MetroCluster

Enterprise applications

NetApp
February 10, 2026

Inhalt

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.

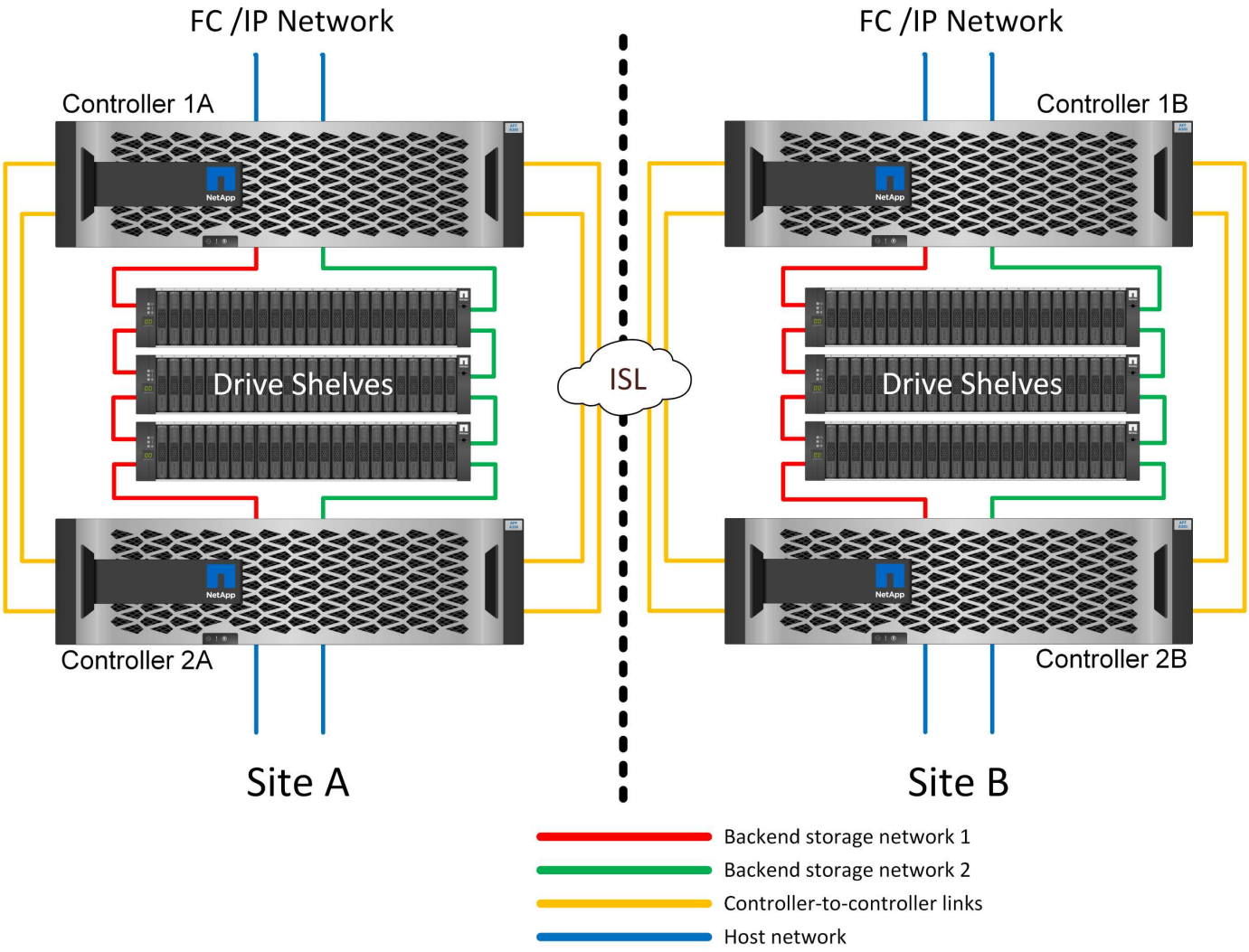
MetroCluster IP

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.

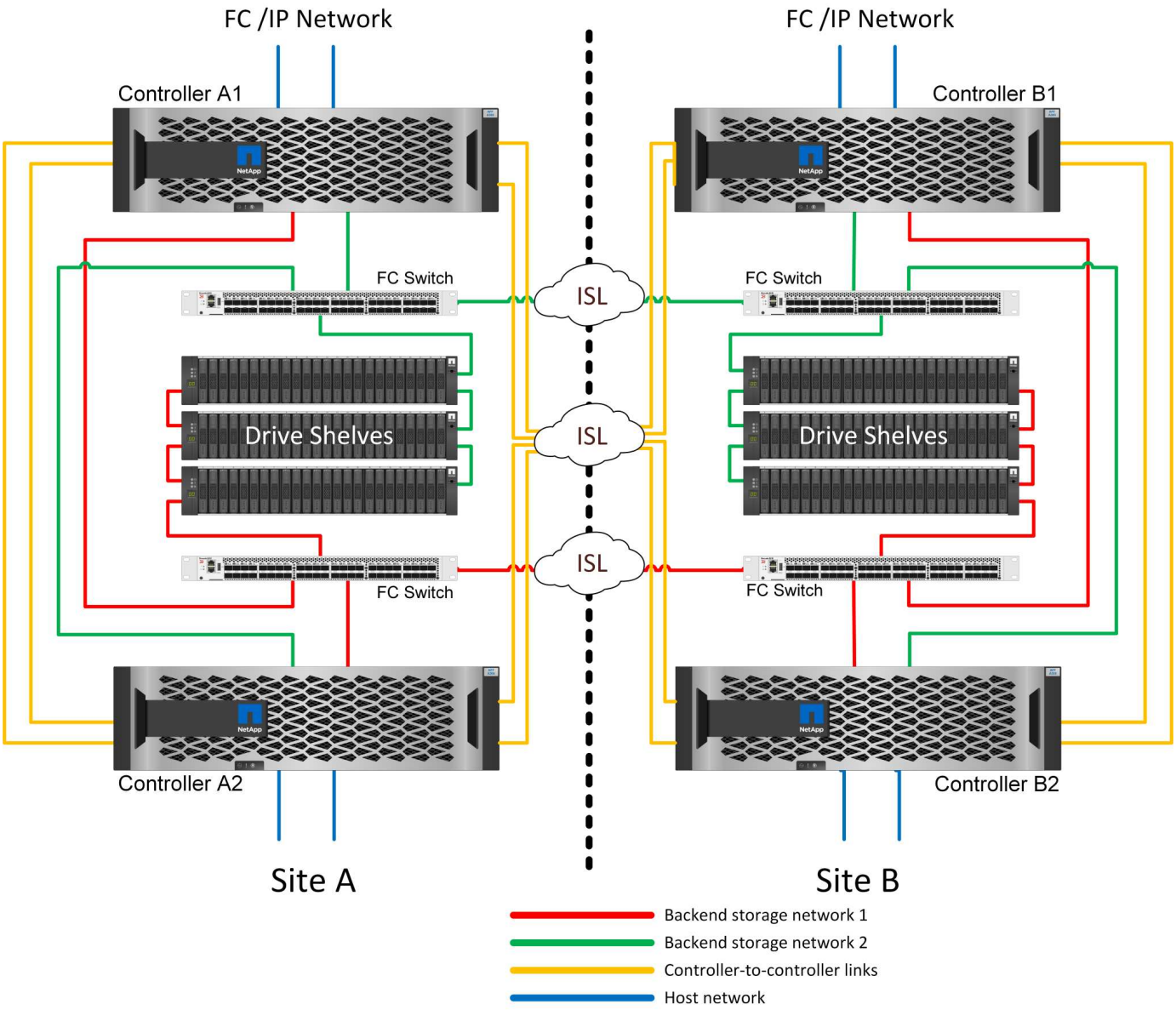
Auch die Architektur ist einfacher, da die einzigen standortübergreifenden Verbindungen für die Controller gelten. Bei FC SAN Attached MetroCluster schreibt ein Controller direkt auf die Laufwerke am entgegengesetzten Standort und benötigt somit zusätzliche SAN-Verbindungen, Switches und Bridges. Ein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



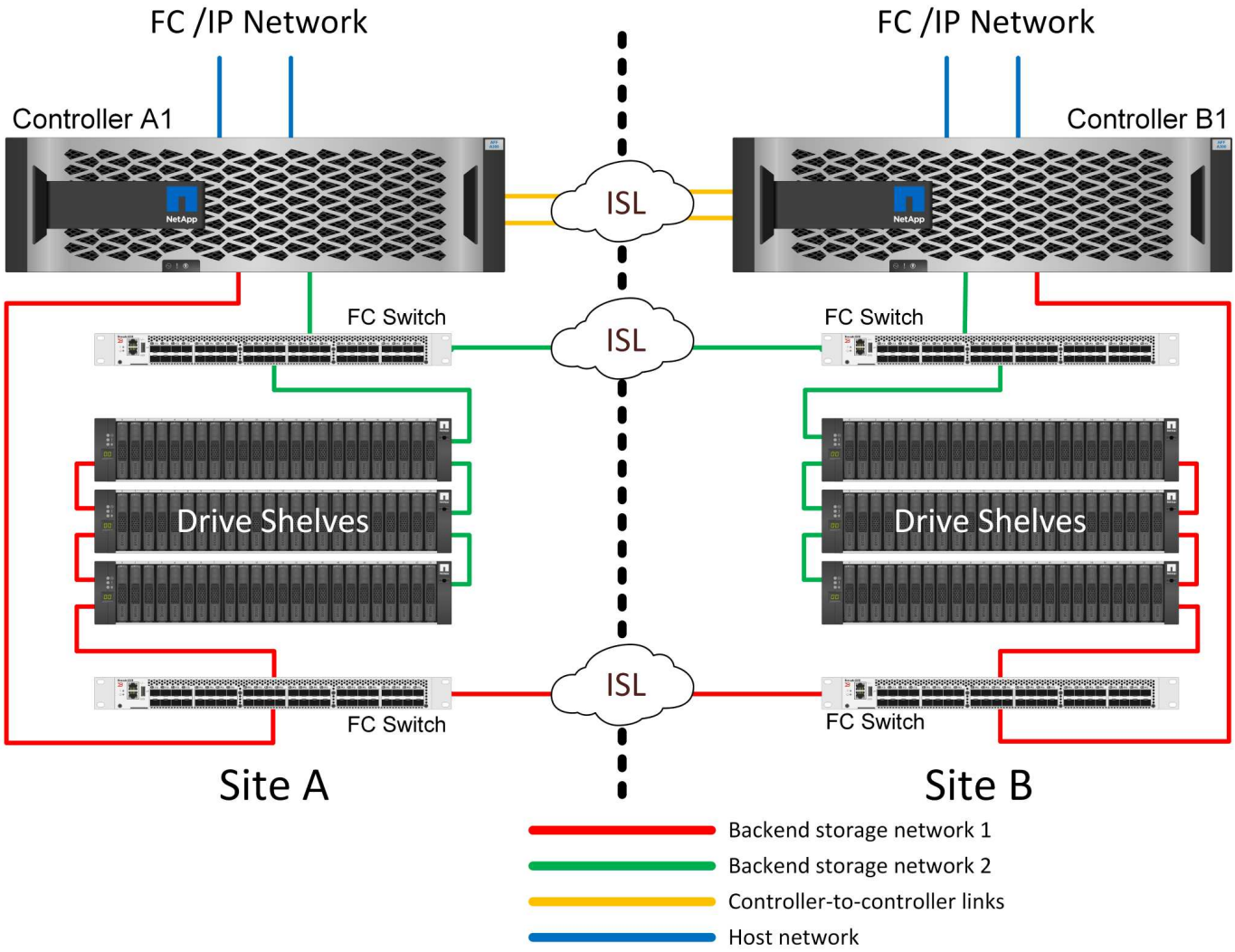
Einige Infrastrukturen mehrerer Standorte sind nicht für den aktiv/aktiv-Betrieb konzipiert, sondern werden eher als primärer Standort und Disaster-Recovery-Standort genutzt. In dieser Situation ist eine MetroCluster-Option für HA-Paare aus den folgenden Gründen im Allgemeinen vorzuziehen:

- 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

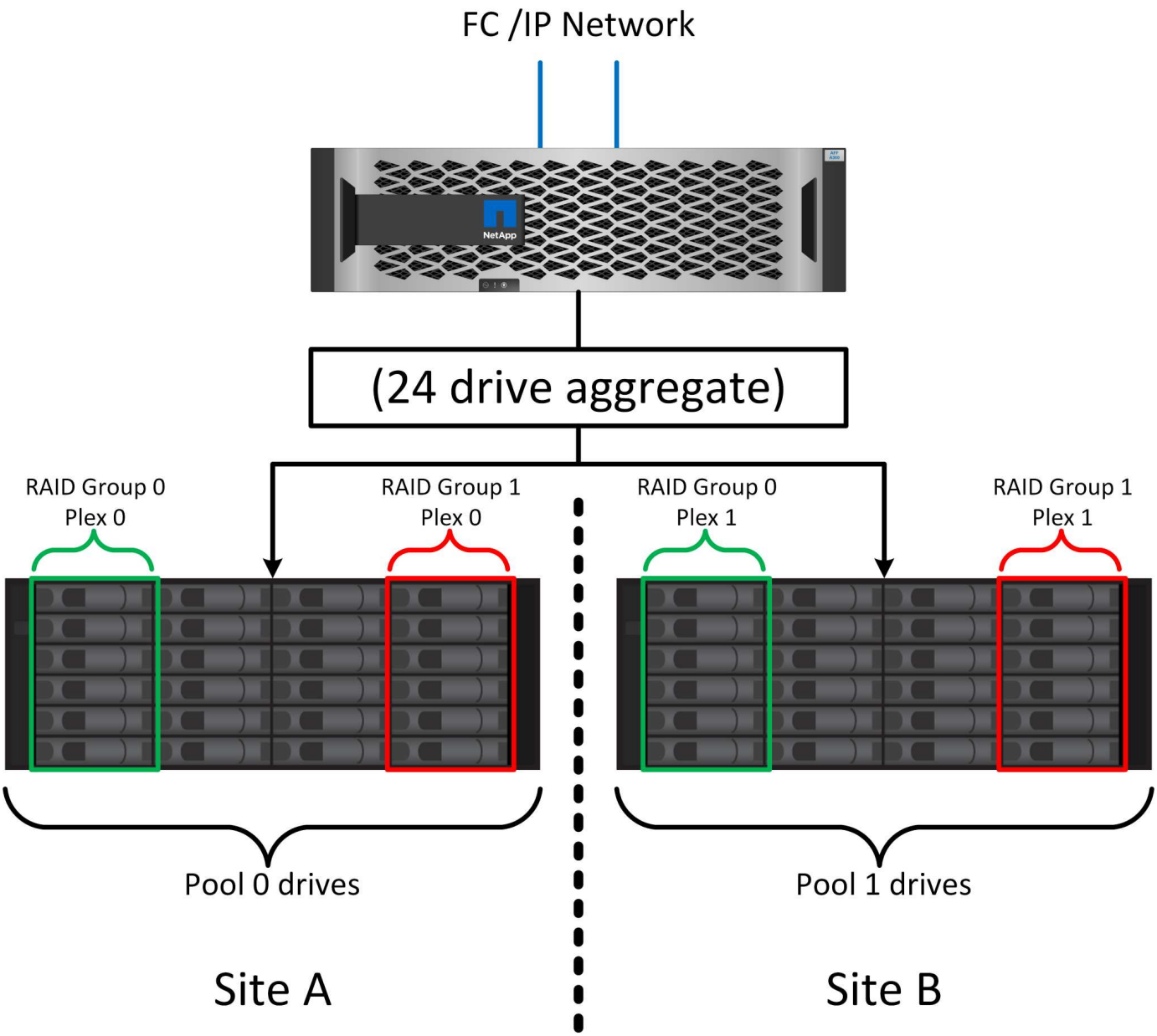
MetroCluster erweitert die NVRAM-Datensicherung auf folgende Weise:

- 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

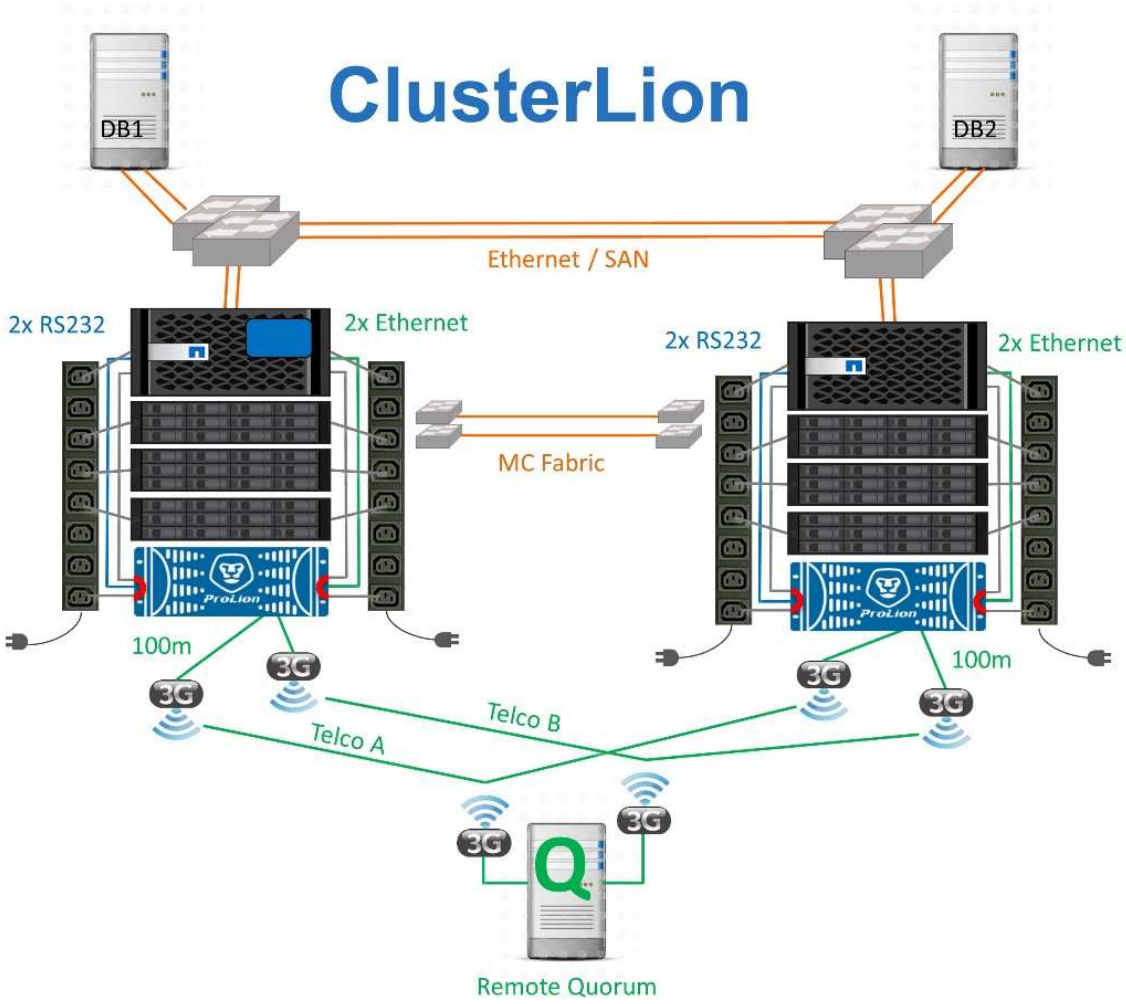
Der ONTAP Mediator wird mit MetroCluster IP und bestimmten anderen ONTAP-Lösungen verwendet. Es fungiert als herkömmlicher Tiebreaker Service, ähnlich wie die oben beschriebene MetroCluster Tiebreaker Software, verfügt aber auch über eine wichtige Funktion zum automatisierten, unbeaufsichtigten Switchover.

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.