator bietet, nimmt er an, dass der Partner entweder tatsächlich nicht verfügbar ist oder isoliert war und seine LUN-Pfade offline geschaltet hat. Der nicht bevorzugte Standort setzt dann den Status RPO=0 frei und verarbeitet sowohl Lese- als auch Schreib-I/O weiter. Er übernimmt die Rolle der Replikationsquelle und wird der neue bevorzugte Standort.

Wenn der Mediator vollständig nicht verfügbar ist:

- Ein Ausfall der Replikationsdienste führt aus irgendeinem Grund dazu, dass der bevorzugte Standort den Zustand RPO=0 freigibt und die Lese- und Schreib-I/O-Verarbeitung wieder aufgenommen wird. Der nicht bevorzugte Standort nimmt seine Pfade offline.
- Ein Ausfall des bevorzugten Standorts führt zu einem Ausfall, da der nicht bevorzugte Standort nicht verifizieren kann, dass der gegenteilige Standort wirklich offline ist. Daher ist es für den nicht bevorzugten Standort nicht sicher, die Services wieder aufzunehmen.

Servicewiederherstellung

SnapMirror bietet Selbstreparatur. SnapMirror Active Sync erkennt automatisch eine fehlerhafte Replikationsbeziehung und versetzt sie zurück in den Zustand RPO=0. Sobald die synchrone Replikation wiederhergestellt ist, werden die Pfade wieder online geschaltet.

In vielen Fällen erkennen Cluster-Applikationen automatisch die Rückgabe ausgefallener Pfade, und diese Applikationen sind ebenfalls wieder online. In anderen Fällen ist möglicherweise ein SAN-Scan auf Host-Ebene erforderlich oder Applikationen müssen manuell wieder online geschaltet werden.

Es hängt von der Anwendung und ihrer Konfiguration ab, und im Allgemeinen können solche Aufgaben leicht automatisiert werden. Die SnapMirror Active Sync Software selbst wird automatisch behoben und sollte nach der Wiederherstellung der Stromversorgung und Konnektivität keinen Benutzereingriff erfordern, um die RPO=0-Speichervorgänge wiederaufzunehmen.

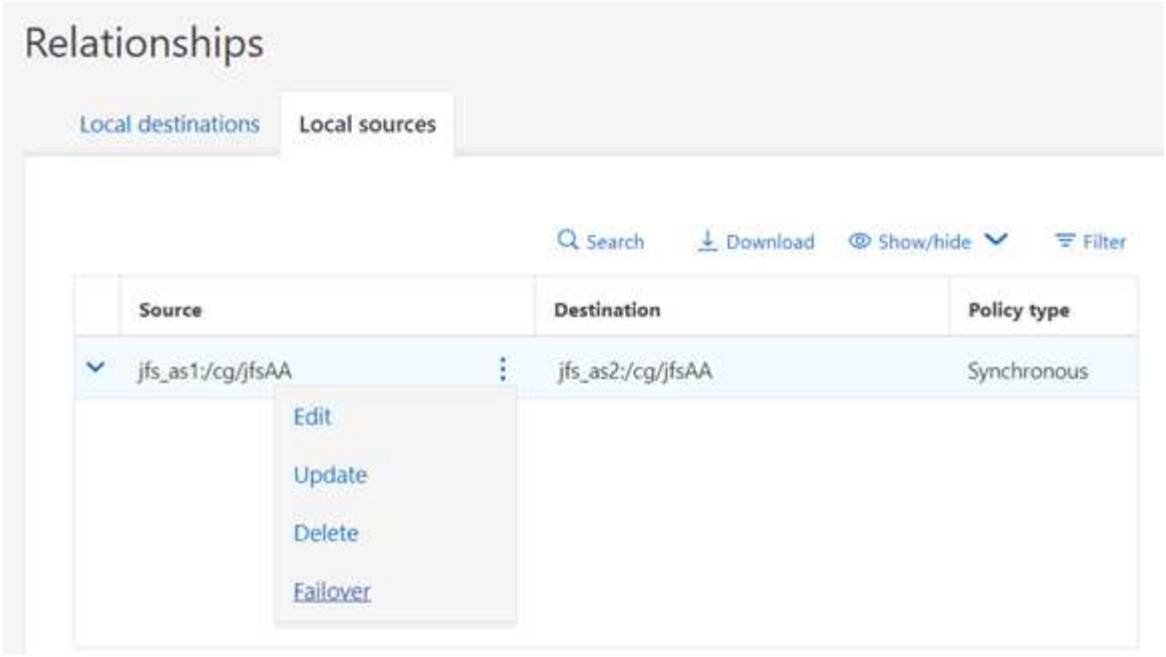
Manueller Failover

Der Begriff „Failover“ bezieht sich nicht auf die Richtung der Replizierung mit SnapMirror Active Sync, da es sich um eine bidirektionale Replizierungstechnologie handelt. Stattdessen bezieht sich „Failover“ darauf, welches Speichersystem bei einem Ausfall der bevorzugte Standort ist.

Möglicherweise möchten Sie beispielsweise ein Failover ausführen, um den bevorzugten Standort zu ändern, bevor Sie einen Standort zu Wartungszwecken herunterfahren oder bevor Sie einen DR-Test durchführen.

Das Ändern des bevorzugten Standorts erfordert eine einfache Bedienung. I/O-Vorgänge werden für eine oder zwei Sekunden angehalten, da zwischen den Clustern die Berechtigung für das Replikationsverhalten wechselt, die E/A-Vorgänge sind jedoch ansonsten nicht betroffen.

GUI-Beispiel:



Beispiel für eine Rückänderung über die CLI:

```
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
```